Towards Clear Expectations for Uncertainty Estimation

Victor Bouvier, Simona Maggio, Alexandre Abraham & Léo Dreyfus-Schmidt
Dataiku
{name.surname}@dataiku.com

Abstract

If Uncertainty Quantification (UQ) is crucial to achieve trustworthy Machine Learning (ML), most UQ methods suffer from disparate and inconsistent evaluation protocols. We claim this inconsistency results from the unclear requirements the community expects from UQ. This opinion paper offers a new perspective by specifying those requirements through five downstream tasks where we expect uncertainty scores to have substantial predictive power. We design these downstream tasks carefully to reflect real-life usage of ML models. On an example benchmark of 7 classification datasets, we did not observe statistical superiority of state-of-the-art intrinsic UQ methods against simple baselines. We believe that our findings question the very rationale of why we quantify uncertainty and call for a standardized protocol for UQ evaluation based on metrics proven to be relevant for the ML practitioner.

1 Introduction

Uncertainty Quantification (UQ) is a critical requirement for trustworthy Machine Learning (ML) in high-risk applications (Hüllermeier & Waegeman, 2021). However the very concept of uncertainty has no agreed upon formal definition. A myriad of definitions and quantifiers for uncertainty exist based on different paradigms. For instance, calibration aims to minimize the *Expected Calibration Error* (ECE) (Naeini et al., 2015) or the *Brier Score* (Brier et al., 1950), Conformal Prediction seeks to produce the smallest set of predictions (Angelopoulos et al., 2020; Angelopoulos & Bates, 2022), while intrinsic UQ measures the ability to separate uncertainty sources, whether it comes from the data (aleatoric) or the model (epistemic) (Malinin et al., 2020; Hüllermeier & Waegeman, 2021).

If those methods cater to different needs, we found the requirements for the UQ they represent not well specified and often heterogeneous, as supported by Table 1. We initiate to fill this gray area of the literature by proposing clear expectations for UQ through a collection of downstream tasks that reflect realistic cases the practitioner encounters when needing to deploy an ML model reliably. Crucially, these downstream tasks enable comparison and validation of UQ methods. To this purpose, the contributions of this opinion paper are:

- A set of natural downstream UQ tasks where uncertainty is expected to play a decisive role, such as classification with rejection, error detection, out-of-distribution (OoD) detection, shift detection and performance drop prediction.
- A methodology to evaluate the predictive power of uncertainty scores on those downstream tasks. To assess the usefulness of uncertainty scores, we always propose simple baselines for the task at hand.
- An example of the above methodology through a benchmark of UQ methods on 7 tabular classification datasets, an under-explored modality for UQ. This shows that, quite surprisingly, for any task simple baselines are often sufficient alternatives to more advanced uncertainty quantification.

Reference	Dataset	OOD Dataset	Downstream tasks
Lakshminarayanan et al. (2017)	UCI Regression MNIST, SVHN, ImageNet	No OOD Left-out class	Calibration Calibration, OoD detection
Ovadia et al. (2019)	MNIST CIFAR10, ImageNet 20 NG (even classes)	Rotated, Translated, NotMNIST Gaussian blurred, SVHN 20 NG odd classes, One billion word	Calibration
Filos et al. (2019)	UCI Regression Diabete retinopathy	No OOD APTOS 2019 blindness detection	Calibration Retention
Malinin et al. (2020)	UCI regression-classification	Year MSD	Retention, OoD detection
Shaker & Hüllermeier (2020)	spect diabetes	No OOD	Retention
Nado et al. (2021)	ImageNet	ImageNet-C/A/V2	Calibration, OoD detection
Daxberger et al. (2021)	WILDS	WILDS shifts	Calibration
Malinin et al. (2021)	weather ante 2019-04, not snow Machine translation EN-RU newspaper	weather post 2019-07, snow Other vocabulary (Reddit)	Retention

Table 1: Evaluation Metrics and Tasks in SOTA works on UQ. Prior works in the literature do not share a standard evaluation protocol, using different datasets or metrics, making it difficult to assess the actual progress of Uncertainty Quantification for Machine Learning.

