Uncertainty estimation via ROCKET

Valerii Kornilov [1]

Abstract

In this project I compare performance of several popular uncertainty estimators applied to time series classifiers based on ROCKET. Uncertainty is estimated either on ensemble of ROCKETs or via dropout.

1. Introduction

Uncertainty estimation has already found applications in different areas of machine learning such as computer vision (Kendall & Gal, 2017; Kendall et al., 2015), natural language processing (Xiao & Wang, 2019), etc. Nowadays, we have lots of classical and deep learning models, which solve huge variety of tasks quite accurately, however, all these models still make mistakes. Thus, models that provide uncertainty in their answer are highly valuable in cases where the cost of mistake is significant, for example, in medicine (Gulshan et al., 2016), autonomous driving (Loquercio et al., 2020). Uncertainty of a model also will be useful in situations when the prediction influences many people's behaviour, such as wheather forecasting or investment recommendations.

In this project, uncertainty is quantified on a large number of time series datasets. The goal is to compare several popular uncertainty estimators to understand, which approach works the best for time series classification task, when we use ROCKET with linear classifier as the base model.

2. Data and preprocessing

2.1. Data

In this work UCR archive (Dau et al., 2019) is used. The current version contains 128 datasets with univariate time series data. 45 of the datasets contain time series of varying length.

[1]Skolkovo Institute of Science and Technology (Skoltech). Correspondence to: <Valerii.Kornilov@skoltech.ru>.

2.2. Preprocessing

The preprocessing step is similar to (Dempster et al., 2020). Fitsly, I've normalized time series (per observation) and interpolated missing values to address the problem of unequal observations lengths for some of the datasets. Also, I've deleted constant observations as uninformative. After generating features with ROCKET the features were normalized and constant features were deleted.

3. Methods

3.1. ROCKET

ROCKET (Dempster et al., 2020) (for RandOm Convolutional KErnel Transform) is an approach for fast generation of features from time series data. Combined with linear classifier (logistic regression or ridge classifier) it produces state-of-the-art results for time series classification task. Basically, it exploits descriptive power of large number of random kernels to aggregate information about time series. In default configuration it uses 10000 random convolutional kernels which vary in the following characteristics:

1. *Length:* One of the values from the set $\{7, 9, 11\}$, sampled with equal probability.
2. *Bias:* Value sampled from uniform distribution $b \sim U[-1, 1]$
3. *Weights:* Values sampled from normal distribution $W \sim \mathcal{N}(\mathbf{0}, I)$ and additionally centered after sampling: $W - \bar{W}$
4. *Dilation:* Value d sampled on an exponential scale ($d = \lfloor 2^x \rfloor$), where $x \sim U[0, \log_2 \frac{\text{length input}-1}{\text{length kernel}-1}]$
5. *Padding:* With equal probability there can be no padding, or padding to keep initial size of input after convolution.

After applying 1d convolution to a time series, the authors aggregate response with two methods: taking the maximum value of response and calculating *ppv* - proportion of positive values. Both these features ensure spatial invariance.

In total, default ROCKET is a two-step procedure:

1. Generate 10000 random kernels and convolve each time series with them.

2. Calculate *max* and *ppv* of response, which produces 20000 features for each time series.

Finally, we can apply some simple classifier to generated ROCKET's features.

3.2. Baseline Uncertainty estimators based on dropout

- Predictive entropy (PE): classic approach (Shannon, 1948) to estimate uncertainty based on the idea of amount of information contained in the received message. If the classifier is absolutely uncertain about class label of incoming observation we can expect that it will produce uniform class probability distribution. If, on the contrary, it is absolutely sure, one-hot vector is expected as the output.

$$H[P(y|x)] = -\sum_{y \in \mathcal{Y}} P(y|x) \log P(y|x)$$

- Mutual Information (MI) is an estimator of epistemic uncertainty (Smith & Gal, 2018) - the one, which appears when posterior of parameters is broad due to the lack of data. It quantifies uncertainty as a gain we could obtain if the label for a new observation x would be known:

$$MI(w, y|D, x) = H[p(y|x, D)] - E_{p(w|D)} H[p(y|x, w)] \quad (1)$$

Both predictive entropy and MI are intractable in general Bayesian framework. But in the case of dropout posterior approximation we could get estimates of these quantities with sampling (Gal & Ghahramani, 2016):

Predictive distribution:

$$p(y|D, x) \simeq \frac{1}{T} \sum_{i=1}^{T} p(y|w_i, x) =: p_{MC}(y|D, x) \quad (2)$$

Predictive Entropy:

$$H[p(y|D, x)] \simeq H[p_{MC}(y|D, x)] \quad (3)$$

MI:

$$MI(w, y|D, x) = H[p_{MC}(y|D, x)] - \quad (4)$$

$$-\frac{1}{T} \sum_{i=1}^{T} H[p(y|w_i, x)] \quad (5)$$

Computationally, it is equivalent to aggregating results of several forward passes.

- Std averaged over runs: this is standard deviation of classes probabilities, which is calculated for each run independently and averaged after that.

The following three metrics are calculated on averaged probabilities from several forward passes. They are rather heuristic, so I don't comment on them.

- Margin: difference between the probability of the most confident class and the probability of the second most confident one.

- Std: minus standard deviation of classes probabilities

- Maxprob: maximum of predicted class probability (Actually, 1-Maxprob is calculated to keep the same logic across metrics: the more is value of the metric, the more is uncertainty.)

3.3. Baseline Uncertainty estimators on ensemble

Here ensemble consists of several ROCKETs feature generators, followed by linear classifiers. No dropout layers are in use. Uncertainty estimators are the same as in the previous subsection but averaging is made over ensemble outputs.

3.4. Approach to evaluate uncertainty estimators

To compare quality of uncertainty estimators I've used rejection rate curves, where on the x-axis there is rejection rate and on the y-axis some metric of model performance (accuracy in the experiments). Under rejection rate (r) I mean the percentage of observations dropped from the sample (of size N) based on the value of the uncertainty proxy.

The whole pipeline is the following:

1. Estimate uncertainty proxy (ex. entropy) for all observations.

2. Drop fraction of observations with the largest values of uncertainty proxy.

3. Calculate accuracy on the remaining part, increase the fraction and repeat from the previous step.

The logic is as following: with more iterations the number of observations, where the model is uncertain, reduces and we can expect to obtain more accurate results for the remaining ones. So, the more misclassified observations are filtered off with an uncertainty measure, the better it is. On the plots that corresponds to the curves, which are lying closer to the ideal rejection curve (IRC):

$$IRC(r) = \frac{TP + TN}{N - \min(N \cdot r, FP + FN)} \quad (6)$$

However, the number of the datasets, on which I compare uncertainty estimators is huge and uncertainty estimators can be compared only within the same dataset. So, I report area under rejection curves (AURC) instead of plotting them. AURC is normalized with the square under IRC. The closer AURC is to 1, the better is the uncertainty measure. You can find an example of rejection curves on fig.1

Also I calculate an analogue of AUC-ROC, which uses classification correctness instead of usual binary labels and penalizes model mistakes for each case, when uncertainty of misclassified observation is lower than for correctly classified observation. This metric can be preferable in the cases, when the number of observations is small and a rejection curve has poor approximation.

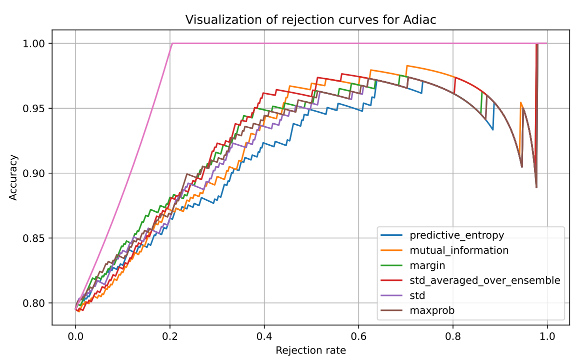

Figure 1. Example of rejection curves and IRC

4. Experiments

First task was to reproduce results of (Dempster et al., 2020). So, the procedure was the same as in the original work. I calculated mean accuracy on ensemble of 10 ROCKETs, followed by ridge classifier.

The second experiment was to get uncertainty estimates based on ensemble of rockets. Here also 10 ROCKETs were in use but with logistic regression as classifier.

The third experiment was to calculate uncertainty using dropout. I've used averaging on 500 runs of ROCKET followed by logistic regression with dropout rate equals 0.1.

