Uncertainty Quantification for Machine Learning: One Size Does Not Fit All

Paul Hofman[1,2], **Yusuf Sale**[1,2], **Eyke Hüllermeier**[1,2,3]
[1]LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]German Research Center for Artificial Intelligence (DFKI, DSA)
paul.hofman@ifi.lmu.de, yusuf.sale@ifi.lmu.de, eyke@ifi.lmu.de

Abstract

Proper quantification of predictive uncertainty is essential for the use of machine learning in safety-critical applications. Various uncertainty measures have been proposed for this purpose, typically claiming superiority over other measures. In this paper, we argue that there is no single best measure. Instead, uncertainty quantification should be tailored to the specific application. To this end, we use a flexible family of uncertainty measures that distinguishes between total, aleatoric, and epistemic uncertainty of second-order distributions. These measures can be instantiated with specific loss functions, so-called proper scoring rules, to control their characteristics, and we show that different characteristics are useful for different tasks. In particular, we show that, for the task of selective prediction, the scoring rule should ideally match the task loss. On the other hand, for out-of-distribution detection, our results confirm that mutual information, a widely used measure of epistemic uncertainty, performs best. Furthermore, in an active learning setting, epistemic uncertainty based on zero-one loss is shown to consistently outperform other uncertainty measures.

1 Introduction

Uncertainty quantification (UQ), the assessment of a model's uncertainty in predictive tasks, has become an increasingly prominent topic in machine learning research and practice. A common distinction is made between *aleatoric* and *epistemic* uncertainty (Hüllermeier and Waegeman 2021). Broadly speaking, aleatoric uncertainty originates from the inherent stochastic nature of the data-generating process, while epistemic uncertainty is due to the learner's incomplete knowledge of this process. The latter can therefore be reduced by acquiring additional information, such as more training data, whereas aleatoric uncertainty, as a characteristic of the data-generating process, is non-reducible.

Due to inherent challenges in representing epistemic uncertainty, higher-order formalisms, most notably second-order distributions (i.e., distributions over distributions), are typically employed. Given a suitable uncertainty *representation*, the key question that follows is how to appropriately quantify (total) uncertainty in terms of a numerical value, and how to decompose it into an aleatoric and an epistemic component. This choice has important consequences for downstream tasks, e.g., it determines which examples are abstained on, which inputs are flagged as out-of-distribution, or which unlabeled points are queried, and poor uncertainty quantification can thus mask true performance or misguide decisions even if the base predictor is strong.

For second-order uncertainty representations, entropy-based measures have long been the default choice (Depeweg et al. 2018). Yet recent work (Wimmer et al. 2023) questions whether these metrics truly satisfy the core criteria of sound uncertainty quantification. Consequently, exploring alternative uncertainty measures is a natural and necessary step toward overcoming the limitations of existing approaches. However, much of the prior work treats these alternatives as general competitors rather than asking which uncertainty measure is appropriate for a specific downstream objective, leading to conflicting or opaque conclusions.

Uncertainty measures in the machine learning literature have largely been treated as a one-size-fits-all solution, with little emphasis on adapting to specific tasks. However, recent work (Mucsányi, Kirchhof, and Oh 2024) suggests that different tasks may require tailored uncertainty measures. Crucially, in the absence of an observable baseline, uncertainty measures are typically evaluated empirically through (downstream) tasks such as selective prediction, out-of-distribution (OoD) detection, or active learning, each of which may require different uncertainty measures.

These considerations highlight the need for a more flexible approach to uncertainty quantification, one that aligns with the specific requirements of underlying tasks. Accordingly, leveraging a classical decomposition of proper scoring rules, we adopt a loss-based family of total, aleatoric, and epistemic uncertainty measures that subsumes traditional measures as a special case (Sale et al. 2024b; Hofman, Sale, and Hüllermeier 2024b; Kotelevskii et al. 2025).

While recent work has explored this family of measures theoretically and cast them as alternatives to the static entropy-based approach, it has largely overlooked the most critical factor: the downstream machine learning *task* used to (empirically) *evaluate* the entire uncertainty pipeline. By tying uncertainty measures directly to each task's evaluation loss, we demonstrate that instantiating an uncertainty measure with that same loss yields optimal alignment with the task's objectives. We demonstrate this both theoretically and empirically. Theoretically, we establish a formal connection

between task losses and the losses used to construct uncertainty measures, showing that optimal uncertainty quantification requires alignment between these components. Empirically, we validate our framework across important downstream tasks, including selective prediction, OoD detection, and active learning, confirming that different tasks benefit from different uncertainty measures. Together, these results expose why common one-size-fits-all practices can be misleading and motivate more deliberate, task-aware uncertainty evaluation.

2 Uncertainty in Machine Learning

In this paper, we consider a standard supervised learning setting, in which a learner is given access to a set of i.i.d. training data $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, where $\mathcal{X}$ denotes the instance space and $\mathcal{Y}$ the set of outcomes. We focus on the classification scenario, where $\mathcal{Y} = \{1, \dots, K\}$ consists of a finite set of class labels. Additionally, we denote by $\mathbb{P}(\mathcal{Y})$ the set of all probability measures on $\mathcal{Y}$, which can be identified with the $(K-1)$-simplex Δ_K. We consider a hypothesis space $\mathcal{H}$, where each hypothesis $h \in \mathcal{H}$ maps instances $\boldsymbol{x}$ to probability distributions on outcomes. For brevity, we write $h(\boldsymbol{x}) = \hat{\boldsymbol{\theta}}$ for the probabilistic prediction produced by the hypothesis $h \in \mathcal{H}$, where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_K) \in \Delta_K$. Similarly, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ denotes the *ground-truth* (conditional) probability distribution on the outcomes given a query instance $\boldsymbol{x} \in \mathcal{X}$. Finally, we denote the extended real number line by $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$.

