Could We Relieve AI/ML Models of the Responsibility of Providing Dependable Uncertainty Estimates? A Study on Outside-Model Uncertainty Estimates

Lisa Jöckel[1] and Michael Kläs[1]

[1] Fraunhofer Institute for Experimental Software Engineering IESE,
Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
{lisa.joeckel, michael.klaes}@iese.fraunhofer.de

Abstract. Improvements in Artificial Intelligence (AI), especially in the area of neural networks, have led to calls to use them also in the context of safety-critical systems. However, current AI-based models are data-driven, so we cannot assure that they will provide the intended outcome for any input. To obtain information about the uncertainty remaining in their outcome, uncertainty estimation capabilities can be integrated during model building. However, the approach of providing accurate outcomes and dependable uncertainty estimates using the same model has limitations. Among others, estimates of such 'in-model' approaches are provided without statistical confidence, tend to be overconfident if not calibrated, and are hard to interpret and review by domain experts. An alternative 'outside-model' approach is the use of model-agnostic uncertainty wrappers (UWs). To investigate how well they perform in comparison to in-model approaches, we benchmarked them against deep ensembles, which can be considered the gold standard for in-model uncertainty estimation, as well as to the softmax outputs of a deep neural network as a baseline. Despite a slightly higher Brier score, the UW provides other benefits that are important in a safety-critical context, like considering a statistical confidence level and providing explainable uncertainty estimates through a decision tree considering human-interpretable semantic factors. Furthermore, in-model uncertainty estimates can be forwarded into an UW, combining advantages of both approaches.

Keywords: Uncertainty Wrapper, Data-Driven Model, Machine Learning, Benchmarking Study, Traffic Sign Recognition, Automated Driving, Deep Ensemble, Uncertainty Calibration, Uncertainty Quantification.

1 Introduction

The use of Machine Learning (ML) and other Artificial Intelligence (AI) approaches can provide solutions for tasks that are difficult to tackle with traditional software development approaches. In recent years, in particular, deep neural networks have massively improved the performance of various tasks related to perception and understanding [1] [2]. Thus, it is not surprising that there is an intention to use modern ML approaches also in the context of systems with safety requirements as they promise new or massively improved functionalities. However, it is still open how best to deal with

components that rely on data-driven models (DDMs) such as deep neural networks, in a safety context considering their hard-to-predict behavior.

One major issue is that we can neither assume nor demonstrate that such data-driven components will provide the "correct" outcome for any input. Contrary to traditional components, the behavior of data-driven components is specified by example data and thus we need to live with a certain degree of uncertainty when using them.

Although uncertainty estimation is an active research area in AI, many DDMs do not provide dependable uncertainty estimates. Research in this area is dominated by benchmarks and proposals for ML methods that lead to DDMs that provide uncertainty estimates together with their main outcome. Yet, we see – especially from a safety perspective – limitations with these kinds of 'in-model' uncertainty estimates [3]. In particular, they violate the 'separation of concerns' principle [4] since they make the DDM itself responsible for providing dependable assessments of its performance.

In the best case, this leads to black-box uncertainty estimates that cannot be checked for plausibility by experts. Commonly, however, such estimates also ignore features that would help to provide better uncertainty estimates because these features do not improve the accuracy of the primary model outcome. For example, knowledge about the amount of precipitation does not help to decide whether a camera picture shows a traffic sign of type A or B, but it helps to assess the uncertainty in the provided result. In the worst case, the bad practice of providing uncertainty estimates that are either not calibrated or calibrated on the data used to train the model can make these estimates systematically overconfident.

To provide an alternative to existing 'in-model' uncertainty estimation approaches, Kläs and Sembach proposed the 'uncertainty wrapper' (UW) concept as an 'outside-model' approach [5]. The concept of model-agnostic UWs does not only avoid the above issues, but also allows addressing all three types of uncertainty sources considered in the onion shell model [6], namely model fit, data quality, and scope compliance. Furthermore, a confidence level can be set for the provided uncertainty estimates, which we consider essential for the use in a safety-critical application.

In previous work, Kläs and Jöckel illustrated how UWs can be applied in the context of pedestrian detection [7]. Compared to a naïve baseline approach, the UW provided more dependable uncertainty estimates improving all three components of the Brier score, a common measure for the quality of probabilistic predictions [8] [9].

Although the 'outside-model' uncertainty estimates of UWs have advantages from a conceptual perspective, it is still open how well UWs perform with respect to estimation quality in comparison to state-of-the-art 'in-model' approaches. Moreover, it is open how well UWs deal with limitations commonly observed in practices and whether it is possible to leverage synergies between 'in-model' and 'outside-model' approaches. To address these questions, this paper presents an experimental study in which DDMs with and without an UW were benchmarked under different settings.

The remainder of this paper is structured as follows: Section 2 gives an overview of related work on uncertainty estimation including some background on the concept of UWs. Section 3 elaborates the addressed research questions and introduces the study design and the benchmarking metrics we used. Section 4 presents and discusses the study results, and Section 5 concludes the paper.

