ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Evaluation methodology for disentangled uncertainty quantification on regression models

Kevin Pasini[1], Clement Arlotti[1], Milad Leyli-abadi[1], Marc Nabhan[2] and Johanna Baro[1] *

1- IRT SystemX, Paris, France 2- Air Liquide, Les-Loges-en-Josas, France

Abstract. A practical way to enhance the confidence of the predictions made by Machine Learning (ML) models is to enrich them with trustworthiness add-ons such as Uncertainty Quantification (UQ). Existing UQ paradigms capture two intertwined components (epistemic and aleatoric), but few of them evaluate their disentanglement, even less on real data. We thus propose and implement a methodology to assess the effectiveness of uncertainty disentanglement despite the absence of ground truth in real datasets. To do so, we use a data withdrawal-based strategy to simulate Out-of-Distribution (OOD) data and evaluate four state-of-the-art UQ approaches.

1 Introduction and related work

Complex industrial systems are now equipped with multiple sensors to process massive amounts of dynamic data and deploy AI-based monitoring models. However, to enhance their trustworthiness, these models should be complemented by an uncertainty management framework[2]. In the field of ML-UQ, several paradigms claim to produce models able to separate and quantify two distinct components contributing to the total uncertainty [3, 10, 6, 5]. The aleatoric component captures irreducible variability in a model prediction due to inherent noise in the data, while the epistemic component is related to the model relevance when facing an atypical input and can be reduced by observing more data during the training step [8].

However, *disentangled Uncertainty Quantification* (dUQ) faces both technical and methodological difficulties. On real data with noisy and limited observations, the epistemic and aleatoric components are strongly entangled [8]. Plus, no ground truth allows the evaluation of the quantification and even less the possibility of the decomposition. Our contribution thus addresses the methodological challenge of epistemic model confidence evaluation, in the absence of ground truth in real data. We will first compare recent works on *Uncertainty Quantification* (UQ) regression in *Machine Learning* (ML), then propose a dUQ evaluation methodology based on a data withdrawal-based strategy that simulates OOD, aiming to assess the effectiveness of aleatoric and epistemic uncertainty disentanglement.

Three main UQ paradigms are extensively studied in the literature for deep learning models: The *Bayesian* formalism [4] is used to develop probabilistic methods for UQ (e.g., *Monte Carlo Dropout*, *Monte Carlo Markov chain*, ...); The *ensemble models* [9] are also widely used due to their simple implementation (namely *Deep Ensemble*); Finally, *Evidential Deep Learning* (EDL) [3] learns a distribution over the parametric space of model outcomes and collects evidence regarding the model predictions.

*This work has been supported by the French government under the France 2030 program, as part of the SystemX Technological Research Institute within the Confiance.ai program (www.confiance.ai).

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

We propose to analyze the state-of-the-art approaches for UQ through different angles in Table 1. We can observe that few papers suggest UQ decomposition and many of them are applied on synthetic dataset without comprehensive interpretation of the results. There is also a lack of cross-comparison among various paradigms. In this paper, we seek to bridge this gap by providing a comprehensive benchmark of these UQ paradigms using both real and synthetic datasets.

Table 1: Summary and comparison of UQ approaches

		Features			Problem support		Environment setup			
	Methods	UQ paradigm	Prior	UQ Decomposition	Regression	Classification	Dataset	Evaluation criteria	Interpretation	Baselines
UQ Papers	MCDP [7]	Drop-out	No	No	Yes	Yes	Public / Synthetic (diverse)	NLL / RMSE	Local	Yes (diverse)
	DeepEnsemble [9]	Ensemble	No	No	Yes	Yes	Public / Synthetic (stationary)	NLL / RMSE Brier/Calibration	Local Qualitative	Yes (diverse)
	AutoDEUQ [6]	Ensemble	No	Yes	Yes	No	Public / Synthetic (stationary)	NLL/RMSE	No	Yes (diverse)
	EDL [3]	Evidential	Yes	No	Yes	No	Public / Synthetic (stationary)	NLL/RMSE	Partial Quantitative	Yes (diverse)

Color codes (green: satisfying, orange: partial, red: ignorance)

2 Disentangled Uncertainty Quantification framework

To enable the cross-comparison and benchmarking of state-of-the-art approaches, we manipulate a UQ decomposition formalism to obtain a unified set of indicators. Next, a set of experiments based on statistical tests are designed to evaluate the model epistemic confidence (see Figure 1). In the following, we first introduce the context and notations and describe these components in more detail.