2 Challenges in Uncertainty Evaluation

While prior works in UQ propose a plethora of different techniques and specific tasks to evaluate them, there is no shared requirement an uncertainty score should fulfill. We outline the diverse landscape of UQ evaluation and highlight the requirements we believe an uncertainty estimation should comply with to be used to guarantee trustworthy ML.

UQ Evaluation in the literature. Many methods has been designed for uncertainty estimation (Eyke, 2020; Abdar et al., 2021), yet only limited work has focused on how to evaluate and benchmark these approaches (Nado et al., 2021; Malinin et al., 2021), mainly focusing on deep classification of images (Filos et al., 2019; Atanov et al., 2020). As summarized in Table 1, research papers proposing new UQ estimation methods do not share the same evaluation protocol, some focusing on calibration metrics, some on downstream tasks metrics. Error and Out-of-Distribution (OoD) detection are the most common downstream tasks, but the validation remains often partial with either a single task evaluated or a qualitative assessment that uncertainty estimates increase in specific situations (Ovadia et al., 2019). Importantly, the lack of comparison to a baseline tailored to the task at hand makes it difficult to assess the actual practicality of an uncertainty score compared to a simpler alternative, such as using an anomaly detection method for the OoD detection task. Additionally, designing downstream tasks is problematic too. From defining what, an OoD sample is, for which there is no standard in particular for tabular data, to comparing performances on the error detection task that could be misleading (Atanov et al., 2020).

Requirements for reliable ML. We believe that robust uncertainty estimation is a cornerstone towards trustworthy ML that naturally leads to the following five downstream tasks where a good uncertainty score should perform well.

Retention. Uncertainty scores should help deferring the prediction to a human expert, which is a typical case of medical diagnosis with ML where difficult samples are deferred to the medical expert. Formally, for a budget $r \in [0, 1]$, an oracle provides the ground-truth for $r\%$ of the data with higher uncertainty score. As a metric, we report the area under the $F1$-score between oracle-augmented predictions and the true label, when varying $r \in [0, 1]$.

Error detection. Uncertainty scores should help detecting model errors, as higher uncertainty should correlate with higher error probability. It reflects the situation where model errors are highly detrimental, and one should rather abstain from predicting. As a metric, we report the Area Under the ROC curve (AUROC) when the uncertainty score is used as the prediction score of the test set errors.

Out-of-Distribution (OoD) detection. We expect uncertainty scores to support detection of OoD samples from different domains, that are generated via splits along specific feature values. As the model is expected to be more uncertain on OoD samples, we report the AUROC when the uncertainty score is used as the prediction score of the test set errors to detect the domain, *i.e.* whether the sample comes from the shifted or unshifted dataset.

Shift detection. Uncertainty scores are used as proxy for detecting data shift (Rabanser et al., 2019). In practice, we perform a Kolmogorov-Smirnov test between the distribution of uncertainty scores on the test set and of shifted datasets, also generated via splits along specific feature values. We report the accuracy metric of detecting shift with a signifiance level $\alpha = 0.05$ based on 100 bootstrap of shifted datasets.

Performance Drop Prediction. We expect that the ratio of uncertainty scores above a given threshold reflect the error rate on several shifted datasets, generated as in the task above. The threshold is selected so that the ratio of samples with uncertainty above the threshold matches the error rate on the test set, similarly to the Averaged Thresholded Confidence (ATC) technique (Garg et al., 2022). We report as metric the mean absolute error between the estimated performance drop and the true performance drop.

It is worth noting the described tasks share similarities. Both retention and error detection assess how valuable is an uncertainty score for detecting difficult in-distributions samples, *i.e.* for which the model is likely to fail to predict, and depend on the primary model. Both OoD detection and shift detection assess how valuable is an uncertainty score for detecting samples that are far from the data distribution the model was trained on. The task of the performance drop prediction combines both assessments and is inspired by recent work (Garg et al., 2022).

3 Benchmarking UQ Methods for Reliable ML

As a validation protocol, we use uncertainty score derived from UQ methods on the five downstream tasks together with a simple baseline to assess the actual practicality of an uncertainty score.