2 Related Work on Uncertainty Predictions

Uncertainty is a topic of increasing relevance in the field of ML. The objective is to better understand the sources of uncertainty and provide dependable information on how much we can rely on an outcome given by a specific DDM.

How the concept of uncertainty as discussed in ML is related to safety standards, such as IEC 61508, and the application rule VDE-AR-E 2842-61 is detailed by Kläs et al. in their recent work [10], in which they also illustrate how the UW pattern can help to handle uncertainty in compliance with safety constraints.

Sources of uncertainty. A number of classifications have been proposed for potential sources of uncertainty. The best-known is probably the distinction between aleatoric and epistemic uncertainty. In general, *aleatoric* uncertainty means uncertainty due to randomness, which is not systematic but rather unavoidable "noise". *Epistemic* uncertainty, on the other hand, means uncertainty that is systemic in the way it refers to phenomena that could be known in principle but are not considered [11]. The idea is that we need to accept aleatoric uncertainty but should try to reduce epistemic uncertainty, e.g., by collecting more and better data. In a concrete setting, however, the distinction which part of the uncertainty is aleatoric and which is epistemic depends on the viewpoint [11]. Moreover, quantifications are usually not comparable between modeling approaches due to different interpretations and hypothesis spaces.

A classification that is orthogonal to aleatoric and epistemic is proposed by the *onion shell model* [6], which distinguishes between model fit, input quality, and scope compliance uncertainty. It allows mathematically separating uncertainty attributed to (a) limitations in the DDM, (b) differences in the quality of the model input, and (c) the possibility that the model is applied outside its application scope [5]. Type (a) can be reduced by improving the DDM and can be measured by traditional model testing and performance metrics (e.g., mean absolute error or true positive rate). Type (b) can be tackled by modeling the influence of input quality on the quality of the DDM outcomes (e.g., how the performance changes given a certain input quality). Type (c) relies on defining the application scope and monitoring compliance (e.g., how similar the current situation is to the situations considered during model training and testing).

Outside-model estimations. Building upon this classification, an 'uncertainty wrapper' (UW) framework has been proposed [7]. UWs provide 'outside-model' estimations for uncertainty following the separation of concerns principle. They are model-agnostic and consider the DDM they encapsulate as a black box (cf. Fig. 1). Factors that may influence input quality and scope compliance uncertainty are modeled in a quality and a scope model, respectively. In the context of traffic sign recognition (TSR), e.g., the obstruction of vision by rain or fog would represent quality factors and the GPS coordinates could indicate as a scope factor whether the model is being applied outside its target application scope (e.g., a specific country). Using a decision-tree-based approach, the quality impact model decomposes the target application scope into areas with similar uncertainties based on the quality factors and safeguarded with a statistical confidence level that can be freely chosen. The scope compliance model uses the data

provided by the scope factors to calculate the probability that the DDM is applied outside its target application scope. This can include checking some fixed boundaries as well as calculations on similarity between the current input and the inputs considered during model development.

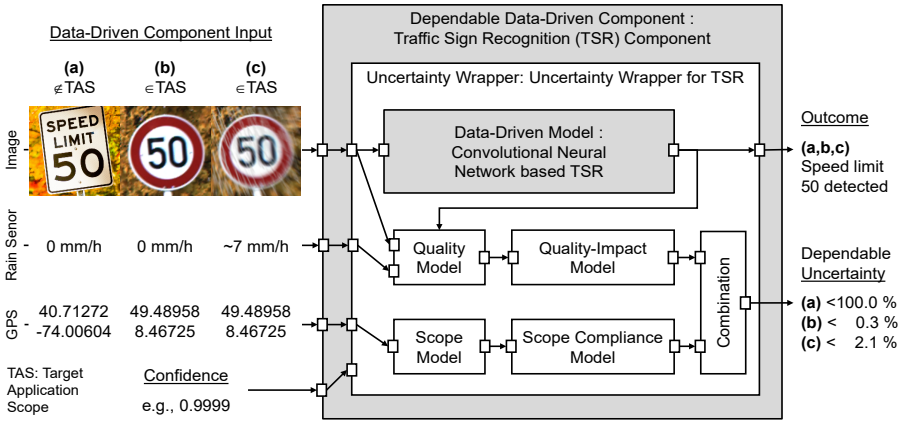

Fig. 1. Uncertainty Wrapper architecture together with sample inputs and outputs. [12]

In-model estimations. To date, the more established way to realize uncertainty estimation is to design the DDM itself such that it returns not only its categorical or binary outcome, but also the probability *p* that the provided outcome is correct. In the following, we will refer to approaches following this pattern as *'in-model'* approach.

Understanding *uncertainty* as the likelihood that the DDM outcome is not correct, the most naïve approach to estimate uncertainty is to determine the *overall error rate* on a sample as the global uncertainty estimate for all outcomes. Obviously, this estimation approach considers only model fit uncertainty.

There are classes of DDMs (e.g., logistic regression, Naïve Bayes, support vector machines) that provide by default a *preference value* between zero and one in addition to their outcome. Although these values are commonly interpreted as uncertainty estimates, there are limitations: Preference values do commonly nor represent real probabilities and are determined on training data, which favors overconfidence [3].