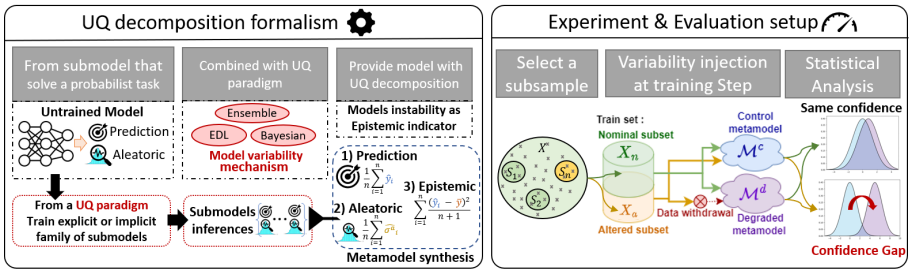

Fig. 1: Disentangled UQ framework and the experimental setup

Context and notations A dataset D (real-world or synthetic) in a supervised learning domain is composed of pairs $\{(\mathbf{x}_i, y_i)\}_{i\in 1:N}$ representing the input features and outputs. In this work, we focus on the regression task; therefore, the output variable y is continuous. The input vector $\mathbf{x}_i$ includes contextual variables, latent variables h_i and past timestamps of target variables in the case of time series data. The partitions of the dataset into various samples are noted as $D = \{S_1, S_2, \ldots, S_n\}$ with $(n < N)$. We assume that the data is generated by a function $y_i = f(x_i) + \varepsilon_i$ with: $\varepsilon_i \sim \mathcal{N}(0, \sigma_i(\mathbf{x}_i, \mathrm{h}_i))$ being a Gaussian noise not explained by the model and associated to irreducible variability. The objective of a ML model $\hat{f}(x)$ is to approximate the true function $f(x)$ by minimizing the mean squared error $(y_i - \hat{f}(x_i))^2$.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

UQ decomposition formalism According to a common view in the ML-UQ community [10, 5], dUQ approaches are often composed of two parts. An explicit or implicit family of submodels (approached via an ensemble of submodels), each one providing a prediction and an aleatoric estimation. A metamodel synthesizing these outputs and producing an epistemic confidence level based on the variability of prediction in the family. The metamodel, denoted $\mathscr{M}$, learns and manipulates diverse submodels $\hat{f}_\theta$ to combine their inferences. The learning phase aims to capture the explainable variability (true function f) and estimate irreducible variability by exploring a diversity of submodel candidates. To ensure diversity and avoid redundancy, a variability infusion mechanism (denoted δ and depending on the UQ paradigm) is needed during the learning phase (e.g., MC-dropout). According to the bias-variance trade-off, we can decompose the metamodel expected error:

$$\boldsymbol{E}_{\boldsymbol{\theta}}\left[(y_i-\hat{f}_\theta(\mathbf{x}_i))^2\right]=\underbrace{E_y\left[(y_i-f(\mathbf{x}_i))^2\right]}_{\text{Intrinsic variability}}+\underbrace{\boldsymbol{E}_{\boldsymbol{\theta}}\left[\hat{f}_\theta(\mathbf{x}_i)-f(\mathbf{x}_i)\right]^2}_{\text{Bias}}+\underbrace{\boldsymbol{E}_{\boldsymbol{\theta}}\left[\left(\boldsymbol{E}_{\boldsymbol{\theta}}\left[\hat{f}_\theta(\mathbf{x}_i)\right]-\hat{f}_\theta(\mathbf{x}_i)\right)^2\right]}_{\text{Variance}} \quad (1)$$

Intrinsic variability correspond to the gap between the deterministic true function $f(x_i)$ and the real observation y_i due to randomness. It is pure aleatoric that cannot be caught by any model. The *bias* term corresponds to the average gap between submodels and the true function f at the value x_i. It acts as irreducible variability that can't be handled by our modeling (due to constraints and bad assumptions) and so, is assimilated to aleatoric too. Then the *variance* term corresponds to the dispersion of submodels prediction. It can be assimilated to epistemic variability due to lack of observations and could be reduced in our modeling scope by gathering more data. Finally, the metamodel provides a *regression with dUQ* based on the average of the predictions $\hat{y}_{i,\theta}$ and the aleatoric estimations $\hat{\sigma}^a_{i,\theta}$ of submodels, plus an epistemic indicator $\widehat{\sigma}^e_i$ estimated from the dispersion of submodels predictions.