For the last two experiments I've calculated in total 24 metric of uncertainty estimators quality per dataset (6 estimators x 2 quality metrics x 2 experiments)

I've used UCR split on train and test as is. For ridge classifier I've used internal scikit-learn leave-one-out procedure for cross-validation. When using logistic regression as classifier, I've implemented stratified split with 30% observations going to validation and used this part for early stopping determination. The time consumption of training logistic regression on large number of datasets was prohibitive for best regularization parameter searching, so in all experiments it was fixed (1e-2).

After evaluation uncertainty estimators I compared their performance across datasets, using standard two-sided t-criteria on two related samples (with H_0: no difference in performance).

5. Results and discussion

From tab.1 it can be said that among metrics calculated via dropout the best one is Maxpob, but the difference is negligible. The whole batch of metrics calculated on ensemble (Margin, Maxprob, Predictive Entropy, Std) are statistically better than all metrics calculated on dropout. From tab.2 MI is the worst estimator on ensemble, the other ones are close in performance. So, if measuring performance of uncertainty estimator with ROC-AUC, the estimators, calculated on ensemble of ROCKETs are slightly better. Also, Maxprob is statistically no worse than any other uncertainty estimator. When measuring performance with AURC (tab.3, 4) conclusions about better uncertainty via ensemble and dominance of Maxprob are hold. Expanded results (per dataset) on quality (ROC-AUC or AURC) of uncertainty estimators measured on ensemble or via dropout can be found in tabs [5, 6, 7,8].

6. Conclusions

In this project several uncertainties estimators were compared on multiple time series datasets. Small increase of uncertainty estimation can be achieved with the usage of ROCKETs ensemble. Among the compared estimators Maxprob has shown the best performance.

Table 1. Difference in quality (ROC-AUC) of estimators. In the first block of rows I compare performance of estimators for the same model (dropout). In the second block I compare uncertainty metrics, calculated on ensemble, with the ones, calculated on dropout. Stars denotes significance level *-0.1, **-0.05, ***-0.001 (two-sided t-test)

Model	Metric	Margin	Maxprob	MI	PE	Std (ens)	Std
Dropout	Margin	-	-0.001^{***}	0.003	0.001	-0.001	-0.001
	Maxprob	0.001^{***}	-	0.005	0.003^{***}	0.000	0.000^{*}
	MI	-0.003	-0.005	-	-0.002	-0.004^{***}	-0.004
	PE	-0.001	-0.003^{***}	0.002	-	-0.002	-0.002^{***}
	Std (ens)	0.001	-0.000	0.004^{***}	0.002	-	-5.794
	Std	0.001	-0.000^{*}	0.004	0.002^{***}	5.794	-
Ensemble	Margin	0.015^{***}	0.013^{**}	0.018^{***}	0.016^{***}	0.014^{**}	0.013^{**}
	Maxprob	0.016^{***}	0.014^{**}	0.019^{***}	0.017^{***}	0.015^{**}	0.015^{***}
	MI	-0.005	-0.007	-0.001	-0.003	-0.006	-0.006
	PE	0.012^{**}	0.010^{*}	0.015^{***}	0.013^{**}	0.011^{**}	0.011^{**}
	Std (ens)	0.006	0.004	0.010	0.007	0.005	0.005
	Std	0.015^{***}	0.013^{**}	0.018^{***}	0.016^{***}	0.014^{**}	0.014^{**}

Table 2. Difference in quality (ROC-AUC) of estimators. In the second block of rows I compare performance of estimators for the same model (ensemble). In the fisrt block I compare uncertainty metrics, calculated via dropout, with the ones, calculated on ensemble. Stars denote significance level *-0.1, **-0.05, ***-0.001 (two-sided t-test)

Model	Metric	Margin	Maxprob	MI	PE	Std (ens)	Std
Dropout	Margin	-0.015^{***}	-0.016^{***}	0.005	-0.012^{**}	-0.006	-0.015^{***}
	Maxprob	-0.013^{**}	-0.014^{**}	0.007	-0.010^{*}	-0.004	-0.013^{**}
	MI	-0.018^{***}	-0.019^{***}	0.001	-0.015^{***}	-0.010	-0.018^{***}
	PE	-0.016^{***}	-0.017^{***}	0.003	-0.013^{**}	-0.007	-0.016^{***}
	Std (ens)	-0.014^{**}	-0.015^{**}	0.006	-0.011^{**}	-0.005	-0.014^{**}
	Std	-0.013^{**}	-0.015^{***}	0.006	-0.011^{**}	-0.005	-0.014^{**}
Ensemble	Margin	-	-0.001^{**}	0.020^{***}	0.002^{*}	0.008^{***}	-0.000
	Maxprob	0.001^{**}	-	0.021^{***}	0.004^{***}	0.009^{***}	0.000
	MI	-0.020^{***}	-0.021^{***}	-	-0.017^{***}	-0.011^{***}	-0.020^{***}
	PE	-0.002^{*}	-0.004^{***}	0.017^{***}	-	0.005^{**}	-0.003^{***}
	Std (ens)	-0.008^{***}	-0.009^{***}	0.011^{***}	-0.005^{**}	-	-0.008^{***}
	Std	0.000	-0.000	0.020^{***}	0.003^{***}	0.008^{***}	-

Table 3. Difference in quality (AURC) of estimators. In the first block of rows I compare performance of estimators for the same model (dropout). In the second block I compare uncertainty metrics, calculated on ensemble, with the ones, calculated on dropout. Stars denotes significance level *-0.1, **-0.05, ***-0.001 (two-sided t-test)

Model	Metric	Margin	Maxprob	MI	PE	Std (ens)	Std
Dropout	Margin	-	-0.002^{**}	-0.002	0.008	-0.003	-0.000
	Maxprob	0.002^{**}	-	-0.000	0.010	-0.000	0.001^{**}
	MI	0.002	0.000	-	0.011	-0.000	0.002
	PE	-0.008	-0.010	-0.011	-	-0.011^{**}	-0.009
	Std (ens)	0.003	0.000	0.000	0.011^{**}	-	0.002
	Std	0.000	-0.001^{**}	-0.002	0.009	-0.002	-
Ensemble	Margin	0.026^{**}	0.024^{*}	0.024^{*}	0.035^{***}	0.023^{**}	0.026^{*}
	Maxprob	0.027^{**}	0.025^{*}	0.024^{**}	0.036^{***}	0.024^{**}	0.027^{**}
	MI	-0.002	-0.004	-0.005	0.006	-0.005	-0.003
	PE	0.020	0.018	0.017	0.028^{**}	0.017	0.019
	Std (ens)	0.017	0.015	0.014	0.026^{*}	0.014	0.016
	Std	0.025^{*}	0.023	0.022^{*}	0.033^{***}	0.022^{*}	0.024^{*}

Table 4. Difference in quality (AURC) of estimators. In the second block of rows I compare performance of estimators for the same model (ensemble). In the fisrt block I compare uncertainty metrics, calculated via dropout, with the ones, calculated on ensemble. Stars denote significance level *-0.1, **-0.05, ***-0.001 (two-sided t-test)

Model	Metric	Margin	Maxprob	MI	PE	Std (ens)	Std
Dropout	Margin	-0.026^{**}	-0.027^{**}	0.002	-0.020	-0.017	-0.025^{*}
	Maxprob	-0.024^{*}	-0.025^{*}	0.004	-0.018	-0.015	-0.023
	MI	-0.024^{*}	-0.024^{**}	0.005	-0.017	-0.014	-0.022^{*}
	PE	-0.035^{***}	-0.036^{***}	-0.006	-0.028^{**}	-0.026^{*}	-0.033^{***}
	Std (ens)	-0.023^{**}	-0.024^{**}	0.005	-0.017	-0.014	-0.022^{*}
	Std	-0.026^{*}	-0.027^{**}	0.003	-0.019	-0.016	-0.024^{*}
Ensemble	Margin	-	-0.000	0.029^{***}	0.006^{**}	0.009^{**}	0.001
	Maxprob	0.000	-	0.030^{***}	0.007^{***}	0.010^{**}	0.002
	MI	-0.029^{***}	-0.030^{***}	-	-0.022^{***}	-0.020^{***}	-0.027^{***}
	PE	-0.006^{**}	-0.007^{***}	0.022^{***}	-	0.002	-0.004^{***}
	Std (ens)	-0.009^{**}	-0.010^{**}	0.020^{***}	-0.002	-	-0.007^{*}
	Std	-0.001	-0.002	0.027^{***}	0.004^{***}	0.007^{*}	-

References

Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., and Keogh, E. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Dempster, A., Petitjean, F., and Webb, G. I. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22): 2402–2410, 2016.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kendall, A., Badrinarayanan, V., and Cipolla, R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Loquercio, A., Segu, M., and Scaramuzza, D. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Smith, L. and Gal, Y. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Xiao, Y. and Wang, W. Y. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7322–7329, 2019.