Uncertainty Representation. A probabilistic model h predicts a probability distribution that captures *aleatoric* uncertainty about the outcome $y \in \mathcal{Y}$, but pretends full certainty about the distribution $\boldsymbol{\theta}$ itself. In order to represent *epistemic* uncertainty, we consider a Bayesian representation of uncertainty. Hence, we assume that we have access to a posterior distribution $q(h \mid \mathcal{D})$. The posterior distribution gives rise to a distribution over distributions $\boldsymbol{\theta}$ through

$$Q(\boldsymbol{\theta}) = \int_{\mathcal{H}} [\![h(\boldsymbol{x}) = \boldsymbol{\theta}]\!] \, dq(h \mid \mathcal{D}) \,,$$

with $Q \in \Delta_K^{(2)}$ and $\Delta_K^{(2)}$ denotes the set of all probability distributions on Δ_K (*viz.* second-order distributions). To make predictions, a representative first-order distribution is generated by model averaging $\bar{\boldsymbol{\theta}} = \int_{\mathcal{H}} h(\boldsymbol{x}) \, dq(h \mid \mathcal{D})$. In practice, we usually only have access to samples of the posterior, e.g., through an ensemble of predictors. Thus, we use a finite approximation $\bar{\boldsymbol{\theta}} = \frac{1}{M} \sum_{m=1}^M h^m(\boldsymbol{x})$, where M denotes the number of ensemble members or samples drawn from the posterior in the case of e.g. variational inference.

Uncertainty Quantification. Given a second-order distribution $Q \in \Delta_K^{(2)}$, the task of uncertainty quantification is to specify functionals (namely, uncertainty measures) $\mathrm{TU}, \mathrm{AU}, \mathrm{EU} : \Delta_K^{(2)} \to \mathbb{R}_{\geq 0}$ that quantify total, aleatoric, and epistemic uncertainty, respectively. Particularly, a well-known decomposition of proper scoring rules yields a theoretically principled family of uncertainty measures, flexibly instantiated by the choice of the loss function.

Proper scoring rules, originating in Savage's elicitation framework (Savage 1971) and developed further by Gneiting and Raftery (2007), assign numerical scores to probabilistic forecasts and incentivize *truthful* reporting. A scoring rule is proper if a forecaster's expected score is optimized exactly when the announced distribution equals their true belief, and strictly proper if this optimizer is unique. Further, a function $\mathcal{Y} \to \overline{\mathbb{R}}$ is called Δ_K-quasi-integrable if it is measurable with respect to $2^{\mathcal{Y}}$, and is quasi-integrable with respect to all $\boldsymbol{\theta} \in \Delta_K$. We assume scoring rules to be *negatively* oriented, thus taking a standard machine learning perspective where we wish to minimize the corresponding loss.

Definition (Proper scoring rule)**.** A scoring rule is a measurable function $\ell : \Delta_K \times \mathcal{Y} \to \overline{\mathbb{R}}$ such that, for all $\hat{\boldsymbol{\theta}} \in \Delta_K$, the expectation

$$L_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) := \mathbb{E}_{Y \sim \boldsymbol{\theta}} \left[\ell(\hat{\boldsymbol{\theta}}, Y) \right] \tag{1}$$

is well defined for every $\boldsymbol{\theta} \in \Delta_K$. The scoring rule ℓ is *proper* if, for all $\hat{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Delta_K$,

$$L_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}) \;\leq\; L_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}), \tag{2}$$

and *strictly proper* if equality in (2) holds only when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.

It is well-known that (strictly) proper scoring rules (and with them their corresponding expected losses) can be decomposed into a *divergence* term and an *entropy* term, respectively (Gneiting and Raftery 2007; Kull and Flach 2015):

$$D_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = L_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) - L_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}), \quad H_\ell(\boldsymbol{\theta}) = L_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}).$$

The latter captures the expected loss that materializes even when the ground truth $\boldsymbol{\theta}$ is predicted, whereas the former represents the "excess loss" that is caused by predicting $\hat{\boldsymbol{\theta}}$ and hence deviating from the optimal prediction $\boldsymbol{\theta}$. This decomposition naturally aligns with the distinction between *irreducible* (aleatoric) and *reducible* (epistemic) uncertainty: $H_\ell(\boldsymbol{\theta})$ is the irreducible part of the risk, and hence relates to aleatoric uncertainty, whereas $D_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ is purely due to the learner's imperfect knowledge, or epistemic state, and could in principle be reduced by additional information.

So far, our discussion has focused on first-order distributions, assuming access to the true conditional distribution $\boldsymbol{\theta}$. In practice, however, and as previously motivated, uncertainty about $\boldsymbol{\theta}$ is represented through a second-order distribution Q. Consequently, it is sensible to define

$$\begin{aligned} \mathrm{EU}(Q) &= \mathbb{E}_{\boldsymbol{\theta} \sim Q}[D_\ell(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta})] && (3) \\ &= \mathbb{E}_{\boldsymbol{\theta} \sim Q}[L_\ell(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) - L_\ell(\boldsymbol{\theta}, \boldsymbol{\theta})] && (4) \\ &= \underbrace{\mathbb{E}_{\boldsymbol{\theta} \sim Q}[L_\ell(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta})]}_{\mathrm{TU}(Q)} - \underbrace{\mathbb{E}_{\boldsymbol{\theta} \sim Q}[L_\ell(\boldsymbol{\theta}, \boldsymbol{\theta})]}_{\mathrm{AU}(Q)} . && (5) \end{aligned}$$

That is, EU is the *gain* — in terms of loss reduction — the learner can expect when predicting, not on the basis of the uncertain knowledge Q, but only after being revealed the true $\boldsymbol{\theta}$. Intuitively, this is plausible: The more uncertain the learner is about the true $\boldsymbol{\theta}$ (i.e., the more dispersed Q), the

more it can gain by getting to know this distribution. The connection to proper scoring rules is also quite obvious: Total uncertainty in (5) is the expected loss of the learner when predicting optimally ($\bar{\boldsymbol{\theta}}$) on the basis of its uncertain belief Q. It corresponds to the expectation (with regard to Q) of the expected loss (1). Broadly speaking, we average the score of the prediction $\bar{\boldsymbol{\theta}}$ over the potential ground-truths $\boldsymbol{\theta} \sim Q$. Aleatoric uncertainty is the expected loss that remains, even when the learner is perfectly informed about the ground-truth $\boldsymbol{\theta}$ before predicting. Again, we average over the potential ground-truths $\boldsymbol{\theta} \sim Q$. Epistemic uncertainty is the difference between the two, i.e., the expected loss reduction due to information about $\boldsymbol{\theta}$. When ℓ is taken to be a strictly proper scoring rule, (3) is also known as the Bregman information (Banerjee, Guo, and Wang 2004). We discuss three important loss-instantiations of the uncertainty measures:

(1) *Log loss:* Instantiating the uncertainty measures (5) with the log loss $\ell(\hat{\boldsymbol{\theta}}, y) = -\log(\hat{\theta}_y)$ yields

$$\underbrace{\mathrm{S}(\bar{\boldsymbol{\theta}})}_{\mathrm{TU}(Q)} = \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim Q}[\mathrm{S}(\boldsymbol{\theta})]}_{\mathrm{AU}(Q)} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim Q}[\mathrm{KL}(\boldsymbol{\theta} \parallel \bar{\boldsymbol{\theta}})]}_{\mathrm{EU}(Q)},$$

where $\mathrm{S}(\cdot)$ and $\mathrm{KL}(\cdot \parallel \cdot)$ denote the Shannon entropy and Kullback-Leibler divergence, respectively. Clearly, the log loss instantiations correspond to the entropy-based measures (Depeweg et al. 2018). Although these measures have been criticized (Wimmer et al. 2023), they remain the most commonly used uncertainty measures in the classification setting.