Experiment and evaluation In this section, we propose a set of experiments to evaluate the model epistemic confidence, despite the absence of ground truth by using a training data withdrawal strategy that simulate OOD data. They aim at highlighting epistemic confidence gaps between models prediction on nominal and OOD data. This strategy consists in ablating partially or totally a selected subset of data (called the altered subset) from the training set (see Figure 1). Two instances of the metamodel are trained on nominal (X_n) and altered (X_a) datasets (called control $\mathscr{M}^c$ and degraded $\mathscr{M}^d$ metamodels respectively). The epistemic confidence gap is quantified by comparing the predictions of these instances on test sets corresponding to nominal and altered queries and through statistical hypothesis tests. We used the negative epistemic ratio under total log-likelihood ($I^e=-\ln(1+\frac{\hat{\sigma}^a}{\widehat{\sigma}^e})$ called *disentangled Epistemic indicator* (dE-Ind)) which act as a relative epistemic indicator.

Our statistical framework compares dE-ind distribution using Wilcoxon-Mann-Whitney and Wilcoxon signed rank tests (denoted T1 and T2), with the null hypothesis H_0 that the distribution of dE-Ind computed on nominal and altered sets are identical. The alternative hypothesis H_1 is that the distribution of dE-Ind over altered data dominates the distribution computed on the nominal dataset.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

3 Experimental settings and results

The benchmark compare four UQ paradigms on univariate time series: Random Forest disentangled Uncertainty Quantification (RF-dUQ)[1], Probabilistic Neural Network Monte Carlo Dropout (PNN-MCDP) [7], Probabilistic Neural Network Deep Ensemble (PNN-DE)[9] and Evidential Deep Learning regression (EDL)[3].

The experiments are performed on both real and synthetic datasets. The real dataset corresponds to weekly demand of industrial gas deliveries (continuous target variable) issued from 29 time series collected across different areas during 7 years ($29 \times 7 \times 53$ weekly observations). It comprises 29 features including context, calendar and history of gas demand. We can split the time series into 3 levels of variability due to their heterogeneity and heteroscedasticity. The synthetic dataset corresponds to times series of 16000 observations and 16 features. The data are generated following a local time-dependent Gaussian distribution including white noise.

UQ-regression performance with data-alteration: We evaluate our four approaches on a nominal setup (without training data removal) to ensure comparable accuracy and calibrated variance. We use root mean squared errors (RMSE) and negative log-likelihood (NLL) metrics to assess respectively regression, and regression under uncertainty performances. We consider Sharpness and Coverage (i.e. size and % of data in the confidence interval respectively) to evaluate UQ-relevance. Table 2 summarizes the results between mentioned models and a simple Multi-Layer perceptron (MLP) model without UQ. Each approach obtains a coverage close to the theoretical one, although PNN-DE yields narrower confidence intervals for similar coverage.

Table 2: Test set performance of *nominal* setup using our two forecasting datasets.