7. Appendix

Table 5: ROC-AUC quality of uncertainty estimators calculated on model with dropout. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
Adiac	0.798	0.806	0.820	0.840	**0.842**	0.829	0.837
ArrowHead	0.754	0.754	**0.763**	0.754	0.760	0.754	0.755
Beef	0.800	0.722	0.715	0.722	**0.736**	0.715	0.715
BeetleFly	0.700	**0.405**	0.357	**0.405**	0.381	**0.405**	**0.405**
BirdChicken	0.900	0.639	**0.694**	0.639	**0.694**	0.639	0.639
CBF	0.960	0.952	**0.954**	0.952	0.953	0.952	0.952
Car	0.817	**0.822**	0.818	0.818	0.814	0.818	0.816
ChlorineConcentration	0.610	0.664	0.620	0.669	0.643	0.666	**0.670**
CinCECGTorso	0.712	0.745	**0.754**	0.747	0.751	0.746	0.746
Coffee	0.964	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Computers	0.772	0.701	**0.703**	0.701	0.701	0.701	0.701
CricketX	0.751	0.851	0.794	0.840	0.821	**0.853**	0.849
CricketY	0.746	0.796	0.789	0.785	**0.797**	0.793	0.791
CricketZ	0.769	0.866	0.832	0.871	0.851	0.868	**0.871**
DiatomSizeReduction	0.840	0.906	**0.915**	0.904	0.908	0.905	0.905
DistalPhalanxOutlineAgeGroup	0.727	0.671	0.678	0.673	**0.679**	0.675	0.674
DistalPhalanxOutlineCorrect	0.808	0.758	**0.763**	0.758	0.760	0.758	0.758
DistalPhalanxTW	0.676	0.784	**0.785**	0.763	0.778	0.776	0.773
ECG200	0.900	0.731	**0.736**	0.731	0.730	0.731	0.731
ECG5000	0.936	0.870	**0.881**	0.865	0.875	0.868	0.867
ECGFiveDays	0.958	0.957	**0.959**	0.957	0.958	0.957	0.957
Earthquakes	0.532	**0.614**	0.585	**0.614**	0.611	**0.614**	**0.614**
ElectricDevices	0.698	0.685	0.664	0.682	0.682	**0.691**	0.689
FaceAll	0.792	0.529	**0.692**	0.558	0.624	0.542	0.543
FaceFour	0.920	0.921	**0.954**	0.921	0.949	0.917	0.917
FacesUCR	0.916	0.912	**0.919**	0.916	0.918	0.915	0.917
FiftyWords	0.780	0.868	0.873	0.860	**0.882**	0.876	0.873
Fish	0.926	0.854	**0.915**	0.861	0.896	0.859	0.859
FordA	0.944	**0.897**	0.894	**0.897**	0.897	**0.897**	**0.897**
FordB	0.791	**0.779**	0.778	**0.779**	0.779	**0.779**	**0.779**
GunPoint	0.993	0.987	**0.993**	0.987	0.987	0.987	0.987
Ham	0.781	0.724	**0.727**	0.724	0.725	0.724	0.724
HandOutlines	0.949	0.728	**0.753**	0.728	0.748	0.728	0.728
Haptics	0.536	0.634	0.612	0.630	0.613	0.630	**0.635**
Herring	0.672	**0.649**	0.642	**0.649**	0.649	**0.649**	**0.649**
InlineSkate	0.402	**0.652**	0.629	0.642	0.636	0.651	0.649
InsectWingbeatSound	0.610	0.692	0.686	0.706	0.702	0.703	**0.706**
ItalyPowerDemand	0.948	0.864	**0.865**	0.863	0.864	0.863	0.863
LargeKitchenAppliances	0.864	0.801	**0.850**	0.802	0.841	0.803	0.802
Lightning2	0.689	0.712	**0.718**	0.708	0.714	0.708	0.709
Lightning7	0.767	0.694	**0.720**	0.703	0.709	0.701	0.702
Mallat	0.928	0.897	0.889	0.897	0.895	0.897	**0.898**
Meat	0.967	**0.948**	0.940	**0.948**	0.940	**0.948**	**0.948**
MedicalImages	0.701	0.793	0.727	**0.815**	0.776	0.807	0.814
MiddlePhalanxOutlineAgeGroup	0.552	0.544	0.529	0.549	0.544	**0.553**	0.549
MiddlePhalanxOutlineCorrect	0.811	0.767	0.768	0.767	**0.770**	0.767	0.767
MiddlePhalanxTW	0.539	0.800	0.815	0.806	**0.819**	0.807	0.811

Continued on next page

Table 5: ROC-AUC quality of uncertainty estimators calculated on model with dropout. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
MoteStrain	0.843	0.806	**0.813**	0.805	0.809	0.805	0.805
NonInvasiveFetalECGThorax1	0.952	0.916	0.923	0.923	**0.930**	0.925	0.926
NonInvasiveFetalECGThorax2	0.953	0.907	**0.946**	0.928	0.944	0.921	0.925
OSULeaf	0.909	0.918	**0.921**	0.914	0.921	0.919	0.916
OliveOil	0.900	0.765	0.765	0.765	**0.778**	0.765	0.765
PhalangesOutlinesCorrect	0.829	0.737	0.723	0.737	**0.739**	0.737	0.737
Phoneme	0.263	0.663	0.545	0.653	0.569	0.665	**0.666**
Plane	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
ProximalPhalanxOutlineAgeGroup	0.810	0.686	**0.697**	0.692	0.696	0.689	0.690
ProximalPhalanxOutlineCorrect	0.880	**0.859**	0.845	**0.859**	0.849	**0.859**	**0.859**
ProximalPhalanxTW	0.795	0.705	**0.712**	0.705	0.708	0.704	0.704
RefrigerationDevices	0.520	0.609	0.606	**0.629**	0.618	0.622	0.625
ScreenType	0.379	0.545	0.564	**0.591**	0.569	0.553	0.576
ShapeletSim	0.950	**0.970**	0.966	**0.970**	0.968	**0.970**	**0.970**
ShapesAll	0.858	0.882	0.887	**0.893**	0.893	0.888	0.891
SmallKitchenAppliances	0.808	**0.773**	0.747	0.765	0.769	0.770	0.769
SonyAIBORobotSurface1	0.952	0.944	**0.947**	0.944	0.946	0.944	0.944
SonyAIBORobotSurface2	0.834	0.761	**0.768**	0.761	0.765	0.761	0.761
StarLightCurves	0.967	0.788	**0.797**	0.785	0.793	0.785	0.785
Strawberry	0.868	**0.859**	0.837	**0.859**	0.848	**0.859**	**0.859**
SwedishLeaf	0.926	0.901	**0.934**	0.917	0.932	0.914	0.918
Symbols	0.971	0.710	**0.777**	0.707	0.744	0.709	0.709
SyntheticControl	0.970	0.987	0.985	0.989	0.986	0.988	**0.989**
ToeSegmentation1	0.917	0.898	**0.906**	0.898	0.903	0.898	0.898
ToeSegmentation2	0.877	0.669	**0.719**	0.669	0.696	0.669	0.669
Trace	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
TwoLeadECG	0.976	0.905	**0.918**	0.906	0.912	0.906	0.907
TwoPatterns	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
UWaveGestureLibraryAll	0.959	0.926	0.929	0.932	**0.933**	0.931	0.932
UWaveGestureLibraryX	0.817	0.837	0.824	0.841	0.837	0.842	**0.843**
UWaveGestureLibraryY	0.752	0.773	0.745	0.777	0.770	0.780	**0.782**
UWaveGestureLibraryZ	0.748	0.779	0.773	0.783	**0.791**	0.785	0.787
Wafer	0.945	0.745	**0.759**	0.745	0.755	0.745	0.745
Wine	0.889	0.688	**0.757**	0.688	0.736	0.688	0.688
WordSynonyms	0.658	0.824	0.801	0.833	0.816	0.834	**0.837**
Worms	0.753	0.661	0.685	**0.687**	0.685	0.670	0.673
WormsTwoClass	0.792	**0.625**	0.613	**0.625**	0.616	**0.625**	**0.625**
Yoga	0.844	0.728	**0.751**	0.728	0.743	0.728	0.728
ACSF1	0.830	0.792	0.768	0.799	0.775	**0.807**	0.806
AllGestureWiimoteX	0.623	0.785	0.783	**0.812**	0.800	0.800	0.808
AllGestureWiimoteY	0.700	**0.768**	0.719	0.744	0.734	0.762	0.758
AllGestureWiimoteZ	0.556	0.715	0.718	0.699	**0.726**	0.712	0.710
BME	0.987	**0.980**	0.976	**0.980**	**0.980**	**0.980**	**0.980**
Chinatown	0.977	0.822	**0.823**	0.822	0.822	0.822	0.822
Crop	0.719	0.846	0.807	0.844	0.828	0.853	**0.854**
DodgerLoopDay	0.595	0.766	0.746	0.741	**0.769**	0.767	0.757
DodgerLoopGame	0.853	0.706	**0.709**	0.706	0.707	0.706	0.706
DodgerLoopWeekend	0.919	0.556	0.588	0.607	0.561	0.607	**0.612**