To address these limitations, *calibration methods* such as isotonic regression and Platt scaling [13] are applied as a kind of post-processing on the preference values. This, however, cannot solve the limitation that information sources are ignored that are only relevant for the uncertainty but not for the outcome.

Out-of-distribution and *novelty detection* methods [14] such as SafeML [15] provide means to detect whether a DDM is applied outside its intended application context. However, they are also limited to scope-compliance-related uncertainty, which on the other hand is largely ignored by the 'in-model' approaches discussed above.

In the context of deep neural networks, which are our focus, Bayesian neural networks [16] and deep ensembles [17] [16] are commonly proposed to provide uncertainty estimates. Since both approaches are computationally expensive, they are often approximated by using Monte Carlo Dropout [16]. Benchmarks and comparisons of

state-of-the-art approaches for 'in-model' uncertainty estimation indicate that deep ensembles are currently the gold standard for neural networks considering estimation performance [18] [19]. However, their computational demands increase linearly with the number of ensemble members during training and operation, which commonly challenges current hardware [16] [18].

3 Study Planning and Execution

This chapter introduces and concretizes the addressed research questions, presents the derived study design, and explains the study execution.

3.1 Research Questions

A key question regarding the use of outside-model uncertainty estimation approaches is whether such model-agnostic approaches, which make no assumption about the internals of the encapsulated DDM, can achieve an estimation performance comparable to in-model uncertainty estimation approaches. In particular, we want to compare UW-based estimates with estimates based on existing in-model uncertainty estimation approaches that are either state-of-the-practice or state-of-the-art:

RQ1: How does the uncertainty estimation performance of UWs and DDMs differ when trained and calibrated on data with sufficient examples of quality deficits?

Instead of considering in-model and outside-model approaches as competitors, we can also think about means to leverage synergies between the two kinds of approaches:

RQ2: Can the uncertainty estimation performance of an UW be improved by also considering the uncertainty estimates of the DDM as a factor, besides other factors?

In practice, we commonly need to deal with imperfections. Thus, we investigate how the UW approach performs under conditions that can occur in real-world applications:

What are the implications if…
RQ3.1: no uncertainty factors with semantic meaning can be defined?
RQ3.2: the DDM is not calibrated?
RQ3.3: the DDM is trained on data that insufficiently covers relevant quality deficits?
RQ3.4: a state-of-the-art DDM is not used due to limited computational resources?

Answering these questions can provide important insights for practical applications. RQ3.1 is linked to the question of whether it may be reasonable to build an UW even if no information sources on uncertainty other than the DDM are available. Answering RQ3.2 can help to decide whether a DDM has to be calibrated if we want to consider its uncertainty estimates as a factor in an UW. RQ3.3 investigates possible benefits of encapsulating a DDM that has blind spots on the quality issues it will face during its usage. Finally, RQ3.4 may help to identify cases where we can substitute a deep ensemble, i.e., the gold standard, with a more resource-efficient DDM.

3.2 Study Design and Variation Points

This section describes the context, design, and decisions in our study execution plan as summarized in Fig. 2 and detailed in Section 3.3. First, we will introduce the task of the DDM and the target application scope to which we refer in our evaluation. Next, we will describe the datasets used in the study and the augmentation used to enrich available datasets with quality deficits.

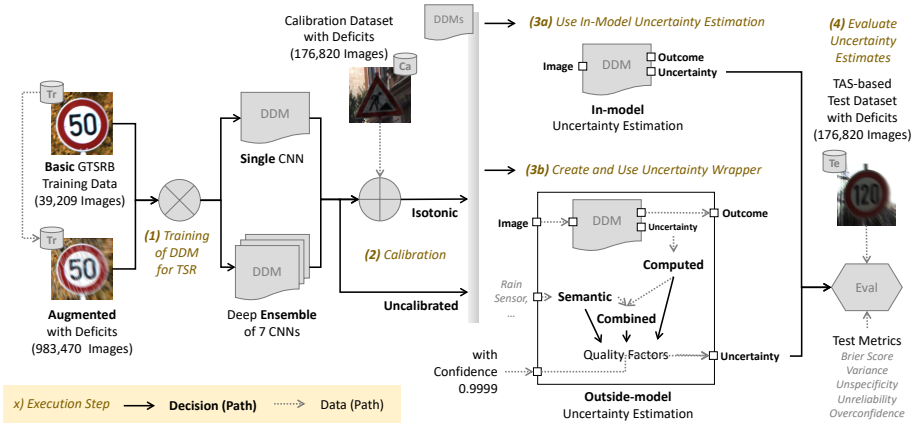

Fig. 2. Summary on study execution plan with execution steps, design decisions, and data flow.

Motivated by our research questions, Table 1 summarizes relevant variation points in the study design, including the respective decisions and planned comparisons.

Table 1. Overview on the decisions and planned comparisons for each research question.