Comparing UQ. We aim to compare both agnostic UQ, *i.e.* quantifying uncertainty of a black-box model, and intrinsic UQ, *i.e.* a model that natively quantifies its uncertainty. First, we study two agnostic UQ; *Isotonic Calibration* (IC) and *Conformal Prediction*[1] (CP). Various uncertainty scores can be derived from conformal predictions; the *confidence* which measures how certain the model is that the prediction set is a singleton, the *credibility* which measures how certain the model is that the prediction set is not empty, and the *p-value* that is the quantile of the non-conformity score associated with prediction in the calibration samples with label equal to the prediction (Shafer & Vovk, 2008). For the isotonic calibration module, we use as uncertainty score the *Max-Confidence* uncertainty score as 1 minus the maximal predicted probability over classes. Second, we study intrinsic UQ that aims to separate the aleatoric from the epistemic uncertainty. To this purpose, we compare three different scores: the total, aleatoric and epistemic uncertainties (Hüllermeier & Waegeman, 2021). To compare as fairly as possible intrinsic and agnostic UQ, we fix the base model of intrinsic UQ, considering it as a black-box model on which we plug agnostic UQ. Such methodology is motivated by the conclusion from (Atanov et al., 2020) which stresses that error-based tasks (in our particular setup: retention, error detection and performance drop prediction) depend on the primary model, thus are not directly comparable across different primary models.

[1]We used the confidence as non-conformity score. We refer the reader to (Angelopoulos et al., 2020) for a comprehensive overview of conformal predictions.

Primary model	Agnostic UQ	Score	Retention	Error Detection	OoD Detection	Shift Detection	Perf. Drop Pred.
Logistic Regression	Isotonic Calibration	Max-Confidence	**0.942**	**0.763**	0.460	0.489	**0.096**
	CP (Confidence)	p-value	0.937	0.719	0.486	0.537	**0.075**
		Credibility	0.929	0.665	0.480	0.409	**0.088**
		Confidence	0.929	0.658	0.466	0.355	**0.091**
	None	Baseline	**0.942***	**0.768***	**0.639**†	**0.956**†	**0.080***
Random Forest	Isotonic Calibration	Max-Confidence	**0.953**	**0.789**	0.560	0.877	0.053
	CP (Confidence)	p-value	0.946	0.738	0.576	0.847	**0.036**
		Credibility	0.941	0.694	0.572	0.812	**0.041**
		Confidence	0.941	0.697	0.567	0.836	**0.046**
	None	Baseline	**0.953***	**0.797***	**0.639**†	**0.956**†	**0.035***
Gradient Boosting Trees	Isotonic Calibration	Max-Confidence	**0.951**	**0.785**	0.506	0.811	0.032
	CP (Confidence)	p-value	0.945	0.736	0.512	0.796	**0.018**
		Credibility	0.940	0.695	0.510	0.782	0.028
		Confidence	0.940	0.695	0.503	0.759	0.028
	None	Baseline	**0.951***	**0.794***	**0.639**†	**0.956**†	**0.016***
Multi-Layers Perceptron	Isotonic Calibration	Max-Confidence	**0.953**	**0.789**	0.459	0.812	**0.091**
	CP (Confidence)	p-value	0.945	0.718	0.461	0.761	**0.080**
		Credibility	0.942	0.694	0.459	0.748	**0.080**
		Confidence	0.942	0.692	0.453	0.717	**0.082**
	None	Baseline	**0.953***	**0.793***	**0.639**†	**0.956**†	**0.079***
Deep Ensemble (Intrinsic UQ)	None	Total	**0.955**	**0.798**	0.506	0.899	**0.056**
		Aleatoric	**0.955**	**0.796**	0.470	0.893	0.092
		Epistemic	0.947	0.748	0.579	0.932	0.092
	Isotonic Calibration	Max-Confidence	**0.955**	**0.792**	0.500	0.824	**0.070**
	CP (Confidence)	p-value	0.947	0.725	0.506	0.749	**0.056**
		Credibility	0.944	0.698	0.503	0.734	**0.059**
		Confidence	0.944	0.695	0.495	0.721	**0.060**
	None	Baseline	**0.955***	**0.798***	**0.639**†	**0.956**†	**0.056***
CatBoost (Intrinsic UQ)	None	Total	**0.957**	**0.806**	0.512	0.894	**0.046**
		Aleatoric	**0.957**	**0.806**	0.512	0.874	**0.046**
		Epistemic	0.939	0.676	0.551	0.895	0.080
	Isotonic Calibration	Max-Confidence	**0.956**	0.796	0.505	0.818	**0.066**
	CP (Confidence)	p-value	0.950	0.740	0.537	0.773	**0.046**
		Credibility	0.946	0.706	0.510	0.769	**0.052**
		Confidence	0.946	0.709	0.502	0.782	**0.054**
	None	Baseline	**0.957***	**0.806***	**0.639**†	**0.956**†	**0.046***