(2) *Brier loss* (or quadratic loss): Similarly, fixing $\ell(\hat{\boldsymbol{\theta}}, y) = \sum_{k=1}^{K} (\hat{\theta}_k - [\![k = y]\!])^2$ yields

$$\underbrace{1 - \sum_{k=1}^{K} \bar{\theta}_k^2}_{\mathrm{TU}(Q)} = \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim Q}\left[1 - \sum_{k=1}^{K} \theta_k^2\right]}_{\mathrm{AU}(Q)} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim Q} \sum_{k=1}^{K} (\bar{\theta}_k - \theta_k)^2}_{\mathrm{EU}(Q)}$$

The Brier loss (Brier 1950) is another strictly proper scoring rule, which is often used as a measure of calibration (Minderer et al. 2021; Clarté et al. 2023). The decomposition generates the measures of expected Gini impurity for aleatoric uncertainty. The Gini impurity quantifies the probability of misclassification when predicting randomly according to the ground-truth distribution, i.e. $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. The measure of epistemic uncertainty is the expected squared difference. This measure has also been proposed by Smith and Gal (2018).

(3) *Zero-one loss:* With $\ell(\hat{\boldsymbol{\theta}}, y) = 1 - [\![\arg\max_k \hat{\theta}_k = y]\!]$, we get the following instantiations:

$$\underbrace{1 - \max_k \bar{\theta}_k}_{\mathrm{TU}(Q)} = \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim Q}[1 - \max_k \theta_k]}_{\mathrm{AU}(Q)} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim Q}[\max_k \theta_k - \theta_{\arg\max_k \bar{\theta}_k}]}_{\mathrm{EU}(Q)}.$$

The aleatoric component is the expected complement of the confidence. Assuming $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, it quantifies the probability of misclassification when predicting the class with the highest probability. Quantifying uncertainty on the basis of the confidence of the model is common (Hendrycks and Gimpel 2017), but in a second-order representation, this measure has not been used before. Interestingly, this component aligns with the measure of aleatoric uncertainty proposed and axiomatically justified by Sale et al. (2024a). The epistemic component of this decomposition has, to the best of our knowledge, not been used in machine learning so far. It is minimized when all first-order distributions $\boldsymbol{\theta}$ in the support of the second-order distribution Q have the same argmax as the Bayesian model average $\bar{\boldsymbol{\theta}}$, where the support is defined as $\mathrm{supp}(Q) = \{\boldsymbol{\theta} \in \Delta_K : Q(\boldsymbol{\theta}) > 0\}$. Minimizing this component guarantees that every first-order distribution in the support of Q agrees on the class with highest probability (i.e., consensus on the most likely class), but it does not eliminate all epistemic uncertainty about the true conditional distribution $\boldsymbol{\theta}$. In particular, variability in the assigned probability mass, such as how confident the predictors are about that top class or how much weight is placed on runner-up classes, can persist. Put differently, the zero-one loss epistemic measure only captures label-level disagreement about which class is most likely and is blind to finer-grained uncertainty in the shape of the distributions; even when it is zero, Q may still be diffuse and reflect unresolved uncertainty about $\boldsymbol{\theta}$ beyond the predicted label.

Remark. Let $\ell : \Delta_K \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a strictly proper scoring rule with associated convex potential $G : \Delta_K \to \mathbb{R}$ such that $D_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = L_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) - L_\ell(\boldsymbol{\theta}, \boldsymbol{\theta})$ is the *Bregman divergence* induced by G. For a second-order distribution $Q \in \Delta_K^{(2)}$, let $\bar{\boldsymbol{\theta}} := \mathbb{E}_{\boldsymbol{\theta}\sim Q}[\boldsymbol{\theta}]$ denote the (Bayesian) model average. Then, the measure of epistemic uncertainty (3) is exactly a Jensen gap of the convex potential G, which measures how spread out the posterior over first-order predictions is:

$$\mathrm{EU}(Q) = \mathbb{E}_{\boldsymbol{\theta}\sim Q}\left[G(\boldsymbol{\theta})\right] - G\left(\bar{\boldsymbol{\theta}}\right).$$

This is easy to see: By the convex-potential representation of strictly proper scoring rules (Gneiting and Raftery 2007), $D_\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = G(\boldsymbol{\theta}) - G(\hat{\boldsymbol{\theta}}) - \langle \nabla G(\hat{\boldsymbol{\theta}}), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \rangle$. With the prediction $\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$ and taking the expectation over Q,

$$\begin{aligned}\mathrm{EU}(Q) &= \mathbb{E}[D_\ell(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})] \\ &= \mathbb{E}[G(\boldsymbol{\theta})] - G(\bar{\boldsymbol{\theta}}) - \left\langle \nabla G(\bar{\boldsymbol{\theta}}),\, \mathbb{E}[\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}] \right\rangle \\ &= \mathbb{E}[G(\boldsymbol{\theta})] - G(\bar{\boldsymbol{\theta}}),\end{aligned}$$

since $\mathbb{E}[\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}] = 0$. Thus, $\mathrm{EU}(Q) = 0$ iff Q is a point mass at $\bar{\boldsymbol{\theta}}$; more generally, EU is monotone in the convex order, if Q' is more dispersed than Q but has the same mean, then $\mathrm{EU}(Q') \geq \mathrm{EU}(Q)$. This characterization is epistemically appealing: EU grows precisely with posterior dispersion about the unknown ground-truth distribution and vanishes only when the learner's belief collapses to a single $\boldsymbol{\theta}$, i.e., when no epistemic uncertainty remains.