Research Question	Training Data	DDM	Calibration Approach	UW Quality Factors	Comparison
RQ1	Augmented	Single + Ensemble	Isotonic	Semantic	***In-model vs. UW***
RQ2	Augmented	Single + Ensemble	Isotonic	***Combined***	vs. RQ1
RQ3.1	Augmented	Single + Ensemble	Isotonic	***Computed***	vs. RQ1 & RQ2
RQ3.2	Augmented	Single + Ensemble	***Uncalibrated***	Combined + Computed	vs. RQ2 & RQ3.1
RQ3.3	***Basic***	Single + Ensemble	Isotonic + Uncalibrated	All	vs. RQ1 to RQ3.2
RQ3.4	Augmented	***Single + Ensemble***	Isotonic + Uncalibrated	Semantic	Single vs. Ensemble

Task and Target Application Scope (TAS). We investigated our research questions using the example of traffic sign recognition, where the main task of the DDM is to correctly classify traffic signs on given images. The assumed target application scope was a roadworthy passenger car traveling in Germany at different points in time and faced with different weather conditions and related operation conditions, such as dirt

on the camera lens. Note that our interest in the study is not on the performance of the DDM with respect to its primary outcome but the provided uncertainty estimate.

Datasets. To address our research questions, we need different kinds of data: *training data* to build the DDM and to identify situations that differ in their degree of uncertainty; *calibration data* to derive unbiased uncertainty estimators on unseen data; and representative *test data* to evaluate the quality of the uncertainty estimates provided by the investigated in-model and outside-model uncertainty approaches.

The foundation of the datasets is the German Traffic Sign Recognition Benchmark (GTSRB) dataset [20], an established dataset with 51,839 images of German traffic signs annotated with sign type labels as ground truth (i.e., the intended outcome). Because the GTSRB data does not provide any further information that can be used to define semantic factors with a potential influence on uncertainty, e.g., location, time of day, or weather conditions, we had to augment the available data with a selection of realistic quality deficits and annotate them respectively.

To accomplish this, we applied the data augmentation framework proposed by Jöckel and Kläs [21], which allows augmenting an image with a selection of photorealistic quality deficits based on a given *situation setting*. About 2.7 million realistic situation settings were generated based on historical weather data from Deutscher Wetterdienst [22] and street locations within the TAS from OpenStreetMap [23].

In our study, we considered 9 types of *quality deficits* that could affect an image [24]: rain, darkness, haze, natural backlight, artificial backlight, dirt on the traffic sign, dirt on the sensor lens, steamed-up sensor lens, and motion blur. The intensity of each deficit was normalized to a scale between 0 (no effect) and 1 (maximum effect).

Each image of the original training dataset of GTSRB with 39,209 samples, which we refer to as the ***basic*** *training dataset*, was augmented for each quality deficit with low, medium, and high intensity by sampling from all appropriate situation settings and applying respective augmentations. This means each original image was augmented 9 x 3 = 27 times, provided sufficient situation settings where available. Besides the resulting 944,261 augmented and annotated images, the ***augmented*** *training dataset* additionally included the original 39,209 images of the basic training dataset.

The 12,630 images available in the GTSRB test dataset were randomly split into two equally sized, disjoint partitions for calibration and evaluation. To keep the original ratio between training and test data consistent, we augmented each original image 28 times based on settings randomly sampled from the 2.7 million realistic situation settings available. This resulted in 176,820 augmented samples each for the *calibration dataset* and the *evaluation dataset*. As the settings were generated based on the emulated target application scope, we assume that the random samples in this ***TAS-based*** *test dataset* has a distribution representative for the target application scope.

3.3 Study Execution

This section provides details on the four study execution steps illustrated in Fig. 2.

(1) Training of DDMs. To investigate our research questions, we considered two types of architectures for the DDMs: a single state-of-the-art convolution neural network

(CNN) architecture with a softmax output layer, and a deep ensemble architecture including multiple CNNs running in parallel to represent what can be considered the current gold standard for in-model uncertainty estimations.

For the ***single CNN***, we chose a model architecture roughly based on the model that currently performs best in the GTSRB [1] in a variant without spatial transformers and using batch normalization in combination with spatial dropout instead of local contrast normalization after each convolution layer. These modifications were mainly motivated by reducing the computation resources that are required to build and evaluate deep ensembles based on several of these CNNs.

For the ***deep ensemble***, we chose an architecture that combines multiple CNNs of the same architecture and with the same hyper-parameter settings as the single CNN but with different weight initializations during model training. Following the conclusions of Henne et al. [19], who examined the effect of the number of ensemble members on the quality of uncertainty estimates, we decided to use seven ensemble members, which were identified as an adequate number in their study.

Motivated by our research questions, each DDM was trained either with the basic or the augmented training dataset. For the basic training dataset, prediction accuracy stabilized on holdout validation data for the trained CNNs around 0.995 after 30 epochs. Due to the much larger number of samples in our augmented training dataset, accuracy stabilized there after only 10 epochs at around 0.891.

(2) Calibration of DDMs. In uncertainty estimation research, calibration is seen as an important post-processing technique for in-model approaches to make their uncertainty estimates more reliable. In our study, we decided to apply the scikit-learn implementations of Isotonic Regression and Platt's logistic model, which are both model-agnostic, well-established calibration approaches, using the calibration dataset we had prepared for this purpose. When we subsequently talk about ***calibrated*** DDMs, we report the result based on Isotonic Regression, since Isotonic Regression consistently outperformed Platt's logistic model on our datasets. DDMs using a deep ensemble architecture are calibrated by calibration of the final outputs after combining the preference values of all ensemble members as proposed by [23].