Table 2: Results of the example 7 datasets benchmark. In bold, metrics with overlapping confidence interval with best performer based on statistics over 10 different seeds, as described in **Experimental setup**. We observe strong performance from the simple baselines, whether for error detection or OoD and shift detection (†), questioning the relative benefit of specific UQ methods.

Baseline. To assess the utility of an uncertainty score in a downstream task, we compare it against simpler alternatives. For the error-based tasks that are model dependent, we used the *Max-Confidence* score from the black-box primary model, *i.e.* without any UQ. In table 2, we flag results using this score with $\star$ symbol. For OoD-based tasks, we report a naive anomaly detection score based on the mean distance of a sample to its ten nearest neighbors with same label on training data building an uncertainty score which is independent on the primary model. Being model independent, it is worth noting the results based on the anomaly detection score are shared for all models. In table 2, we flag results using this score with † symbol.

Experimental setup. We focus on binary classification tasks from seven tabular datasets. We used the UCI datasets[2] *Adult Income*, *Video Games*, *Default Credit Card*, *Bank*, *Heart*, *Forest Covertype* and *BNG Zoo*[3] We detail in Appendix A how OoD data is obtained. For agnostic UQ, we studied four different primary models: Logistic Regression, Random Forest, Gradient Boosting Trees and a Multi-Layer Perceptron (MLP) implemented using the `scikit-learn` library (Pedregosa et al., 2011) with default parameters for each estimator. For intrinsic UQ, we used two influential methods: *Deep Ensemble* (Lakshminarayanan et al., 2017) with an ensemble of 10 MLPs, and *CatBoost* (Malinin et al., 2020) with `RMSEWithUncertainty` loss. We compute mean metrics m and their standard deviation σ over 10 random seeds and aggregated on the seven datasets. Bold metrics are such that their confidence interval $[m - \sigma, m + \sigma]$ lower bound is higher than any other confidence

[2] `archive.ics.uci.edu/ml/datasets`

[3] For datasets that are not binary tasks (Forest Covertype and BNG Zoo), we derive a binary task in a one-vs-all fashion using the majority class as the positive class.

interval higher bound. As error-based tasks (retention, error detection and performance drop prediction) are not directly comparable (Atanov et al., 2020), we bold metrics for each primary model separately.

Analysis. Although primary models are not directly comparable on error-based tasks, results presented in Table 2 show that metrics range in a similar way. More importantly, we find that no UQ method clearly outperforms others and that tailored baselines performs similarly on error-based tasks than UQ methods, and even outperforms them on OoD detection and shift detection. Thus, if the utility of UQ methods for trustworthy ML is to be judged on those downstream tasks, we advocate for the use of those simple baselines.