Continued on next page

Table 5: ROC-AUC quality of uncertainty estimators calculated on model with dropout. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
EOGHorizontalSignal	0.544	0.748	0.759	0.745	**0.768**	0.745	0.744
EOGVerticalSignal	0.497	**0.767**	0.749	0.746	0.758	0.761	0.755
EthanolLevel	0.536	0.789	0.782	0.759	**0.798**	0.780	0.773
FreezerRegularTrain	0.990	0.931	**0.943**	0.931	0.940	0.931	0.931
FreezerSmallTrain	0.909	0.807	**0.831**	0.807	0.823	0.807	0.807
Fungi	0.532	0.593	**0.626**	0.585	0.607	0.587	0.587
GestureMidAirD1	0.723	0.751	0.778	0.763	**0.791**	0.758	0.762
GestureMidAirD2	0.669	0.731	0.665	0.721	0.705	**0.740**	0.737
GestureMidAirD3	0.462	0.595	0.630	0.575	**0.635**	0.601	0.596
GesturePebbleZ1	0.814	**0.808**	0.797	0.790	0.790	0.801	0.796
GesturePebbleZ2	0.734	0.782	0.788	0.782	**0.788**	0.783	0.785
GunPointAgeSpan	0.953	0.938	**0.940**	0.938	0.940	0.938	0.938
GunPointMaleVersusFemale	0.987	0.985	**0.986**	0.985	**0.986**	0.985	0.985
GunPointOldVersusYoung	0.994	**0.984**	0.974	**0.984**	0.979	**0.984**	**0.984**
HouseTwenty	0.899	0.847	**0.854**	0.847	0.850	0.847	0.847
InsectEPGRegularTrain	0.988	**0.997**	0.982	0.996	0.991	**0.997**	0.996
InsectEPGSmallTrain	0.924	0.953	**0.957**	0.953	0.956	0.934	0.953
MelbournePedestrian	0.883	0.903	0.910	0.907	**0.913**	0.908	0.909
MixedShapesRegularTrain	0.945	0.918	**0.939**	0.920	0.936	0.920	0.920
MixedShapesSmallTrain	0.908	0.901	**0.922**	0.902	0.916	0.902	0.901
PLAID	0.762	0.854	0.819	0.853	0.847	0.859	**0.860**
PickupGestureWiimoteZ	0.660	**0.813**	0.718	0.772	0.743	0.795	0.791
PigAirwayPressure	0.144	0.674	0.490	0.746	0.572	0.731	**0.772**
PigArtPressure	0.813	0.822	**0.870**	0.830	0.863	0.829	0.831
PigCVP	0.683	0.580	**0.607**	0.599	0.605	0.594	0.595
PowerCons	0.961	0.903	**0.904**	0.903	0.903	0.903	0.903
Rock	0.660	0.768	**0.784**	0.763	0.777	0.763	0.761
SemgHandGenderCh2	0.872	**0.855**	0.836	**0.855**	0.845	**0.855**	**0.855**
SemgHandMovementCh2	0.584	0.771	0.580	0.754	0.661	0.775	**0.775**
SemgHandSubjectCh2	0.813	**0.895**	0.769	0.861	0.819	0.890	0.881
ShakeGestureWiimoteZ	0.720	0.815	0.817	0.819	0.817	0.812	**0.819**
SmoothSubspace	0.940	0.828	**0.850**	0.832	0.842	0.832	0.832
UMD	0.972	0.825	0.820	**0.830**	0.827	0.829	**0.830**

Table 6: AURC quality of uncertainty estimators calculated on model with dropout. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
Adiac	0.798	0.692	0.722	0.740	**0.749**	0.726	0.736
ArrowHead	0.754	0.599	**0.615**	0.598	0.610	0.599	0.599
Beef	0.800	0.532	0.528	0.535	**0.553**	0.524	0.524
BeetleFly	0.700	**-0.214**	-0.297	**-0.214**	-0.252	**-0.214**	**-0.214**
BirdChicken	0.900	0.553	**0.646**	0.553	**0.646**	0.553	0.553
CBF	0.960	0.950	**0.952**	0.950	0.951	0.950	0.950
Car	0.817	**0.772**	0.769	0.770	0.765	0.768	0.766
ChlorineConcentration	0.610	0.451	0.357	0.447	0.411	0.452	**0.455**
CinCECGTorso	0.712	0.618	**0.641**	0.623	0.633	0.620	0.620
Coffee	0.964	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Computers	0.772	0.543	**0.570**	0.543	0.559	0.543	0.543
CricketX	0.751	0.773	0.714	0.759	0.743	**0.775**	0.771
CricketY	0.746	0.707	0.704	0.692	**0.713**	0.702	0.700
CricketZ	0.769	0.779	0.771	0.783	0.775	0.782	**0.785**
DiatomSizeReduction	0.840	0.898	**0.910**	0.896	0.902	0.897	0.897
DistalPhalanxOutlineAgeGroup	0.727	0.452	**0.495**	0.459	0.482	0.459	0.458
DistalPhalanxOutlineCorrect	0.808	0.590	**0.603**	0.590	0.598	0.590	0.590
DistalPhalanxTW	0.676	0.587	**0.595**	0.563	0.579	0.578	0.574
ECG200	0.900	0.530	0.529	0.530	**0.531**	0.530	0.530
ECG5000	0.936	0.837	**0.853**	0.830	0.844	0.834	0.834
ECGFiveDays	0.958	0.956	**0.958**	0.956	0.957	0.956	0.956
Earthquakes	0.532	**0.421**	0.310	**0.421**	0.375	**0.421**	**0.421**
ElectricDevices	0.698	0.313	0.305	0.325	**0.327**	0.326	0.324
FaceAll	0.792	-0.210	**0.451**	-0.092	0.192	-0.165	-0.164
FaceFour	0.920	0.922	**0.960**	0.921	0.953	0.919	0.919
FacesUCR	0.916	0.888	**0.899**	0.893	0.897	0.892	0.894
FiftyWords	0.780	0.821	0.837	0.815	**0.845**	0.831	0.828
Fish	0.926	0.801	**0.903**	0.811	0.873	0.808	0.808
FordA	0.944	0.867	0.865	0.867	**0.868**	0.867	0.867
FordB	0.791	0.704	**0.707**	0.704	0.706	0.704	0.704
GunPoint	0.993	0.993	**1.000**	0.993	0.993	0.993	0.993
Ham	0.781	0.436	**0.464**	0.436	0.452	0.436	0.436
HandOutlines	0.949	0.607	**0.665**	0.607	0.655	0.607	0.607
Haptics	0.536	**0.226**	0.217	0.220	0.212	0.219	0.222
Herring	0.672	0.348	0.388	0.348	**0.395**	0.348	0.348
InlineSkate	0.402	**0.361**	0.351	0.351	0.356	0.359	0.356
InsectWingbeatSound	0.610	0.502	0.492	**0.524**	0.519	0.519	0.523
ItalyPowerDemand	0.948	0.816	**0.822**	0.814	0.817	0.814	0.815
LargeKitchenAppliances	0.864	0.673	**0.770**	0.674	0.746	0.674	0.673
Lightning2	0.689	0.536	**0.562**	0.495	0.542	0.495	0.512
Lightning7	0.767	0.547	**0.588**	0.560	0.577	0.557	0.558
Mallat	0.928	0.876	0.864	0.875	0.871	0.876	**0.876**
Meat	0.967	**0.952**	**0.952**	**0.952**	**0.952**	**0.952**	**0.952**
MedicalImages	0.701	0.685	0.559	**0.718**	0.647	0.704	0.711
MiddlePhalanxOutlineAgeGroup	0.552	0.065	0.087	0.087	**0.093**	0.084	0.081
MiddlePhalanxOutlineCorrect	0.811	0.639	**0.660**	0.639	0.653	0.639	0.639
MiddlePhalanxTW	0.539	0.740	0.755	0.745	**0.761**	0.746	0.752