(3a) In-model uncertainty estimation. In cases where we used in-model uncertainty estimates, the (calibrated) CNN or deep ensemble does not only report the predicted traffic sign type but also the uncertainty. In these cases, the uncertainty estimate is one minus the ultimately calibrated (and aggregated) preference value(s) as calculated by the softmax-layer(s) of the CNN(s).

(3b) Outside-model uncertainty estimation. Motivated by our research questions, we built UWs with three different sets of quality factors as input.

The default variant includes only ***semantic*** quality factors, meaning we consider as input the bounding box size around the detected traffic sign (assuming that images with fewer pixels make the task more difficult), the category of the predicted traffic sign (assuming that traffic signs of certain categories are more difficult to distinguish), and all nine types of augmented deficits (including rain, darkness, etc.).

The ***combined*** variant uses as an additional quality factor the 'in-model' uncertainty estimates computed by the encapsulated black-box DDM. Depending on the DDM, these uncertainty estimates can be raw softmax values or can be calibrated.

The third variant, the ***computed*** one, only considers quality factors as input that can be derived directly – without any additional information source – from the DDM input or output. In our setting, this comprises the size of the bounding box around the detected traffic sign, the DDM predicted traffic sign category and uncertainty.

The quality impact models of all UWs were trained as a decision tree built with CART algorithm optimized based on entropy with no pruning while training. After training, the quality impact models were calibrated on the calibration dataset considering a confidence level of 0.9999 and pruning all leaves containing less than 200 data samples in the case of UWs considering only semantic factors, and 700 otherwise (based on the results of a grid search).

Since our investigation focused on uncertainty related to input quality, which can be considered as the key strength of in-model uncertainty estimation approaches, a scope compliance model was not included and data points are considered inside TAS.

(4) Evaluation of uncertainty estimates. All uncertainty estimation approaches were evaluated on the TAS-based test dataset. To measure the uncertainty estimation performance, we computed the Brier score (bs), which measures the mean squared difference between the predicted probability of an outcome and the actual outcome [8]. The Brier score can be decomposed into variance (var), resolution (res), and unreliability (unr) [9] with $bs = var - res + unr$. A high *variance* corresponds to a high error rate of the DDM, i.e., more overall uncertainty. *Resolution* describes how much the case-specific uncertainty estimates differ from the overall uncertainty. As the res is bounded by the var and higher res values are better, we report instead $var - res$ as *unspecificity*. Finally, *unreliability* measures how well the estimated uncertainty is calibrated to the observed error rate of the DDM, i.e. smaller unreliability means better calibration. Additionally, we report as a metric of *overconfidence* the part of the unreliability attributed to uncertainty estimates that underestimate the observed error rate, which is the more serious case in a safety-critical setting.

4 Study Results and Discussion

We organized this section along the identified research questions RQ1 to RQ3, for which we will first present and then discuss the obtained evaluation results.

4.1 RQ1: Comparing UW performance with in-model approaches

This section addresses the question of how uncertainty estimation performance between UWs and DDMs differs when they were trained and calibrated on appropriate data with good coverage of potential quality deficits. Accordingly, Table 2 presents the evaluation metrics for both DDM architectures – a *single* CNN and deep *ensemble* of CNNs – in comparison to the performance of UWs encapsulating these DDMs and relying in their estimates instead on *semantic* factors.

As Table 2 shows, the UWs based on semantic factors performed worse than the corresponding in-model approaches if we consider the Brier score as a global measure of performance. The main reason can be seen in their higher unspecificity, which is not completely compensated by their improved reliability. Please note that unreliability for the UWs was calculated considering a confidence level of .9999, which results in an intended unreliability increase as overconfident estimates are penalized. Had we not demanded this confidence level, the unreliability of the UWs would be two magnitudes lower (single = 0.00043, ensemble = 0.00039). Overconfidence was strongly decreased by UWs compared to in-model approaches.

Interpretation: In settings where appropriate training data is available, using UWs seems to be a trade-off between (a) reducing the resolution of uncertainty estimates and (b) obtaining higher interpretability, low overconfidence based on defined confidence levels, and separation of concerns through an outside-model approach.

Table 2. Study results on the performance of in-model approaches and UWs.

Research Question	DDM Archit.	U. Estimation Approach	Brier Score	Vari-ance	Un-speci-ficity	Unreli-ability	Over-confi-dence
Baseline	Single	In-model	.09048	.19553	.00033	.09016	4.5e-02
RQ1		UW/*semantic*	.14931	.19553	.14630	.00301	2.4e-07
RQ2		UW/*combined*	.09065	.19553	.09018	.00048	8.9e-08
RQ3.1		UW/*computed*	.09076	.19553	.09048	.00028	1.6e-08
Baseline	Ensemble	In-model	.08584	.18696	.00034	.08550	3.8e-02
RQ1		UW/semantic	.14501	.18696	.14236	.00264	9.6e-07
RQ2		UW/*combined*	.08585	.18696	.08549	.00037	1.0e-06
RQ3.1		UW/*computed*	.08594	.18696	.08563	.00031	1.0e-06

Fig. 3 provides an impression of the human interpretability of UW due to the use of semantic factors and their decision tree structure. For instance, the decision paths can be checked for plausibility (e.g., whether it is reasonable to assume that traffic signs are harder to classify if they are covered by dirt).