3 Customized Uncertainty Quantification

Throughout, we write $\mathcal{L}(\Delta_K, \mathcal{Y})$ to denote the collection of all proper scoring rules $\ell : \Delta_K \times \mathcal{Y} \to \overline{\mathbb{R}}$. Moreover, let U_ℓ

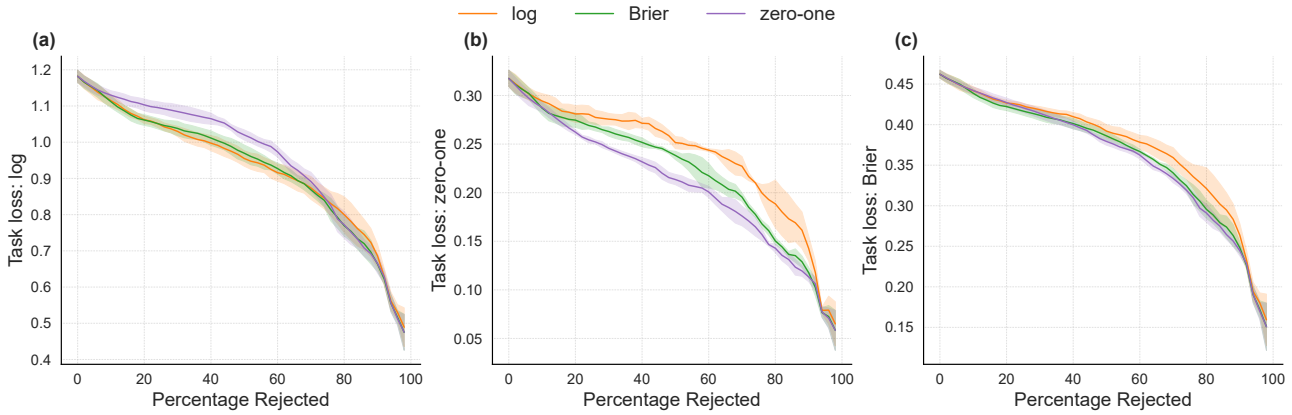

Figure 1: Selective prediction with different task losses using the total uncertainty component as the rejection criterion where **(a)** uses the *log loss* as task loss, **(b)** *zero-one loss*, and **(c)** the *Brier loss*, respectively. *Results are averaged over three runs.*

denote a mapping $\Delta_K^{(2)} \longrightarrow \mathbb{R}_{\geq 0}$, where $\ell \in \mathcal{L}(\Delta_K, \mathcal{Y})$.

Task Loss vs. Uncertainty Loss. In real-world settings, we do not measure uncertainty purely for its own sake, but to understand to what extent it informs and improves performance on the downstream task. In typical machine learning tasks, performance is evaluated on held-out test data $\mathcal{D}_{\text{test}}$ using a loss function ℓ_{task}, which we call the *task loss*. To distinguish it from the scoring rule ℓ that parametrizes the uncertainty measure U_ℓ, we refer to that rule as the *uncertainty loss*. The task loss may have the same structure as the uncertainty loss, i.e., $\ell_{\text{task}} : \Delta_K \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ quantifies the cost associated with predicting the probability distribution $\hat{\boldsymbol{\theta}} \in \Delta_K$ when the true outcome is $y \in \mathcal{Y}$, and the overall loss is the average over the predictions on $\mathcal{D}_{\text{test}}$. In general, however, ℓ_{task} can be a complex loss function that is neither defined in an instance-wise manner nor decomposable over the data points in $\mathcal{D}_{\text{test}}$. In selective prediction, for example, the performance is determined by the *ordering* of the data points (according to their uncertainty). Thus, a loss cannot be assigned to an individual data point anymore. Instead, the uncertainty score assigned to a data point can only be assessed in comparison to others. We say a scoring rule ℓ is better aligned with the task loss ℓ_{task} than another rule ℓ' if using its induced uncertainty measure U_ℓ yields a strictly lower (expected) task loss than using $U_{\ell'}$.

3.1 Selective Prediction

Selective prediction is a task where the model can abstain from making a prediction on some inputs if it is uncertain about the correct outcome. Usually, performance on this task is measured using a hold-out dataset, for example, a test set. Formally, let the test set be denoted as $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where for each instance $\boldsymbol{x}_i$, a predictive model outputs a second-order distribution $Q_i \in \Delta_K^{(2)}$. Further, for $\alpha \in [0, 1]$ let $k = \lfloor \alpha n \rfloor$ be a (fixed) rejection level, which dictates the number of instances for which the model is allowed to abstain from making a prediction. The permutation π of $\{1, \ldots, n\}$ is defined so that

$$U_\ell(Q_{\pi(1)}) \geq U_\ell(Q_{\pi(2)}) \geq \cdots \geq U_\ell(Q_{\pi(n)}).$$

In other words, the permutation π sorts instances by their uncertainty, as quantified by the measure U_ℓ. Again, let $\ell \in \mathcal{L}(\Delta_K, \mathcal{Y})$ be a loss function. Then, the area under the loss-rejection curve (AULC) is defined as

$$\text{AULC} = \int_0^1 \left(\frac{1}{\lfloor \alpha n \rfloor} \sum_{\text{j}=1}^{\lfloor \alpha n \rfloor} \ell^*(\hat{\boldsymbol{\theta}}_{\pi(\text{j})}, \text{y}_{\pi(\text{j})}) \right) \text{d}\alpha. \tag{6}$$

Taking the expectation (over the randomness in the labels) in (6) yields the *expected* AULC. In the context of selective prediction, AULC can be interpreted as the task loss ℓ_{task}, as it quantifies the expected prediction error over varying levels of instance rejection. We call $\ell^\star$ in (6) *auxiliary* task loss.