Continued on next page

Table 6: AURC quality of uncertainty estimators calculated on model with dropout. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
MoteStrain	0.843	0.722	**0.742**	0.703	0.730	0.703	0.705
NonInvasiveFetalECGThorax1	0.952	0.828	0.861	0.839	**0.865**	0.840	0.841
NonInvasiveFetalECGThorax2	0.953	0.884	**0.941**	0.909	0.936	0.902	0.906
OSULeaf	0.909	0.907	**0.918**	0.906	0.916	0.909	0.906
OliveOil	0.900	**0.668**	0.634	**0.668**	**0.668**	**0.668**	**0.668**
PhalangesOutlinesCorrect	0.829	0.530	0.536	0.530	**0.550**	0.530	0.530
Phoneme	0.263	0.296	0.143	0.292	0.194	**0.302**	0.302
Plane	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
ProximalPhalanxOutlineAgeGroup	0.810	0.442	0.433	**0.450**	0.442	0.446	0.446
ProximalPhalanxOutlineCorrect	0.880	**0.820**	0.801	**0.820**	0.807	**0.820**	**0.820**
ProximalPhalanxTW	0.795	0.563	**0.579**	0.564	0.572	0.563	0.563
RefrigerationDevices	0.520	0.194	0.194	**0.220**	0.205	0.209	0.212
ScreenType	0.379	0.119	**0.182**	0.163	0.181	0.125	0.146
ShapeletSim	0.950	**0.974**	0.970	**0.974**	0.972	**0.974**	**0.974**
ShapesAll	0.858	0.850	0.869	0.871	**0.876**	0.863	0.866
SmallKitchenAppliances	0.808	0.650	0.627	0.641	**0.652**	0.647	0.645
SonyAIBORobotSurface1	0.952	0.939	**0.943**	0.939	0.941	0.939	0.939
SonyAIBORobotSurface2	0.834	0.658	**0.673**	0.658	0.667	0.658	0.658
StarLightCurves	0.967	0.674	0.673	0.649	**0.683**	0.649	0.649
Strawberry	0.868	**0.818**	0.785	**0.818**	0.802	**0.818**	**0.818**
SwedishLeaf	0.926	0.890	**0.930**	0.908	0.927	0.905	0.909
Symbols	0.971	0.521	**0.658**	0.515	0.595	0.520	0.520
SyntheticControl	0.970	0.990	0.989	0.991	0.989	0.990	**0.991**
ToeSegmentation1	0.917	0.867	**0.878**	0.867	0.874	0.867	0.867
ToeSegmentation2	0.877	0.310	**0.507**	0.310	0.432	0.310	0.310
Trace	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
TwoLeadECG	0.976	0.791	**0.850**	0.793	0.809	0.793	0.797
TwoPatterns	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
UWaveGestureLibraryAll	0.959	0.918	0.922	0.924	**0.925**	0.923	0.924
UWaveGestureLibraryX	0.817	0.792	0.778	0.797	0.794	0.798	**0.799**
UWaveGestureLibraryY	0.752	0.669	0.623	0.667	0.662	0.674	**0.676**
UWaveGestureLibraryZ	0.748	0.693	0.685	0.699	**0.710**	0.701	0.702
Wafer	0.945	0.538	**0.566**	0.538	0.557	0.538	0.538
Wine	0.889	0.594	**0.676**	0.594	0.653	0.594	0.594
WordSynonyms	0.658	0.757	0.737	0.771	0.755	0.771	**0.774**
Worms	0.753	0.466	**0.508**	0.489	0.498	0.469	0.474
WormsTwoClass	0.792	**0.325**	0.287	**0.325**	0.321	**0.325**	**0.325**
Yoga	0.844	0.469	**0.519**	0.470	0.499	0.470	0.470
ACSF1	0.830	0.539	0.567	0.554	**0.573**	0.558	0.557
AllGestureWiimoteX	0.623	0.653	0.677	0.690	**0.690**	0.672	0.682
AllGestureWiimoteY	0.700	**0.642**	0.583	0.609	0.604	0.634	0.630
AllGestureWiimoteZ	0.556	0.527	0.536	0.506	**0.548**	0.524	0.520
BME	0.987	**0.986**	0.982	**0.986**	**0.986**	**0.986**	**0.986**
Chinatown	0.977	0.786	**0.788**	0.786	0.787	0.786	0.786
Crop	0.719	0.796	0.749	0.796	0.778	0.806	**0.807**
DodgerLoopDay	0.595	0.677	0.652	0.654	0.679	**0.683**	0.672
DodgerLoopGame	0.853	0.458	**0.504**	0.492	0.482	0.492	0.497
DodgerLoopWeekend	0.919	-0.479	0.033	0.355	-0.473	0.355	**0.387**

Continued on next page

Table 6: AURC quality of uncertainty estimators calculated on model with dropout. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
EOGHorizontalSignal	0.544	0.563	0.583	0.563	**0.592**	0.561	0.560
EOGVerticalSignal	0.497	**0.648**	0.621	0.629	0.638	0.642	0.636
EthanolLevel	0.536	0.675	0.690	0.636	**0.712**	0.664	0.655
FreezerRegularTrain	0.990	0.867	**0.910**	0.867	0.903	0.867	0.867
FreezerSmallTrain	0.909	0.723	**0.774**	0.741	0.749	0.741	0.743
Fungi	0.532	0.318	**0.393**	0.286	0.345	0.299	0.300
GestureMidAirD1	0.723	0.649	0.698	0.681	**0.721**	0.673	0.678
GestureMidAirD2	0.669	0.533	0.446	0.531	0.509	**0.553**	0.551
GestureMidAirD3	0.462	0.248	0.282	0.215	**0.302**	0.247	0.243
GesturePebbleZ1	0.814	**0.751**	0.741	0.737	0.737	0.745	0.740
GesturePebbleZ2	0.734	0.660	**0.680**	0.662	0.675	0.664	0.665
GunPointAgeSpan	0.953	0.934	**0.937**	0.934	0.937	0.934	0.934
GunPointMaleVersusFemale	0.987	0.987	**0.987**	0.987	**0.987**	0.987	0.987
GunPointOldVersusYoung	0.994	**0.987**	0.977	**0.987**	0.982	**0.987**	**0.987**
HouseTwenty	0.899	0.823	**0.832**	0.823	0.828	0.823	0.823
InsectEPGRegularTrain	0.988	**1.000**	0.986	**1.000**	0.993	**1.000**	**1.000**
InsectEPGSmallTrain	0.924	0.952	**0.957**	0.952	0.955	0.916	0.952
MelbournePedestrian	0.883	0.884	0.897	0.889	**0.899**	0.890	0.891
MixedShapesRegularTrain	0.945	0.882	**0.931**	0.886	0.924	0.884	0.885
MixedShapesSmallTrain	0.908	0.859	**0.907**	0.862	0.895	0.861	0.861
PLAID	0.762	0.807	0.762	0.803	0.797	0.812	**0.813**
PickupGestureWiimoteZ	0.660	**0.758**	0.647	0.703	0.684	0.734	0.730
PigAirwayPressure	0.144	0.462	0.122	0.522	0.306	0.522	**0.574**
PigArtPressure	0.813	0.586	**0.799**	0.609	0.725	0.609	0.611
PigCVP	0.683	-0.171	**-0.119**	-0.140	-0.126	-0.149	-0.147
PowerCons	0.961	0.883	**0.885**	0.883	0.884	0.883	0.883
Rock	0.660	0.727	**0.745**	0.722	0.736	0.722	0.720
SemgHandGenderCh2	0.872	**0.821**	0.795	**0.821**	0.808	**0.821**	**0.821**
SemgHandMovementCh2	0.584	0.668	0.312	0.648	0.467	0.676	**0.677**
SemgHandSubjectCh2	0.813	**0.874**	0.715	0.832	0.780	0.867	0.857
ShakeGestureWiimoteZ	0.720	0.727	**0.733**	0.731	0.732	0.723	0.731
SmoothSubspace	0.940	0.796	**0.827**	0.801	0.816	0.801	0.801
UMD	0.972	0.719	0.708	**0.723**	0.716	0.717	**0.723**