Approach	MLP	RF-dUQ	PNN-MCD	PNN-DE	EDL	RF-dUQ	PNN-MCD	PNN-DE	EDL
Dataset	RMSE metrics (lower is better)					NLL metrics (lower is better)			
real	$0.22_{\pm 0.02}$	$0.23_{\pm 0.02}$	$0.22_{\pm 0.2}$	$\mathbf{0.22}_{\pm 0.02}$	$0.22_{\pm 0.01}$	$-0.51_{\pm 0.06}$	$-0.55_{\pm 0.08}$	$\mathbf{-0.57}_{\pm 0.07}$	$-0.55_{\pm 0.08}$
synthetic	$\mathbf{0.43}_{\pm 0.01}$	$\mathbf{0.43}_{\pm 0.01}$	$0.44_{\pm 0.01}$	$0.43_{\pm 0.01}$	$0.44_{\pm 0.01}$	$\mathbf{0.43}_{\pm 0.01}$	$0.46_{\pm 0.01}$	$\mathbf{0.40}_{\pm 0.02}$	$0.44_{\pm 0.01}$
Dataset	Sharpness*(lower is better)					Coverage (Target: 95.65%)			
real	∅*	$0.82_{\pm 0.01}$	$0.81_{\pm 0.02}$	$\mathbf{0.73}_{\pm \mathbf{0.01}}$	$0.75_{\pm 0.02}$	$94.9_{\pm 0.8}$	$94.9_{\pm 1.3}$	$95.1_{\pm 1.4}$	$94.4_{\pm 1.7}$
synthetic	∅*	$1.78_{\pm 0.01}$	$1.86_{\pm 0.05}$	$\mathbf{1.56}_{\pm \mathbf{0.03}}$	$1.80_{\pm 0.03}$	$96.7_{\pm 0.1}$	$96.3_{\pm 0.1}$	$95.0_{\pm 0.01}$	$96.5_{\pm 0.2}$

*NLL, Coverage and sharpness is meaningless for the MLP model.

Detailed d-UQ evaluation on real dataset: The real dataset is composed of three subsets with different variability levels: low, mid (the altered one here) and high. Fig. 2 presents the detailed results for the experiment that withdraws 98% of training data of the mid-var subset. For each approach, performances of the control and degraded models (denoted by *c* and *d* respectively) are shown for each subset. By comparing control vs degraded models, we observe close performances for all metrics on nominal subsets. On the contrary, for the altered subset, the training data withdrawal leads to an increasing RMSE (arrows 1) but no significant change for aleatoric sharpness. Moreover, all approaches (except EDL) display significant increase in their dE-Indicators. Such loss of epistemic confidence combined with unchanged aleatoric metric (A-sharpness) on the altered subset (that simulates OOD data) suggests an effective uncertainty disentanglement. Conversely, the EDL approach does not display any significant increase, suggesting dUQ ineffectiveness in this setup.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

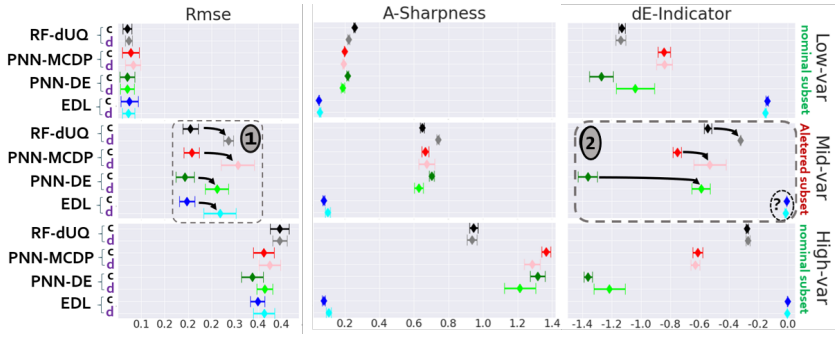

Fig. 2: Subset performances for one experimental setup on real data. Control and degraded models are denoted by *c* and *d*.

Figure 3 shows the RF-UQ predictions with total UQ (blue uncertainty envelopes) and epistemic indicators (point coloration based on indicators). We easily observe that the model shows a lack of confidence when predicting values belonging to the altered context (acting as out-of-distribution data).