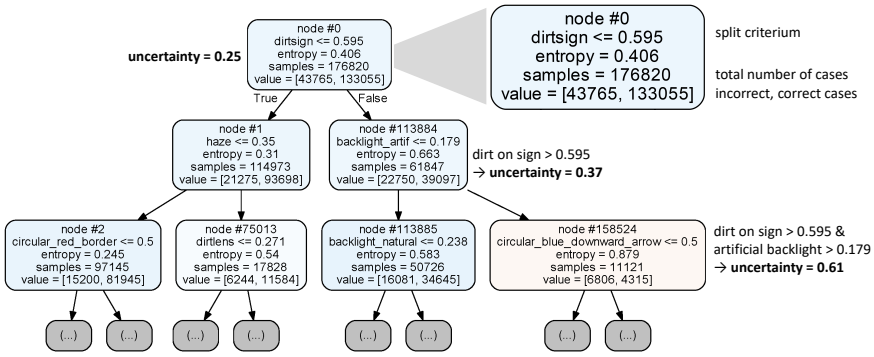

Fig. 3. Calibrated decision tree as part of an UW considering semantic factors.

4.2 RQ2: Synergies between in-model and outside-model approaches

This section addresses the question whether the uncertainty estimation performance of UWs can be improved if the in-model uncertainty estimates of the DDM are also considered as a quality factor, in addition to semantic factors such as precipitation.

The results in Table 2 show that if semantic factors are *combined* with in-model uncertainty estimates as an additional quality factor, the Brier score of the UWs almost reaches the level of the in-model approaches. The combination of in-model uncertainty estimates to the semantic factors not only strongly decreases the unspecificity but also further improves the reliability of the UW estimates.

Interpretation: Considering in-model uncertainty estimates as a factor can improve the uncertainty estimation performance of a UW up to the level of current gold standard in-model predictions. However, this performance improvement comes at the cost of decreased interpretability because a non-semantic "black-box" factor is introduced. Yet, other advantages are preserved, including confidence and separation of concerns.

4.3 RQ3: Performance under common, less than optimal conditions

This section investigates questions regarding the performance of UWs when applied under less than optimal conditions as they commonly occur in real-world settings.

RQ3.1. The results in Table 2 show that the uncertainty estimation performance if no semantic, but only *computed* factors are available in the UW reaches nearly the performance of considering the *combined* factors.

Interpretation: From a performance perspective, UWs with only computed quality factors work surprisingly well. However, we have to remember that neglecting semantic factors reduces interpretability, which is an important advantage of UWs. Thus, we conclude that UWs can also be applied when no semantic factors are available. However, developers should do as much as reasonably practical to assure interpretability, which includes considering as many semantic factors as possible.

RQ3.2. Table 3 shows how estimation performance is affected when the uncertainty estimates of the DDM, which were also used as a quality factor in the combined and computed UW, were not calibrated.

In the case where uncertainty was estimated by the *single* CNN, the Brier score became worse if the DDM was not calibrated. However, if the uncertainty estimates of the uncalibrated DDM were used instead as a quality factor of an UW, we observed no comparable negative impact on the performance of the encapsulating UW. The Brier score of the encapsulating UW (with *combined* and *computed* factors) was even better than the score of the encapsulated uncalibrated as well as calibrated single CNN. The last statement does also hold for the case that the DDM is the considered *deep ensemble*, which appears to have performed quite well even if not calibrated, but to a smaller magnitude.

Interpretation: Encapsulating an uncalibrated DDM within an UW seems to influence the components of the Brier score in a similar way as the calibration of the DDM. When DDMs are used encapsulated by an UW, the calibration of the DDM may thus not be required since its calibration has minor relevance for the performance of the UW.

Table 3. Study results on using uncalibrated instead of calibrated DDMs.

DDM Archit.	DDM Calibration	U. Estimation Approach	Brier Score	Vari-ance	Un-speci-ficity	Unreli-ability	Over-confi-dence
Single	Isotonic	In-model	.09048	.19553	.00033	.09016	4.5e-02
	-	In-model	.09481	.19946	.00155	.09326	6.1e-02
	Isotonic	UW/*semantic*	.14931	.19553	.14630	.00301	2.4e-07
	-	UW/*semantic*	.15071	.19946	.14755	.00316	1.6e-07
	Isotonic	UW/*combined*	.09065	.19553	.09018	.00048	8.9e-08
	-	UW/*combined*	.09013	.19946	.08967	.00046	0.0
	Isotonic	UW/*computed*	.09076	.19553	.09048	.00028	1.6e-08
	-	UW/*computed*	.09020	.19946	.08987	.00033	0.0
Ensemble	Isotonic	In-model	.08584	.18696	.00034	.08550	3.8e-02
	-	In-model	.08547	.18941	.00050	.08497	3.6e-02
	Isotonic	UW/*semantic*	.14501	.18696	.14236	.00264	9.6e-07
	-	UW/*semantic*	.14492	.18941	.14202	.00289	2.8e-07
	Isotonic	UW/*combined*	.08585	.18696	.08549	.00037	1.0e-06
	-	UW/*combined*	.08536	.18941	.08498	.00038	0.0
	Isotonic	UW/*computed*	.08594	.18696	.08563	.00031	1.0e-06
	-	UW/*computed*	.08540	.18941	.08503	.00038	0.0