Table 7: ROC-AUC quality of uncertainty estimators calculated on ensemble. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
Adiac	0.795	0.824	0.837	**0.851**	0.848	0.845	0.850
ArrowHead	0.794	0.736	0.728	0.735	0.732	**0.737**	0.736
Beef	0.800	0.708	0.660	**0.757**	0.694	0.708	0.722
BeetleFly	0.950	**0.526**	0.263	**0.526**	0.421	**0.526**	**0.526**
BirdChicken	0.850	**0.922**	0.882	**0.922**	0.902	**0.922**	**0.922**
CBF	0.960	**0.958**	0.951	0.958	0.951	0.958	0.958
Car	0.783	0.908	**0.936**	0.923	0.931	0.915	0.915
ChlorineConcentration	0.624	0.689	0.630	0.694	0.663	0.693	**0.697**
CinCECGTorso	0.717	0.788	0.804	0.785	**0.804**	0.789	0.787
Coffee	0.964	**1.000**	0.963	**1.000**	0.963	**1.000**	**1.000**
Computers	0.776	**0.683**	0.640	**0.683**	0.663	**0.683**	**0.683**
CricketX	0.741	0.859	0.834	0.855	0.853	**0.863**	0.862
CricketY	0.749	0.789	0.771	0.794	0.790	0.794	**0.799**
CricketZ	0.790	0.853	0.827	0.858	0.849	0.857	**0.860**
DiatomSizeReduction	0.843	0.841	0.829	**0.844**	0.829	0.842	0.843
DistalPhalanxOutlineAgeGroup	0.712	0.746	**0.762**	0.748	0.761	0.749	0.751
DistalPhalanxOutlineCorrect	0.786	0.779	0.778	0.779	**0.783**	0.779	0.779
DistalPhalanxTW	0.676	**0.786**	0.774	0.771	0.779	0.781	0.780
ECG200	0.900	0.750	0.751	0.750	**0.756**	0.750	0.750
ECG5000	0.942	0.868	**0.887**	0.862	0.879	0.865	0.865
ECGFiveDays	0.929	0.969	**0.977**	0.969	0.976	0.969	0.969
Earthquakes	0.561	**0.617**	0.510	**0.617**	0.565	**0.617**	**0.617**
ElectricDevices	0.693	0.699	0.646	0.693	0.672	**0.703**	0.701
FaceAll	0.799	0.658	**0.848**	0.700	0.786	0.679	0.680
FaceFour	0.943	0.940	**0.973**	0.940	0.966	0.940	0.942
FacesUCR	0.922	0.926	0.933	0.932	**0.933**	0.930	0.932
FiftyWords	0.785	0.889	0.874	0.884	0.888	0.894	**0.894**
Fish	0.949	0.935	**0.966**	0.939	0.959	0.938	0.939
FordA	0.942	**0.904**	0.888	**0.904**	0.896	**0.904**	**0.904**
FordB	0.793	**0.792**	0.785	**0.792**	0.791	**0.792**	**0.792**
GunPoint	0.987	**0.986**	**0.986**	**0.986**	0.980	**0.986**	**0.986**
Ham	0.810	0.700	0.707	0.700	**0.716**	0.700	0.700
HandOutlines	0.938	**0.792**	0.781	**0.792**	0.791	**0.792**	**0.792**
Haptics	0.526	0.637	0.620	**0.648**	0.635	0.641	0.645
Herring	0.672	0.642	0.640	0.642	**0.646**	0.642	0.642
InlineSkate	0.396	**0.688**	0.664	0.678	0.671	0.687	0.686
InsectWingbeatSound	0.614	0.694	0.689	0.697	**0.708**	0.700	0.700
ItalyPowerDemand	0.947	0.875	**0.879**	0.875	0.877	0.875	0.875
LargeKitchenAppliances	0.885	0.769	0.784	0.764	**0.785**	0.769	0.767
Lightning2	0.705	0.703	**0.707**	0.703	0.705	0.703	0.703
Lightning7	0.808	**0.613**	0.587	0.610	0.575	0.610	0.610
Mallat	0.926	0.899	0.864	**0.905**	0.892	0.903	0.904
Meat	0.983	**0.932**	0.915	**0.932**	**0.932**	**0.932**	**0.932**
MedicalImages	0.716	0.795	0.755	**0.812**	0.782	0.807	0.811
MiddlePhalanxOutlineAgeGroup	0.552	0.516	0.464	**0.530**	0.498	0.526	0.525
MiddlePhalanxOutlineCorrect	0.814	**0.753**	0.728	**0.753**	0.740	**0.753**	**0.753**
MiddlePhalanxTW	0.558	0.795	0.827	0.769	**0.829**	0.792	0.785

Continued on next page

Table 7: ROC-AUC quality of uncertainty estimators calculated on ensemble. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
MoteStrain	0.828	**0.762**	0.606	0.762	0.627	0.762	0.762
NonInvasiveFetalECGThorax1	0.955	0.917	0.915	0.922	0.922	0.925	**0.926**
NonInvasiveFetalECGThorax2	0.957	0.904	0.936	0.925	**0.941**	0.918	0.922
OSULeaf	0.897	0.933	0.932	0.936	**0.938**	0.938	0.937
OliveOil	0.933	**0.643**	0.625	**0.643**	0.625	**0.643**	**0.643**
PhalangesOutlinesCorrect	0.830	**0.744**	0.709	**0.744**	0.739	**0.744**	**0.744**
Phoneme	0.262	0.663	0.617	0.661	0.634	0.667	**0.670**
Plane	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
ProximalPhalanxOutlineAgeGroup	0.824	0.709	0.685	**0.720**	0.713	0.716	0.716
ProximalPhalanxOutlineCorrect	0.893	**0.861**	0.779	**0.861**	0.836	**0.861**	**0.861**
ProximalPhalanxTW	0.800	0.687	0.682	**0.689**	0.681	0.687	0.688
RefrigerationDevices	0.531	0.590	0.572	**0.594**	0.591	0.593	0.593
ScreenType	0.397	0.553	0.542	**0.571**	0.544	0.564	0.568
ShapeletSim	0.983	**0.960**	0.928	**0.960**	0.947	**0.960**	**0.960**
ShapesAll	0.865	0.895	0.901	0.900	**0.908**	0.900	0.903
SmallKitchenAppliances	0.813	0.777	0.767	0.765	**0.783**	0.773	0.770
SonyAIBORobotSurface1	0.952	0.949	0.947	**0.949**	0.946	**0.949**	**0.949**
SonyAIBORobotSurface2	0.870	0.853	**0.864**	0.853	0.861	0.853	0.853
StarLightCurves	0.971	0.907	**0.923**	0.905	0.921	0.906	0.906
Strawberry	0.949	**0.861**	0.792	**0.861**	0.807	**0.861**	**0.861**
SwedishLeaf	0.936	0.913	0.933	0.919	**0.936**	0.919	0.920
Symbols	0.972	0.865	**0.876**	0.867	0.872	0.866	0.866
SyntheticControl	0.990	**0.988**	0.984	0.983	**0.988**	**0.988**	0.987
ToeSegmentation1	0.934	**0.881**	0.860	**0.881**	0.868	**0.881**	**0.881**
ToeSegmentation2	0.854	0.874	**0.887**	0.874	0.883	0.874	0.874
Trace	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
TwoLeadECG	0.971	0.948	**0.953**	0.949	0.952	0.949	0.949
TwoPatterns	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
UWaveGestureLibraryAll	0.962	0.927	0.926	**0.934**	0.931	0.932	0.934
UWaveGestureLibraryX	0.831	0.835	0.832	0.842	0.843	0.841	**0.843**
UWaveGestureLibraryY	0.752	0.784	0.766	0.790	0.783	0.792	**0.794**
UWaveGestureLibraryZ	0.762	0.781	0.767	0.789	0.787	0.789	**0.791**
Wafer	0.972	0.779	0.734	0.779	0.771	**0.779**	**0.779**
Wine	0.944	0.386	**0.405**	0.386	0.392	0.386	0.386
WordSynonyms	0.674	0.829	0.748	0.833	0.776	0.836	**0.837**
Worms	0.766	0.697	0.676	0.698	0.688	0.698	**0.702**
WormsTwoClass	0.792	**0.597**	0.577	**0.597**	0.584	**0.597**	**0.597**
Yoga	0.847	0.744	0.730	0.744	**0.746**	0.744	0.744
ACSF1	0.830	0.825	0.771	0.817	0.795	**0.830**	0.821
AllGestureWiimoteX	0.631	0.782	0.778	**0.798**	0.793	0.791	0.796
AllGestureWiimoteY	0.698	**0.789**	0.724	0.766	0.754	0.782	0.778
AllGestureWiimoteZ	0.567	**0.712**	0.672	0.701	0.686	0.710	0.708
BME	0.980	0.939	**0.957**	0.937	0.955	0.937	0.937
Chinatown	0.962	0.863	**0.864**	0.863	0.862	0.863	0.863
Crop	0.721	0.847	0.832	0.848	0.846	0.856	**0.857**
DodgerLoopDay	0.595	0.767	0.682	0.718	0.721	**0.773**	0.754
DodgerLoopGame	0.860	**0.746**	0.739	**0.746**	0.737	**0.746**	**0.746**
DodgerLoopWeekend	0.949	0.503	**0.508**	0.503	0.505	0.503	0.501