Proposition 1. *Let $\hat{\boldsymbol{\theta}} \in \Delta_K$ be a (first-order) prediction and $\ell \in \mathcal{L}(\Delta_K, \mathcal{Y})$. Then, the expected AULC is minimized by ordering test instances in non-decreasing order of their (instance-wise) expected loss $\mathbb{E}_{y \sim \theta}\big[\ell(\hat{\boldsymbol{\theta}}, y)\big]$.*

Proof. For $\alpha \in [0, 1]$, the area under the loss-rejection curve (AULC) is defined as

$$\text{AULC} = \int_0^1 \left(\frac{1}{\lfloor \alpha n \rfloor} \sum_{\text{j}=1}^{\lfloor \alpha n \rfloor} \ell(\hat{\theta}_{\pi(\text{j})}, \text{y}_{\pi(\text{j})}) \right) \text{d}\alpha.$$

Define $c_{\pi(j)} := \mathbb{E}\big[\ell(\hat{\theta}_{\pi(j)}, y_{\pi(j)})\big]$. Then, the *expected* area under the loss-rejection curve is given by

$$\begin{aligned} \mathbb{E}[\text{AULC}] &= \int_0^1 \left(\frac{1}{\lfloor \alpha n \rfloor} \sum_{j=1}^{\lfloor \alpha n \rfloor} \mathbb{E}\big[\ell(\hat{\theta}_{\pi(j)}, y_{\pi(j)})\big] \right) d\alpha \\ &= \int_0^1 \left(\frac{1}{\lfloor \alpha n \rfloor} \sum_{j=1}^{\lfloor \alpha n \rfloor} c_{\pi(j)} \right) d\alpha. \end{aligned} \tag{7}$$

Consequently, we can approximate the integral by a Riemann sum with step $\Delta\alpha = \frac{1}{n}$:

$$\mathbb{E}[\text{AULC}] \approx \underbrace{\frac{1}{n} \sum_{k=1}^{n} \left(\frac{1}{k} \sum_{j=1}^{k} c_{\pi(j)} \right)}_{=: S(\pi)}.$$

Dataset	Method	log loss	Brier loss	zero-one loss
CIFAR-100	Dropout	**0.829**$\pm$0.000	0.822$\pm$0.000	0.702$\pm$0.001
	Ensemble	**0.860**$\pm$0.001	0.852$\pm$0.002	0.762$\pm$0.002
	Laplace	**0.845**$\pm$0.002	0.836$\pm$0.001	0.808$\pm$0.001
PLACES365	Dropout	**0.837**$\pm$0.001	0.828$\pm$0.001	0.714$\pm$0.002
	Ensemble	**0.856**$\pm$0.002	0.846$\pm$0.002	0.758$\pm$0.005
	Laplace	**0.863**$\pm$0.003	0.850$\pm$0.003	0.825$\pm$0.004
SVHN	Dropout	**0.835**$\pm$0.000	0.830$\pm$0.000	0.701$\pm$0.002
	Ensemble	**0.872**$\pm$0.005	0.868$\pm$0.005	0.776$\pm$0.007
	Laplace	**0.865**$\pm$0.005	0.856$\pm$0.004	0.826$\pm$0.006

Table 1: OoD detection with CIFAR-10 as in-Distribution data based on epistemic uncertainty. *The mean and standard deviation of the AUROC over three runs are reported.* Best results are highlighted in **bold**.

Then, interchanging the order of summation yields

$$S(\pi) = \sum_{k=1}^{n} \sum_{j=1}^{k} \frac{1}{k}\, c_{\pi(j)} = \sum_{j=1}^{n} c_{\pi(j)} \sum_{k=j}^{n} \frac{1}{k}.$$

With weights $w_j = \sum_{k=j}^{n} \frac{1}{k}$ we finally get $S(\pi) = \sum_{j=1}^{n} w_j\, c_{\pi(j)}$. Since $w_1 \geq w_2 \geq \cdots \geq w_n > 0$, the rearrangement inequality implies that the sum $\sum_{j=1}^{n} w_j\, c_{\pi(j)}$ is minimized when $c_{\pi(1)} \leq c_{\pi(2)} \leq \cdots \leq c_{\pi(n)}$. □

Now, if $\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$, then considering the expectation (with respect to the learner's belief Q) over $\mathbb{E}_{y \sim \boldsymbol{\theta}}\big[\ell(\bar{\boldsymbol{\theta}}, y)\big]$ yields the measure of total uncertainty in (5). This leads to an important observation: In selective prediction, when determining the ordering of test instances (e.g., based on uncertainty measures), the most sensible strategy to minimize the expected AULC, as established in Theorem 1, is to order them according to the (predicted) *total* uncertainty with uncertainty loss ℓ given by ℓ^* in (6).

As an aside, let us note that loss-rejection curves (or, analogously, accuracy-rejection curves) are commonly used as a means to evaluate aleatoric and epistemic uncertainty measures, too, which means the curves are constructed for these measures as selection criteria (Hüllermeier, Destercke, and Shaker 2022; Sale et al. 2024b). In light of our finding that total uncertainty is actually the right criterion, this practice may appear somewhat questionable, as it means that aleatoric and epistemic uncertainty measures are evaluated on a task they are actually not tailored to.

Empirical Results. We generate loss-rejection-curves by rejecting the predictions for instances on which the predictor is most uncertain and computing the loss on the remaining subset (Hühn and Hüllermeier 2008). Given a good uncertainty quantification method, the loss should monotonically decrease with the percentage of rejected instances, because the model misclassifies instances with low uncertainty less often than instances with high uncertainty.

In particular, we train a `RandomForest` classifier, an ensemble of decision trees, on the COVERTYPE dataset (Blackard and Dean 1999). In Figure 1 we show the results for three different task losses using total uncertainty as the rejection criterion. This validates the theory: the rejection ordering is optimal when the uncertainty loss is aligned with the task loss. The effect is most pronounced with zero-one loss, the canonical loss for selective prediction (Nadeem, Zucker, and Hanczar 2010; Geifman and El-Yaniv 2017). Experimental details and additional experiments with varying uncertainty components, models, and datasets are deferred to the supplementary material. The code is at: https://github.com/pwhofman/task-specific-uncertainty.

Finally, two practical caveats are worth noting. The optimal ordering in Theorem 1 presumes that the uncertainty values faithfully reflect the instance-wise expected loss; poor posterior approximations will blur that ranking and weaken the gains, so investing in better second-order beliefs (e.g., via ensembling etc.) improves selective prediction in practice. Conversely, using only aleatoric or only epistemic components as selection criterion, rather than total uncertainty aligned with the task loss, can be misleading.

3.2 Out-of-Distribution Detection

Another complementary downstream task to assess and compare the quality of uncertainty quantification is out-of-distribution (OoD) detection. We first train the model on in-distribution (iD) data and evaluate its uncertainty on held-out iD test instances. Then, we present the model with OoD samples and compute their uncertainties as well. Because the model has not seen the OoD domain during training, it should register *higher* epistemic uncertainty on those inputs. Being able to separate iD from OoD points is essential for reliability, since predictions outside the training distribution are inherently less trustworthy.