RQ3.3. Table 4 summarizes the study results addressing the question of how the uncertainty estimation performance of (un-)calibrated DDMs and encapsulating UWs was affected when the DDMs were trained on data that did not sufficiently cover relevant quality deficits.

Table 4. Results for DDMs trained on a basic dataset insufficiently covering quality deficits.

DDM Archit.	Training Data	U. Estimation Approach	Brier Score	Vari-ance	Unspec-ificity	Unreli-ability	Over-conf.
Single	augmented	In-model/*cal.*	.09048	.19553	.00033	.09016	4.5e-02
	basic	In-model/*cal.*	.12116	.24025	.00001	.12116	4.3e-02
	augmented	In-model/*uncal.*	.09481	.19946	.00155	.09326	6.1e-02
	basic	In-model/*uncal.*	.45116	.23491	.03360	.41756	4.1e-01
	augmented	UW/*semantic*	.14931	.19553	.14630	.00301	2.4e-07
	basic	UW/*semantic*	.16438	.24025	.16127	.00311	1.6e-07
	augmented	UW/*combined*	.09065	.19553	.09018	.00048	8.9e-08
	basic	UW/*combined*	.11620	.24025	.11545	.00075	0.0
	augmented	UW/*computed*	.09076	.19553	.09048	.00028	1.6e-08
	basic	UW/*computed*	.12116	.24025	.12041	.00075	0.0
Ensemble	augmented	In-model/*cal.*	.08584	.18696	.00034	.08550	3.8e-02
	basic	In-model/*cal.*	.12152	.24885	.00000	.12152	3.8e-02
	augmented	In-model/*uncal.*	.08547	.18941	.00050	.08497	3.6e-02
	basic	In-model/*uncal.*	.14389	.24668	.00100	.14290	1.0e-01
	augmented	UW/*semantic*	.14501	.18696	.14236	.00264	9.6e-07
	basic	UW/*semantic*	.17356	.24885	.17058	.00298	3.0e-06
	augmented	UW/*combined*	.08585	.18696	.08549	.00037	1.0e-06
	basic	UW/*combined*	.11498	.24885	.11443	.00055	7.2e-08
	augmented	UW/*computed*	.08594	.18696	.08563	.00031	1.0e-06
	basic	UW/*computed*	.11887	.24885	.11826	.00061	0.0

If the training dataset insufficiently covers relevant quality deficits, as the *basic* training dataset did in our setting, the uncertainty estimation performance as measured by the Brier score gets worse. This applies in particular when the DDM is additionally not calibrated on representative data. The worse Brier score results we obtained were caused by higher variance, higher unspecificity (except calibrated in-model approaches) and higher unreliably. If the in-model uncertainty estimates are calibrated, using an ensemble instead of a single CNN, does not improve the estimation results. The observed performance decrease was less noticeable for UWs compared to in-model uncertainty estimation approaches.

Interpretation: The training dataset is an important ingredient for obtaining reliable uncertainty estimates, especially for in-model approaches. In opposite to the case of uncalibrated DDMs, for which using an ensemble may compensate a missing calibration, in the case of calibrated DDMs that are trained on data insufficiently covering relevant quality deficits using an ensemble seems not to be an effective countermeasure. The observed negative effect of training data can however be partially mitigated by encapsulating the DDM within an UW or at least calibrating it.

RQ3.4. The results in Table 2 to 4 also compare the uncertainty estimation performance obtained when a single CNN or a deep ensemble of CNNs is used. The numbers show an improved Brier score in almost all cases when a deep ensemble was considered. The magnitude of improvement was especially high if the DDM was not calibrated. If the DDM was encapsulated, the performance gap between single and deep ensembles was smaller.

Interpretation: The better performance of deep ensembles in comparison to single CNNs comes at the cost of higher resource consumption during development and operation, which in our setting was seven times higher. The degree to which a deep ensemble exceeds the performance of a single model, and hence its cost-benefit ratio, has to be assessed on the specific application setting.

5 Conclusion

We compared UWs as an outside-model approach with existing in-model uncertainty estimation approaches. Our study focused on uncertainty estimation performance and related overconfidence considering the example task of traffic sign recognition.