Continued on next page

Table 7: ROC-AUC quality of uncertainty estimators calculated on ensemble. Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
EOGHorizontalSignal	0.550	**0.743**	0.735	0.735	0.741	0.738	0.738
EOGVerticalSignal	0.500	**0.745**	0.713	0.744	0.724	0.743	0.744
EthanolLevel	0.546	**0.778**	0.731	0.754	0.764	0.772	0.766
FreezerRegularTrain	0.996	**0.977**	0.972	**0.977**	0.974	**0.977**	**0.977**
FreezerSmallTrain	0.927	0.908	**0.911**	0.908	0.910	0.908	0.908
Fungi	0.780	0.835	0.945	0.913	**0.955**	0.885	0.901
GestureMidAirD1	0.723	0.766	0.676	0.780	0.718	0.775	**0.780**
GestureMidAirD2	0.669	0.763	0.615	0.738	0.678	**0.766**	0.761
GestureMidAirD3	0.415	0.670	0.680	0.687	0.692	0.688	**0.699**
GesturePebbleZ1	0.814	0.817	0.819	0.807	**0.824**	0.815	0.810
GesturePebbleZ2	0.728	0.775	0.780	0.781	**0.791**	0.777	0.780
GunPointAgeSpan	0.946	0.956	**0.971**	0.956	0.966	0.956	0.956
GunPointMaleVersusFemale	0.987	0.967	0.974	0.967	**0.974**	0.967	0.967
GunPointOldVersusYoung	0.987	**0.997**	0.977	**0.997**	0.990	**0.997**	**0.997**
HouseTwenty	0.891	0.919	**0.922**	0.919	0.920	0.919	0.919
InsectEPGRegularTrain	0.992	**1.000**	0.966	0.998	0.974	0.998	0.998
InsectEPGSmallTrain	0.920	0.917	0.871	**0.927**	0.877	0.922	0.927
MelbournePedestrian	0.889	0.913	0.884	0.919	0.898	0.917	**0.919**
MixedShapesRegularTrain	0.946	0.939	**0.949**	0.942	0.948	0.941	0.942
MixedShapesSmallTrain	0.906	**0.918**	0.903	0.914	0.905	0.917	0.916
PLAID	0.795	0.884	0.849	0.884	0.878	0.887	**0.887**
PickupGestureWiimoteZ	0.680	**0.884**	0.798	0.868	0.820	0.875	0.875
PigAirwayPressure	0.130	0.668	0.609	**0.807**	0.649	0.735	0.798
PigArtPressure	0.808	0.857	0.860	**0.882**	0.878	0.870	0.874
PigCVP	0.688	0.578	0.565	**0.599**	0.578	0.595	0.597
PowerCons	0.967	0.914	0.912	0.914	**0.923**	0.914	0.914
Rock	0.660	0.756	0.763	**0.775**	0.772	0.763	0.768
SemgHandGenderCh2	0.897	**0.834**	0.817	**0.834**	0.832	**0.834**	**0.834**
SemgHandMovementCh2	0.622	0.791	0.672	0.774	0.707	**0.793**	0.791
SemgHandSubjectCh2	0.831	**0.904**	0.835	0.880	0.866	0.901	0.894
ShakeGestureWiimoteZ	0.800	0.808	0.755	**0.812**	0.760	**0.812**	0.810
SmoothSubspace	0.947	0.858	**0.882**	0.862	0.875	0.861	0.861
UMD	0.972	0.746	0.746	**0.756**	0.745	0.750	0.753

Table 8: AURC quality of uncertainty estimators calculated on ensemble.
Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
Adiac	0.795	0.709	0.728	**0.747**	0.742	0.740	0.747
ArrowHead	0.794	0.581	0.555	0.580	0.562	**0.583**	0.582
Beef	0.800	0.638	0.578	**0.697**	0.631	0.644	0.658
BeetleFly	0.950	**0.353**	-0.385	**0.353**	0.112	**0.353**	**0.353**
BirdChicken	0.850	**0.941**	0.888	**0.941**	0.912	**0.941**	**0.941**
CBF	0.960	**0.957**	0.949	0.957	0.950	0.957	0.957
Car	0.783	0.913	**0.933**	0.930	0.933	0.921	0.921
ChlorineConcentration	0.624	0.502	0.357	0.501	0.440	0.505	**0.510**
CinCECGTorso	0.717	0.637	**0.664**	0.633	0.662	0.638	0.636
Coffee	0.964	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Computers	0.776	**0.512**	0.453	**0.512**	0.487	**0.512**	**0.512**
CricketX	0.741	**0.782**	0.751	0.769	0.774	0.780	0.779
CricketY	0.749	0.696	0.675	0.698	0.700	0.702	**0.707**
CricketZ	0.790	0.758	0.729	0.762	0.751	0.763	**0.765**
DiatomSizeReduction	0.843	0.803	0.791	**0.807**	0.790	0.804	0.805
DistalPhalanxOutlineAgeGroup	0.712	0.623	**0.677**	0.623	0.666	0.627	0.628
DistalPhalanxOutlineCorrect	0.786	0.634	0.632	0.634	**0.640**	0.634	0.634
DistalPhalanxTW	0.676	**0.647**	0.642	0.630	0.646	0.641	0.639
ECG200	0.900	0.540	0.555	0.540	**0.560**	0.540	0.540
ECG5000	0.942	0.833	**0.855**	0.826	0.846	0.830	0.830
ECGFiveDays	0.929	0.968	**0.977**	0.968	0.975	0.968	0.968
Earthquakes	0.561	**0.395**	0.073	**0.395**	0.281	**0.395**	**0.395**
ElectricDevices	0.693	0.332	0.274	0.341	0.315	**0.343**	0.341
FaceAll	0.799	0.359	**0.800**	0.463	0.678	0.406	0.408
FaceFour	0.943	0.942	**0.981**	0.943	0.972	0.942	0.945
FacesUCR	0.922	0.905	0.914	0.912	**0.914**	0.910	0.912
FiftyWords	0.785	0.848	0.835	0.842	0.850	**0.854**	0.854
Fish	0.949	0.931	**0.968**	0.935	0.959	0.934	0.935
FordA	0.942	**0.879**	0.860	**0.879**	0.870	**0.879**	**0.879**
FordB	0.793	**0.724**	0.708	**0.724**	0.720	**0.724**	**0.724**
GunPoint	0.987	**0.993**	0.989	**0.993**	0.982	**0.993**	**0.993**
Ham	0.810	0.438	0.470	0.438	**0.476**	0.438	0.438
HandOutlines	0.938	0.692	0.687	0.692	**0.701**	0.692	0.692
Haptics	0.526	0.262	0.216	**0.273**	0.253	0.266	0.269
Herring	0.672	0.434	0.434	0.434	**0.438**	0.434	0.434
InlineSkate	0.396	**0.433**	0.414	0.425	0.426	0.433	0.431
InsectWingbeatSound	0.614	0.500	0.487	0.508	**0.519**	0.509	0.510
ItalyPowerDemand	0.947	0.835	**0.840**	0.835	0.839	0.835	0.835
LargeKitchenAppliances	0.885	0.608	**0.650**	0.604	0.645	0.610	0.608
Lightning2	0.705	**0.545**	0.542	**0.545**	0.540	**0.545**	**0.545**
Lightning7	0.808	0.333	0.326	**0.333**	0.290	0.332	0.332
Mallat	0.926	0.879	0.819	**0.886**	0.863	0.884	0.885
Meat	0.983	**0.945**	0.926	**0.945**	**0.945**	**0.945**	**0.945**
MedicalImages	0.716	0.701	0.626	**0.728**	0.676	0.718	0.722
MiddlePhalanxOutlineAgeGroup	0.552	0.076	0.002	**0.107**	0.072	0.091	0.092
MiddlePhalanxOutlineCorrect	0.814	**0.605**	0.575	**0.605**	0.593	**0.605**	**0.605**
MiddlePhalanxTW	0.558	0.729	0.766	0.704	**0.769**	0.725	0.720