4 Conclusion

We challenge the evaluation of *Uncertainty Quantification* (UQ) as a literature review showed inconsistent evaluation protocols. For UQ to help guarantee more reliable use of ML models, we proposed five natural downstream tasks. To our knowledge, this new perspective for UQ evaluation is the first that allows to compare very different paradigms for UQ, from Conformal Prediction to Intrinsic UQ. Quite surprisingly, our example benchmark shows that relying simply on the probability estimates of the primary model is a fair, simple and robust baseline for error-related downstream tasks, while for OoD-related tasks more complex techniques based on separation of sources of uncertainties perform significantly lower than a simple baseline. As concluding words, we draw recommendation for future work towards more reliable UQ evaluation protocol;

1. Adding more datasets would strengthen the statistical significance of results, and allow the use of more advanced ranking methods, such as autorank (Herbold, 2020).
2. Developing more downstream tasks that captures the expectation from uncertainty scores together with simple generic baseline when available.
3. Our analysis is restricted to binary classification task and would benefit being extended to regression tasks.

References

Moloud Abdar et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 15662535. doi: 10.1016/j.inffus.2021.05.008. URL https://doi.org/10.1016/j.inffus.2021.05.008.

Anastasios N Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. 2022.

Anastasios Nikolas Angelopoulos et al. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2020.

Andrei Atanov et al. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. pp. 1–6, 2020.

Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Erik Daxberger et al. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34, 2021.

H Eyke. Aleatoric and Epistemic Uncertainty in Machine Learning : An Introduction to Concepts and Methods. pp. 1–59, 2020.

Angelos Filos et al. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. 2019.

Saurabh Garg et al. Leveraging Unlabeled Data to Predict Out-of-Distribution Performance. (NeurIPS), 2022. URL http://arxiv.org/abs/2201.04234.

Steffen Herbold. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173, 2020.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Balaji Lakshminarayanan et al. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips): 6403–6414, 2017. ISSN 10495258.

S. Maggio and L. Dreyfus-Schmidt. Ensembling shift detectors: an extensive empirical evaluation. *ECML PKDD*, abs/2106.14608, 2021. URL `https://arxiv.org/abs/2106.14608`.

Andrey Malinin et al. Uncertainty in Gradient Boosting via Ensembles. pp. 1–17, 2020. URL `http://arxiv.org/abs/2006.10562`.

Andrey Malinin et al. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks. pp. 1–26, 2021. URL `http://arxiv.org/abs/2107.07455`.

Zachary Nado et al. Uncertainty Baselines: Benchmarks for Uncertainty & Robustness in Deep Learning. pp. 1–12, 2021. URL `http://arxiv.org/abs/2106.04015`.

Mahdi Pakdaman Naeini et al. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Yaniv Ovadia et al. Can You Trust Your Model's Uncertainty ? Evaluating Predictive Uncertainty Under Dataset Shift. (NeurIPS), 2019.

Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Stephan Rabanser et al. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008. ISSN 15324435.

Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In *International Symposium on Intelligent Data Analysis*, pp. 444–456. Springer, 2020.

A Out-of-Distribution data

We detail how we split the dataset to obtain *Out-of-Distribution* (OoD) data for our experiments;

- Adults income: we split in-distribution data and OoD data with respect to the "gender" features; "male" is in-distribution, "female" is out-of distribution.
- Video games: we split in-distribution data and OoD data with respect to the "Genre" features; "Action" is OoD, other values are in-distribution.
- Heart: we split in-distribution data and OoD data with respect to the "gender" features; "2" is out-of-distribution, other values are in-distribution.
- Bank: we split in-distribution data and OoD data with respect to the "job" features; "student" is OoD, other values are in-distribution.
- Default of credit card client: we split in-distribution data and OoD data with respect to the "SEX" features; "1" is OoD, other values are in-distribution.

- Forest Covertype: we split in-distribution data and OoD data with respect to the "Wilderness_Area1" features; "1" is out-of-distribution, other values are in-distribution.
- BNG Zoo: we split in-distribution data and OoD data with respect to the "domestic" features; "True" is OoD, other values are in-distribution.

We then remove the feature used for the split for both the in-distribution and the out-of-distribution datasets. Shifted data for the shift detection and performance drop prediction tasks should be generated via synthetic perturbations as outlined in Section 2. We plan to include synthetically shifted data in the benchmark as future work, but in the experiments presented in the paper we used the OoD dataset as a shifted dataset for shift detection and performance drop prediction as well.