Empirical Results. We train a `ResNet18` (He et al. 2016) on CIFAR-10 and approximate the second-order predictive distribution using three methods: deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017), Monte Carlo Dropout (Gal and Ghahramani 2016), and a Laplace approximation around the trained parameters (Daxberger et al. 2021). Table 1 reports epistemic uncertainty performance on OoD datasets with CIFAR-10 as the in-distribution data, comparing three loss-based instantiations of the epistemic uncertainty measure (3). The results show that epistemic uncertainty measures instantiated with the log loss (i.e., mutual information) achieves the best OoD performance, which

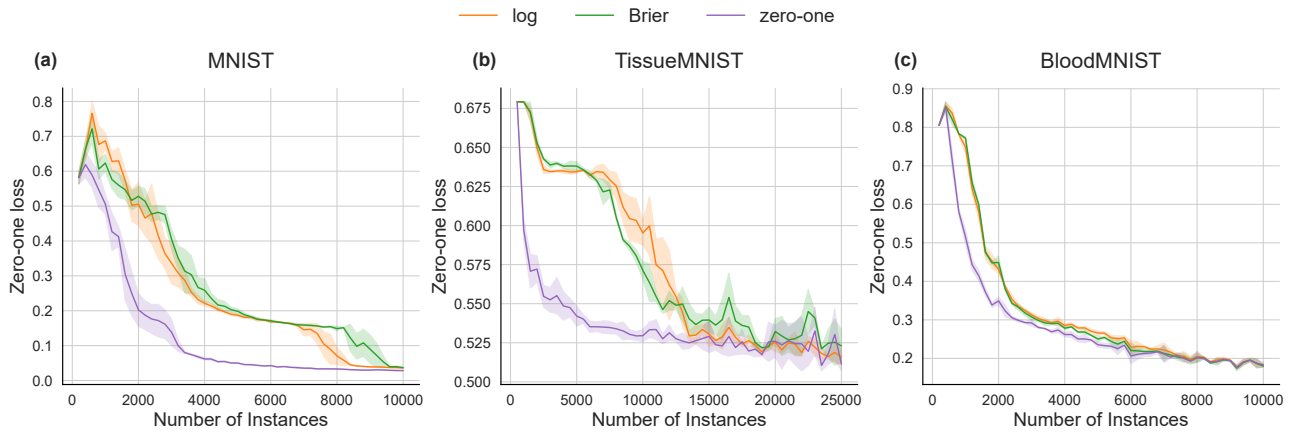

Figure 2: Active learning with different datasets using the epistemic uncertainty component to query new instances, where **(a)** is based on the MNIST dataset, **(b)** TISSUEMNIST, and **(c)** BLOODMNIST, respectively. *Results are averaged over three runs.*

may be explained by the log loss penalizing over-confident predictions, thus improving the separation of iD and OoD samples. This aligns with, and helps justify, its widespread use in second-order uncertainty representations for OoD detection (Mucsányi, Kirchhof, and Oh 2024). On the contrary, the zero-one loss ignores the confidence magnitude. Epistemic uncertainty is zero when the argmaxes "agree", yielding a weaker separation signal. Additional results using IMAGENET (Deng et al. 2009) and FOOD101 (Bossard, Guillaumin, and Gool 2014) as iD data are reported in the supplementary material; they further underscore the superior performance of the log-loss–based instantiation of (3).

We note that OoD detection is qualitatively different from selective prediction or active learning because the "task" itself is shaped by how the OoD examples are constructed. Covariate shifts, semantic shifts, near- versus far-OoD, and synthetic versus natural perturbations all induce different separability structures between in- and out-of-distribution data, so any ranking-based metric, such as the AUROC used here, conflates properties of the uncertainty measure with the particular flavor of shift being evaluated. That mutual information (the log loss instantiated epistemic uncertainty measure) performs best in our experiments indicates it is especially sensitive to the kinds of unfamiliarity present in these benchmarks; however, this performance should not be interpreted as universally dominant without qualification. For example, Li et al. (2025) critically re-examine common OoD detection pipelines and argue that many of them are effectively asking the wrong questions, conflating surrogate signals with the underlying notion of "out-of-distributionness".

This finding, too, echoes the paper's central message: the "best" uncertainty measure depends on the downstream task (there is *no* one size fits all). In selective prediction, aligning the uncertainty loss with the task loss yields optimal behavior, whereas in the OoD detection benchmarks, the log loss instantiation of epistemic uncertainty (mutual information) empirically separates familiar from unfamiliar inputs most reliably, albeit with the caveat that its dominance can depend on the specific distribution shift and evaluation setup.

3.3 Active Learning

Active learning is another popular downstream task that is frequently used to evaluate uncertainties, since its success hinges on identifying examples about which the model is most uncertain in an epistemic sense and thus will benefit most from labeling (Nguyen, Shaker, and Hüllermeier 2022). Its objective is to reach strong performance with as few labels as possible. Beginning from a (small) labeled seed set, the learner iteratively selects unlabeled examples to be annotated by an oracle. Many query strategies leverage epistemic uncertainty (Nguyen, Destercke, and Hüllermeier 2019; Kirsch, van Amersfoort, and Gal 2019; Margraf et al. 2024). The source of (epistemic) uncertainty we care about in this setting is label disagreement: whether plausible predictive distributions disagree on the most likely class. Accordingly, the uncertainty measure should directly reflect that disagreement. Many common measures instead conflate label-level ambiguity with other sources of epistemic uncertainty, for example, uncertainty about the full first-order distribution, even when all plausible predictors agree on the top label. That extra sensitivity can cause less effective queries, whereas the zero-one loss instantiation isolates true disagreement on the predicted label. Here, we run pool-based active learning: in each round, we score candidates using different instantiations of the epistemic uncertainty measure (3) and query the highest-uncertainty examples.