Continued on next page

Table 8: AURC quality of uncertainty estimators calculated on ensemble.
Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
MoteStrain	0.828	**0.568**	0.352	0.568	0.379	0.568	**0.568**
NonInvasiveFetalECGThorax1	0.955	0.807	**0.863**	0.818	0.851	0.818	0.819
NonInvasiveFetalECGThorax2	0.957	0.875	0.930	0.900	**0.933**	0.892	0.896
OSULeaf	0.897	0.923	0.925	0.928	**0.931**	0.928	0.928
OliveOil	0.933	**0.474**	0.424	**0.474**	0.424	**0.474**	**0.474**
PhalangesOutlinesCorrect	0.830	0.553	0.517	0.553	**0.559**	0.553	0.553
Phoneme	0.262	0.307	0.262	0.311	0.287	0.315	**0.317**
Plane	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
ProximalPhalanxOutlineAgeGroup	0.824	0.523	0.514	**0.547**	0.546	0.532	0.532
ProximalPhalanxOutlineCorrect	0.893	**0.817**	0.716	**0.817**	0.795	**0.817**	**0.817**
ProximalPhalanxTW	0.800	0.521	0.494	**0.526**	0.509	0.522	0.523
RefrigerationDevices	0.531	0.174	0.152	**0.184**	0.176	0.179	0.179
ScreenType	0.397	0.093	0.116	**0.125**	0.112	0.109	0.114
ShapeletSim	0.983	**0.964**	0.930	**0.964**	0.950	**0.964**	**0.964**
ShapesAll	0.865	0.849	0.884	0.865	**0.894**	0.860	0.862
SmallKitchenAppliances	0.813	0.663	0.677	0.654	**0.692**	0.661	0.658
SonyAIBORobotSurface1	0.952	**0.945**	0.943	**0.945**	0.942	**0.945**	**0.945**
SonyAIBORobotSurface2	0.870	0.813	**0.825**	0.813	0.822	0.813	0.813
StarLightCurves	0.971	0.880	**0.905**	0.878	0.900	0.879	0.879
Strawberry	0.949	**0.764**	0.679	**0.764**	0.702	**0.764**	**0.764**
SwedishLeaf	0.936	0.902	0.929	0.909	**0.930**	0.909	0.910
Symbols	0.972	0.794	**0.824**	0.798	0.814	0.795	0.796
SyntheticControl	0.990	0.991	0.987	0.986	**0.991**	0.991	0.990
ToeSegmentation1	0.934	**0.843**	0.820	**0.843**	0.828	**0.843**	**0.843**
ToeSegmentation2	0.854	0.848	**0.862**	0.848	0.859	0.848	0.848
Trace	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
TwoLeadECG	0.971	0.867	0.866	0.868	0.863	0.868	**0.874**
TwoPatterns	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
UWaveGestureLibraryAll	0.962	0.917	0.917	**0.925**	0.923	0.923	0.925
UWaveGestureLibraryX	0.831	0.791	0.784	0.799	**0.801**	0.798	0.800
UWaveGestureLibraryY	0.752	0.678	0.661	0.679	0.681	0.685	**0.687**
UWaveGestureLibraryZ	0.762	0.696	0.670	0.708	0.704	0.708	**0.711**
Wafer	0.972	**0.551**	0.450	**0.551**	0.526	**0.551**	**0.551**
Wine	0.944	0.036	**0.091**	0.036	0.056	0.036	0.036
WordSynonyms	0.674	0.761	0.660	0.767	0.700	0.769	**0.770**
Worms	0.766	0.552	0.489	0.536	0.515	0.548	**0.553**
WormsTwoClass	0.792	0.275	**0.295**	0.275	0.271	0.275	0.275
Yoga	0.847	0.527	0.527	0.527	**0.541**	0.527	0.527
ACSF1	0.830	0.719	0.689	0.724	0.707	**0.733**	0.723
AllGestureWiimoteX	0.631	0.643	0.650	0.668	**0.668**	0.657	0.663
AllGestureWiimoteY	0.698	**0.683**	0.588	0.655	0.636	0.674	0.670
AllGestureWiimoteZ	0.567	0.521	0.487	0.513	0.508	**0.523**	0.520
BME	0.980	0.942	**0.959**	0.940	0.956	0.940	0.940
Chinatown	0.962	**0.824**	0.823	**0.824**	0.821	**0.824**	**0.824**
Crop	0.721	0.799	0.780	0.802	0.800	0.810	**0.812**
DodgerLoopDay	0.595	0.671	0.549	0.622	0.617	**0.675**	0.655
DodgerLoopGame	0.860	**0.473**	0.467	**0.473**	0.464	**0.473**	**0.473**
DodgerLoopWeekend	0.949	-0.322	**-0.315**	-0.322	-0.320	-0.322	-0.322

Continued on next page

Table 8: AURC quality of uncertainty estimators calculated on ensemble.
Top-1 metrics for each dataset are in bold

	Accuracy	PE	MI	Margin	Std (ens)	Std	Maxprob
EOGHorizontalSignal	0.550	0.543	**0.571**	0.532	0.564	0.538	0.538
EOGVerticalSignal	0.500	**0.623**	0.579	0.622	0.601	0.621	0.623
EthanolLevel	0.546	0.648	0.579	0.621	**0.650**	0.644	0.636
FreezerRegularTrain	0.996	**0.977**	0.972	**0.977**	0.973	**0.977**	**0.977**
FreezerSmallTrain	0.927	0.860	**0.862**	0.860	0.861	0.860	0.860
Fungi	0.780	0.793	0.940	0.905	**0.953**	0.866	0.887
GestureMidAirD1	0.723	0.675	0.535	0.699	0.610	0.692	**0.700**
GestureMidAirD2	0.669	0.568	0.303	0.530	0.441	**0.576**	0.571
GestureMidAirD3	0.415	0.401	0.402	0.405	**0.421**	0.406	0.418
GesturePebbleZ1	0.814	0.751	0.756	0.746	**0.767**	0.749	0.744
GesturePebbleZ2	0.728	0.646	0.655	0.652	**0.665**	0.648	0.651
GunPointAgeSpan	0.946	0.956	**0.972**	0.956	0.966	0.956	0.956
GunPointMaleVersusFemale	0.987	0.968	0.974	0.968	**0.975**	0.968	0.968
GunPointOldVersusYoung	0.987	**1.000**	0.978	**1.000**	0.992	**1.000**	**1.000**
HouseTwenty	0.891	0.915	**0.917**	0.915	0.915	0.915	0.915
InsectEPGRegularTrain	0.992	**1.002**	0.968	1.000	0.977	1.000	1.000
InsectEPGSmallTrain	0.920	0.910	0.854	**0.920**	0.862	0.914	0.920
MelbournePedestrian	0.889	0.899	0.866	0.906	0.883	0.905	**0.906**
MixedShapesRegularTrain	0.946	0.928	**0.942**	0.931	0.940	0.930	0.931
MixedShapesSmallTrain	0.906	**0.901**	0.890	0.897	0.891	0.900	0.899
PLAID	0.795	0.862	0.809	0.863	0.855	0.866	**0.866**
PickupGestureWiimoteZ	0.680	**0.869**	0.745	0.848	0.783	0.859	0.859
PigAirwayPressure	0.130	0.449	0.369	0.576	0.468	0.517	**0.591**
PigArtPressure	0.808	0.818	0.837	0.859	**0.861**	0.841	0.846
PigCVP	0.688	-0.174	-0.167	**-0.138**	-0.158	-0.147	-0.144
PowerCons	0.967	0.899	0.905	0.899	**0.915**	0.899	0.899
Rock	0.660	0.665	0.676	**0.690**	0.687	0.672	0.677
SemgHandGenderCh2	0.897	**0.795**	0.767	**0.795**	0.791	**0.795**	**0.795**
SemgHandMovementCh2	0.622	0.691	0.502	0.671	0.568	**0.698**	0.697
SemgHandSubjectCh2	0.831	**0.886**	0.795	0.857	0.839	0.881	0.874
ShakeGestureWiimoteZ	0.800	**0.691**	0.625	0.687	0.620	0.685	0.682
SmoothSubspace	0.947	0.840	**0.873**	0.845	0.864	0.844	0.844
UMD	0.972	0.049	0.161	0.307	0.160	0.117	**0.307**