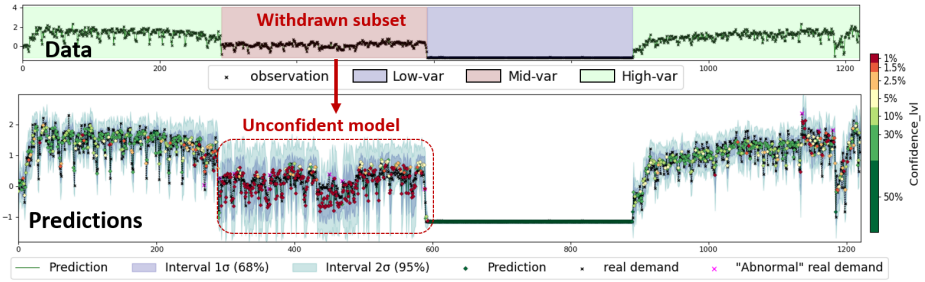

Fig. 3: RF-UQ prediction with dUQ indicators from mid-var altered setup

D-UQ evaluation synthesis: To ensure experimental robustness, 8 variants of the experiments (with 4 cross-validations) were performed on real and synthetic data. Using the statistical framework based on dE-Ind distribution shifts, the goal is to determine if the training data withdrawal affects the epistemic component and if this impact is greater than on the aleatoric component. The statistical framework is based on two tests ($T1$ & $T2$). If an approach passes both of them, we can claim that a lack of confidence is expressed by the epistemic component on the altered subset samples. Figure 4 displays test results for all experiments.

PNN-DE and PNN-MCDP show successful results in almost all configurations for both real and synthetic datasets. RF-dUQ fails on the high-var setup. EDL fails in almost all configurations, illustrating that dUQ is not effective here, either due to the intrinsic behavior of the approach or to parameterization issues in spite of hyper-parameter optimization. The perturbation of the low-variability subset (low-var-98 and low-var-100, indicating 98% and 100% data widthdrawal respectively) leads to small test scores for all approaches, suggesting difficulties to express low confidence in small variability data, even with few observations. We also note that withdrawal

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

percentage does not have a significant impact on dUQ effectiveness. A potential explanation is that neighboring samples belonging to non-removed subsets still retain part of the supporting information for prediction.

		Real dataset						Synthetic dataset	
Injection on		Low-var		Mid_var		High-var		Sub-context	
% Drop		98	100	98	100	98	100	98	100
RF-UQ	T1	7.4	-1.6	19.6	22.0	-2.6	-12.8	55.3	68.1
	T2	10.8	5.0	22.4	22.3	-12.2	-19.0	43.7	44.5
PNN-MCDP	T1	12.7	9.2	19.1	20.3	6.4	12.9	18.6	26.5
	T2	11.6	8.4	14.2	15.3	5.7	13.5	14.9	21.1
PNN-DE	T1	2.8	9.0	29.4	30.9	13.4	21.0	28.6	38.4
	T2	3.1	7.1	23.0	24.8	12.2	17.9	25.5	35.8
EDL	T1	2.7	8.2	-1.6	-1.1	-0.9	-8.0	14.0	2.0
	T2	30.8	31.3	-9.4	-9.0	-7.5	-22.7	-1.0	-33.4
Coloration legend :		: Success (s>4σ)		: Minor failure (0σ > s > 4σ)		: Major Failure (s < 0σ)			

Fig. 4: Results of the statistical tests $T1$ and $T2$ for all experiments.

4 Conclusion and perspectives

We proposed an UQ-decomposition formalism alongside an evaluation methodology. Our formalism was based on a metamodel concept, taking a model as input and providing disentangled UQ indicators as its output. To assess the epistemic confidence in the absence of ground truth, the evaluation methodology was based on training data withdrawal. Experiments performed using four models and a real and a synthetic dataset demonstrated the dUQ relevance and effectiveness on heterogeneous data. Some models (RF, MCDP, DE) show relevant local aleatoric and epistemic indicators, while others (EDL) show limitations for epistemic estimation. As perspectives, we plan to considerate new data alteration techniques, to tackling more complex temporal data with specific NN architecture in order to perform real-time anomaly detection with UQ.

References

[1] Shakeret al. Aleatoric and epistemic uncertainty with random forests. In *International Symposium on Intelligent Data Analysis*, 2020.

[2] Abdar et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021.

[3] Amini et al. Deep evidential regression. *NeurIPS*, 2020.

[4] Blundell et al. Weight uncertainty in neural network. In *ICML*, 2015.

[5] Depeweg et al. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *ICML*, 2018.

[6] Egele et al. Autodeuq: Automated deep ensemble with uncertainty quantification. In *ICPR*, 2022.

[7] Gal et al. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[8] Hüllermeier et al. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 2021.

[9] Lakshminarayanan et AL. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30, 2017.

[10] Liu et al. Accurate uncertainty estimation and decomposition in ensemble learning. *NeurIPS*, 2019.