Empirical Results. We use MNIST (LeCun et al. 1998), FASHIONMNIST (Xiao, Rasul, and Vollgraf 2017), and multiclass subsets of the MEDMNIST collection (Yang et al. 2023). The benchmark includes both color and grayscale tasks; for color inputs we employ a small convolutional architecture based on the `LeNet` (LeCun et al. 1998) architecture, and for grayscale data a small fully connected network. The second-order predictive distribution is approximated via Monte Carlo Dropout, as is standard in image-based active learning (Gal, Islam, and Ghahramani 2017; Kirsch, van Amersfoort, and Gal 2019). Figure 2 shows the task loss (zero-one) versus the number of labeled examples. On all shown datasets, epistemic uncertainty sampling using

the zero-one–loss instantiation delivers the best label efficiency. Also for active learning experiments, further datasets and ablations can be found in the supplementary material.

The zero-one loss instantiation performs well because it targets disagreement over the most likely label among plausible predictive distributions, i.e., the kind of label disagreement that active learning specifically seeks to resolve. The most informative unlabeled examples are those where the model is unsure about the correct label because different plausible (first-order) distributions disagree on which class is most likely; examples where all of them agree on a single top class add little new information. The zero-one instantiation of the epistemic uncertainty measure captures this disagreement: its value is zero when there is consensus on the predicted class and grows only when there is genuine uncertainty about which label should be chosen. In contrast, other instantiations (e.g., log or Brier) may signal uncertainty even though the predicted label would remain unchanged, since they respond to finer variations in the second-order distribution that do not affect the top choice, potentially resulting in less focused queries. Once more, we see that one size does *not* fit all, active learning benefits most from the specific form of uncertainty captured by zero-one loss (disagreement on the predicted label) rather than a broad, undifferentiated uncertainty measure.

3.4 Findings and Insights

Taken together, our theoretical and empirical results form a coherent prescription: uncertainty quantification must be *customized* to the downstream task, not treated as a one-size-fits-all solution. On selective prediction, aligning the uncertainty loss with the task loss yields the optimal rejection ordering; for OoD detection, the log loss instantiation (mutual information) best isolates unfamiliar inputs; and for active learning, resolving label-level disagreement via the zero-one epistemic measure drives the most label-efficient gains. Because these three tasks are among the most widely used evaluation paradigms in both research and practice, this paper serves as a reality check: without careful alignment, empirical comparisons can be misleading.

4 Related Work

Uncertainty Quantification. For second-order distributions, the most commonly used measures are based on information-theoretic decompositions of Shannon entropy (Depeweg et al. 2018). However, these measures have been criticized for violating several properties that uncertainty measures should fulfill (Wimmer et al. 2023). A generalization has been proposed, which considers different instantiations of the predicting model and approximations of the predictive first-order distribution (Schweighofer et al. 2025). Beyond information-theoretic approaches, there have also been proposals of uncertainty measures that are based on a decomposition of risk or loss. Lahlou et al. (2023) propose a method to directly quantify epistemic uncertainty based on the difference between total risk and Bayes risk. Gruber and Buettner (2023) use a general bias-variance decomposition to quantify uncertainty based on the Bregman information (Banerjee, Guo, and Wang 2004) with mutual information as a concrete instantiation. Recently, Hofman, Sale, and Hüllermeier (2024b,a); Schweighofer et al. (2025); Kotelevskii et al. (2025) have introduced an uncertainty quantification framework based on proper scoring rules.

Connection to Downstream Tasks. A complimentary line of work starts from the downstream predictive task. Smith et al. (2024) argue that one should reason about uncertainty by first specifying the predictive task at hand. Other related work compares alternative representations of uncertainty and shows that the chosen representation can substantially affect task performance (Mucsányi, Kirchhof, and Oh 2024; de Jong, Sburlea, and Valdenegro-Toro 2024). Likewise, several works observe that different uncertainty measures exhibit markedly different performance profiles across downstream tasks (Schweighofer et al. 2025; Kotelevskii et al. 2025). To the best of our knowledge, our contribution is the first to explicitly connect, in a single framework, downstream *tasks* to specific uncertainty *measures*, theoretically (via loss-alignment results) and empirically (across selective prediction, OoD detection, and active learning).

5 Concluding Remarks

We have argued and shown that uncertainty quantification is not a one-size-fits-all endeavor: the usefulness of a given uncertainty measure depends critically on the downstream task and how that task evaluates performance. On the theoretical side, we tied the construction of uncertainty measures to proper scoring rules and proved that optimal instance ordering in selective prediction arises when the uncertainty loss is aligned with the task loss. Empirically, we confirmed this principle across three canonical evaluation paradigms. For selective prediction, total uncertainty instantiated with the task-aligned loss yields the best rejection behavior; for out-of-distribution detection, the log loss based epistemic measure (mutual information) most reliably identifies unfamiliar inputs; and for active learning, querying based on the zero-one loss epistemic disagreement delivers the strongest label efficiency. Beyond these specific findings, the broader implication is practical: researchers and practitioners should choose and report uncertainty measures with their target objectives in mind. Blindly applying generic uncertainty scores or mixing selection and evaluation criteria can obscure real performance differences and lead to misleading conclusions.

Limitations and Future Work. All of our downstream benefits hinge on reasonably faithful second-order beliefs; poor posterior approximations can degrade the expected gains, underscoring the value of improved uncertainty representations (e.g., better ensembles, or more expressive posterior approximations). It remains of great interest for both the machine learning community and practitioners to understand how to (empirically) *evaluate* uncertainty itself, an inherently difficult problem, since true uncertainty is unobserved and must be judged indirectly via downstream objectives. Our results *caution* against one-size-fits-all practices and instead point toward task-aligned evaluation protocols and benchmarks, alongside extending the framework beyond multiclass classification to regression, structured outputs, and cost-sensitive or imbalanced settings.

6 Acknowledgments

Yusuf Sale is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

Banerjee, A.; Guo, X.; and Wang, H. 2004. Optimal Bregman prediction and Jensen's equality. In *Proceedings of the 2004 IEEE International Symposium on Information Theory, ISIT 2004, Chicago Downtown Marriott, Chicago, Illinois, USA, June 27 - July 2, 2004*, 169. IEEE.

Blackard, J. A.; and Dean, D. J. 1999. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3): 131–151.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 - Mining Discriminative Components with Random Forests. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, 446–461. Springer.

Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1): 1–3.

Clarté, L.; Loureiro, B.; Krzakala, F.; and Zdeborová, L. 2023. Expectation consistency for calibration of neural networks. In Evans, R. J.; and Shpitser, I., eds., *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216 of *Proceedings of Machine Learning Research*, 443–453. PMLR.