Summarizing our conclusions on the posed research questions, we offer the following preliminary advices. (A1) If general uncertainty estimation performance is the only criterion, we recommend using a state-of-the-art *in-model* approach such as deep ensembles. (A2) If, additionally, computational resources are a limiting factor, a simpler DDM can be an alternative if its uncertainty estimates are calibrated appropriately. (A3) If at least one of the following criteria is relevant in our setting, the use of an UW, as an *outside-model* approach, should be considered:

- separating the concerns of providing good outcome and uncertainty estimates
- assuring interpretability, e.g., to check the plausibility of uncertainty estimates
- providing statistical guaranties based on a given confidence level
- facing a DDM trained on data insufficiently covering relevant quality deficits

- scope compliance is not guaranteed (e.g., causing out-of-distribution issues)

If encapsulating a DDM within a UW, (A4) calibrating the DDM does not appear to be necessarily required, (A5) the kinds of factors that are used in a UW – semantic, computed, or both – is a trade-off decision that should consider the specific needs regarding interpretability, available data, and estimation performance. In a safety-driven setting, we would recommend using semantic factors to the extent that is reasonably practical to keep the uncertainty estimates as transparent as possible.

In summary, we see good reasons to release AI models from the responsibility of providing dependable uncertainty estimates, and, considering the presented results, an uncertainty wrapper can be an option for realizing the required separation of concerns.

For the future, we plan to conduct additional studies to investigate the usefulness of the UW pattern in different settings and integrate it in a structured safety argument.

Acknowledgments. Parts of this work have been funded by the Observatory for Artificial Intelligence in Work and Society (KIO) of the Denkfabrik Digitale Arbeitsgesellschaft in the project "KI Testing & Auditing".

6 References

1. Á. Arcos-García, J. Alvarez-Garcia and L. Soria Morillo, "Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods," *Neural Networks*, no. 99, 2018.
2. A. Garcia-Garcia, S. Orts, S. Oprea, V. Villena Martinez and J. Rodríguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation," arXiv:170406857, 2017.
3. M. Kläs, "Towards Identifying and Managing Sources of Uncertainty in AI and Machine Learning Models - An Overview," arXiv:1811.11669, 2018.
4. E. W. Dijkstra, "On the role of scientific thought," in *Selected writings on Computing: A Personal Perspective*, New York, USA, Springer, 1982, p. 60–66.
5. M. Kläs and L. Sembach, "Uncertainty wrappers for data-driven models – Increase the transparency of AI/ML-based models through enrichment with dependable situation-aware uncertainty estimates," WAISE 2019.
6. M. Kläs and A. M. Vollmer, "Uncertainty in Machine Learning Applications – A Practice-Driven Classification of Uncertainty," WAISE 2018.
7. M. Kläs and L. Jöckel, "A Framework for Building Uncertainty Wrappers for AI/ML-based Data-Driven Components," WAISE 2020.
8. G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.
9. A. H. Murphy, "A new vector partition of the probability score," *Journal of Applied Meteorology*, vol. 12, no. 4, p. 595–600, 1973.
10. M. Kläs, R. Adler, I. Sorokos, L. Joeckel, J. Reich, "Handling Uncertainties of Data-Driven Models in Compliance with Safety Constraints for Autonomous Behaviour," in European Dependable Computing Conference (EDCC), 2021. *(accepted for publication)*
11. A. Der Kiureghian and O. Ditlevsen, "Aleatory or Epistemic? Does It Matter?," *Structural Safety*, vol. 31, no. 2, pp. 105-112, 2009.
12. T. Bandyszak, L. Jöckel, M. Kläs, S. Törsleff, T. Weyer, B. Wirtz, "Handling Uncertainty in Collaborative Embedded Systems Engineering," *Model-Based Engineering of Collaborative Embedded Systems*, Springer, p. 147-170, 2021.

13. C. Guo, G. Pleiss, Y. Sun and K. Weinberger, "On Calibration of Modern Neural Networks," in ICML 2017.
14. M. Pimentel, D. Clifton, L. Clifton and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, p. 215–249, 2014.
15. K. Aslansefat, I. Sorokos, D. Whiting, R. Tavakoli Kolagari and Y. Papadopoulos, "SafeML: Safety Monitoring of Machine Learning Classifiers Through Statistical Difference Measures," IMBSA 2020.
16. F. Arnez, H. Espinoza, A. Radermacher and F. Terrier, "A Comparison of Uncertainty Estimation Approaches in Deep Learning Components for Autonomous Vehicle Applications," AISafety 2020.
17. B. Lakshminarayanan, A. Pritzel and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," NIPS 2017.
18. F. Gustafsson, M. Danelljan and T. Schön, "Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision," CVPR 2020.
19. M. Henne, A. Schwaiger, K. Roscher and G. Weiss, "Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics," SafeAI 2020.
20. "German Traffic Sign Benchmarks," http://benchmark.ini.rub.de/?section=gtsrb. [2021].
21. L. Jöckel and M. Kläs, "Increasing trust in data-driven model validation," SafeComp 2019.
22. "Climate Data Center," https://cdc.dwd.de/portal/.[13.11.2020].
23. "OpenStreetMap," https://www.openstreetmap.de/. [13.11.2020].
24. L. Jöckel, M. Kläs and S. Martínez-Fernández, "Safe Traffic Sign Recognition through Data Augmentation for Autonomous Vehicles Software," in QRS 2019.
25. R. Rahaman and A. Thiery, "Uncertainty Quantification and Deep Ensembles," arXiv:2007.08792, 2020.