Daxberger, E.; Kristiadi, A.; Immer, A.; Eschenhagen, R.; Bauer, M.; and Hennig, P. 2021. Laplace Redux - Effortless Bayesian Deep Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 20089–20103.

de Jong, I. P.; Sburlea, A. I.; and Valdenegro-Toro, M. 2024. How disentangled are your classification uncertainties? *CoRR*, abs/2408.12175.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 248–255. IEEE Computer Society.

Depeweg, S.; Hernandez-Lobato, J.-M.; Doshi-Velez, F.; and Udluft, S. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, 1184–1193. PMLR.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M.; and Weinberger, K. Q., eds., *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 1050–1059. JMLR.org.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep Bayesian Active Learning with Image Data. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1183–1192. PMLR.

Geifman, Y.; and El-Yaniv, R. 2017. Selective Classification for Deep Neural Networks. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4878–4887.

Gneiting, T.; and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477): 359–378.

Gruber, S.; and Buettner, F. 2023. Uncertainty Estimates of Predictions via a General Bias-Variance Decomposition. In Ruiz, F. J. R.; Dy, J. G.; and van de Meent, J., eds., *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, 11331–11354. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Hofman, P.; Sale, Y.; and Hüllermeier, E. 2024a. Quantifying aleatoric and epistemic uncertainty: A credal approach. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*.

Hofman, P.; Sale, Y.; and Hüllermeier, E. 2024b. Quantifying aleatoric and epistemic uncertainty with proper scoring rules. *arXiv preprint arXiv:2404.12215*.

Hühn, J. C.; and Hüllermeier, E. 2008. FR3: A fuzzy rule learner for inducing reliable classifiers. *IEEE Transactions on Fuzzy Systems*, 17(1): 138–149.

Hüllermeier, E.; Destercke, S.; and Shaker, M. H. 2022. Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison. In Cussens, J.; and Zhang, K., eds., *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, 548–557. PMLR.

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

Kirsch, A.; van Amersfoort, J.; and Gal, Y. 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kotelevskii, N.; Kondratyev, V.; Takáč, M.; Moulines, E.; and Panov, M. 2025. From Risk to Uncertainty: Generating Predictive Uncertainty Measures via Bayesian Estimation. In *The Thirteenth International Conference on Learning Representations*.

Kull, M.; and Flach, P. A. 2015. Novel Decompositions of Proper Scoring Rules for Classification: Score Adjustment as Precursor to Calibration. In Appice, A.; Rodrigues, P. P.; Costa, V. S.; Soares, C.; Gama, J.; and Jorge, A., eds., *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I*, volume 9284 of *Lecture Notes in Computer Science*, 68–85. Springer.

Lahlou, S.; Jain, M.; Nekoei, H.; Butoi, V.; Bertin, P.; Rector-Brooks, J.; Korablyov, M.; and Bengio, Y. 2023. DEUP: Direct Epistemic Uncertainty Prediction. *Trans. Mach. Learn. Res.*, 2023.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6402–6413.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11): 2278–2324.

Li, Y. L.; Lu, D.; Kirichenko, P.; Qiu, S.; Rudner, T. G.; Bruss, C. B.; and Wilson, A. G. 2025. Out-of-Distribution Detection Methods Answer the Wrong Questions. *arXiv preprint arXiv:2507.01831*.

Margraf, V.; Wever, M.; Gilhuber, S.; Tavares, G. M.; Seidl, T.; and Hüllermeier, E. 2024. ALPBench: A Benchmark for Active Learning Pipelines on Tabular Data. ArXiv:2406.17322 [cs].

Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; and Lucic, M. 2021. Revisiting the Calibration of Modern Neural Networks. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 15682–15694.

Mucsányi, B.; Kirchhof, M.; and Oh, S. J. 2024. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nadeem, M. S. A.; Zucker, J.; and Hanczar, B. 2010. Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option. In Dzeroski, S.; Geurts, P.; and Rousu, J., eds., *Proceedings of the third International Workshop on Machine Learning in Systems Biology, MLSB 2009, Ljubljana, Slovenia, September 5-6, 2009*, volume 8 of *JMLR Proceedings*, 65–81. JMLR.org.

Nguyen, V.; Shaker, M.; and Hüllermeier, E. 2022. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1): 89–122.

Nguyen, V.-L.; Destercke, S.; and Hüllermeier, E. 2019. Epistemic Uncertainty Sampling. In Kralj Novak, P.; Šmuc, T.; and Džeroski, S., eds., *Discovery Science*, 72–86. Cham: Springer International Publishing. ISBN 978-3-030-33778-0.

Sale, Y.; Bengs, V.; Caprio, M.; and Hüllermeier, E. 2024a. Second-Order Uncertainty Quantification: A Distance-Based Approach. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Sale, Y.; Hofman, P.; Löhr, T.; Wimmer, L.; Nagler, T.; and Hüllermeier, E. 2024b. Label-wise Aleatoric and Epistemic Uncertainty Quantification. In Kiyavash, N.; and Mooij, J. M., eds., *Uncertainty in Artificial Intelligence, 15-19 July 2024, Universitat Pompeu Fabra, Barcelona, Spain*, volume 244 of *Proceedings of Machine Learning Research*, 3159–3179. PMLR.

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336): 783–801.

Schweighofer, K.; Aichberger, L.; Ielanskyi, M.; and Hochreiter, S. 2025. On Information-Theoretic Measures of Predictive Uncertainty. In Chiappa, S.; and Magliacane, S., eds., *Conference on Uncertainty in Artificial Intelligence, Rio Othon Palace, Rio de Janeiro, Brazil, 21-25 July 2025*, volume 286 of *Proceedings of Machine Learning Research*, 3605–3640. PMLR.

Smith, F. B.; Kossen, J.; Trollope, E.; van der Wilk, M.; Foster, A.; and Rainforth, T. 2024. Rethinking Aleatoric and Epistemic Uncertainty. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*.

Smith, L.; and Gal, Y. 2018. Understanding Measures of Uncertainty for Adversarial Example Detection. In Globerson, A.; and Silva, R., eds., *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 560–569. AUAI Press.

Wimmer, L.; Sale, Y.; Hofman, P.; Bischl, B.; and Hüllermeier, E. 2023. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In Evans, R. J.; and Shpitser, I., eds., *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216 of *Proceedings of Machine Learning Research*, 2282–2292. PMLR.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; and Ni, B. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1): 41.