



HAL
open science

On the Calibration of Epistemic Uncertainty: Principles, Paradoxes and Conflictual Loss

Mohammed Fellaji, F. Pennerath, Briec Conan-Guez, Miguel Couceiro

► To cite this version:

Mohammed Fellaji, F. Pennerath, Briec Conan-Guez, Miguel Couceiro. On the Calibration of Epistemic Uncertainty: Principles, Paradoxes and Conflictual Loss. Machine Learning and Knowledge Discovery in Databases. Research Track European Conference, ECML-PKDD 2024, Sep 2024, Vilnius, Lithuania. pp.160-176. hal-04695978

HAL Id: hal-04695978

<https://inria.hal.science/hal-04695978v1>

Submitted on 12 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

On the Calibration of Epistemic Uncertainty: Principles, Paradoxes and Conflictual Loss

Mohammed Fellaji¹ (✉), Frédéric Pennerath¹,
Brieuc Conan-Guez², and Miguel Couceiro^{2,3}

¹ CentraleSupélec, Université Paris Saclay, CNRS, LORIA, France
{mohammed.fellaji, frederic.pennerath}@centralesupelec.fr

² Université de Lorraine, CNRS, LORIA, France
brieuc.conan-guez@univ-lorraine.fr
miguel.couceiro@loria.fr

³ INESC-ID, IST, Universidade de Lisboa, Portugal

Abstract. The calibration of predictive distributions has been widely studied in deep learning, but the same cannot be said about the more specific epistemic uncertainty as produced by Deep Ensembles, Bayesian Deep Networks, or Evidential Deep Networks. Although measurable, this form of uncertainty is difficult to calibrate on an objective basis as it depends on the prior for which a variety of choices exist. Nevertheless, epistemic uncertainty must in all cases satisfy two formal requirements: firstly, it must decrease when the training dataset gets larger and, secondly, it must increase when the model expressiveness grows. Despite these expectations, our experimental study shows that on several reference datasets and models, measures of epistemic uncertainty violate these requirements, sometimes presenting trends completely opposite to those expected. These paradoxes between expectation and reality raise the question of the true utility of epistemic uncertainty as estimated by these models. A formal argument suggests that this disagreement is due to a poor approximation of the posterior distribution rather than to a flaw in the measure itself. Based on this observation, we propose a regularization function for deep ensembles, called *conflictual loss* in line with the above requirements. We emphasize its strengths by showing experimentally that it fulfills both requirements of epistemic uncertainty, without sacrificing either the performance nor the calibration of the deep ensembles.

Keywords: Epistemic Uncertainty · Calibration · Bayesian Deep Learning · Deep Ensembles · Evidential Deep Networks.

1 Introduction

All neural networks, from small discriminative classifiers to large generative models, can be seen as probabilistic models that estimate some distribution. This distribution captures the uncertainty of the predicted variable, induced both by *latent factors* that are inherent in the process that has generated the data,

and by the *model bias*, which reflects the lack of expressiveness of the model to represent the true distribution. The uncertainty related to latent factors is sometimes referred to as *aleatoric or data uncertainty* in contrast to the completely different *epistemic or model uncertainty* that is meant to measure *estimator variance / overfitting, i.e.*, the uncertainty about the output distribution itself, due to the limited size of the training dataset. While every probabilistic model de facto takes into account aleatoric uncertainty, epistemic uncertainty becomes measurable only by models whose output distribution is a random variable. This includes *Bayesian Neural Networks* (BNN) that apply (approximate) Bayesian inference on network weights [17,9,22,6], *Deep Ensembles* (DE) that sample the prior distribution [14], *Prior Networks* [18], *Evidential Deep Learning* (EDL) [28] and derived methods that directly learn parameters of a second-order distribution. All of these models produce not only the *posterior predictive distribution* for a given input, but also a measure of epistemic uncertainty that quantifies the part of uncertainty that can be further reduced by observing more data in the vicinity of the input. Since this measure is usually computed as the *mutual information* between the model output and the parameters (conditioned on the input and the training dataset), we will stick to this choice in the sequel, even if the choice of a better epistemic uncertainty metric remains an arguable subject, as discussed in [32]. Regardless of the choice of metric, epistemic uncertainty appears as a relevant criterion for deciding to label a new example in the context of active learning [4,7], to tackle the exploration-exploitation dilemma in reinforcement learning [29], or to detect OOD examples [15], although some approaches advocate an even finer decomposition to either detect OODs, distinguishing between epistemic and *distributional uncertainty* [18], or to account for *procedural variability* [10], *i.e.*, uncertainty coming from the randomness of the optimization procedure. Accurately quantifying epistemic uncertainty is thus crucial from both theoretical and application perspectives.

Whereas there is a large body of work on the calibration of predictive uncertainty as produced by deep neural networks, including Bayesian ones [13,24], to our knowledge, there is no work dealing specifically with the calibration of epistemic uncertainty. In this paper, we address the question of how to evaluate the quality of epistemic uncertainty produced by deep networks. One difficulty that may explain the lack of research in this field is that the amount of epistemic uncertainty depends on the prior distribution over parameters for which certain freedom of choice exists [5], whether it is an *informative* or an *objective prior* like *Jeffreys prior* [11]. Consequently, the definition of a quantitative score to measure the quality of epistemic uncertainty appears as a questionable objective, and thus we do not consider it. Instead, we adopt a qualitative standpoint by stating two properties that every measure of epistemic uncertainty should ideally fulfill: the first property that we call hereafter *data-related principle*, states that the amount of epistemic uncertainty decreases as the model observes more data. The second property, referred to as *model-related principle*, states that epistemic uncertainty must increase with model complexity, *i.e.*, the number of weights, as a consequence of the *curse of dimensionality*.

While we can check these requirements are de facto true for a simple probabilistic model like Bayesian linear regression, it is not obvious that this still holds for Bayesian deep networks or their alternatives, since the parameters of these more complex non-convex models converge somewhat randomly to one of the many local optima. With this in mind, we conducted an experimental study of epistemic uncertainty as produced by *Deep Ensembles* [14], *MC-Dropout* [6] and *Evidential Deep Learning* [28]. The results are surprising: we observe that in all data regimes and for all tested methods, the average measures of mutual information computed on the test set, completely contradict the *model-related principle*: the larger the model, the smaller the epistemic uncertainty when precisely the opposite is expected. The *data-related principle* seems globally but not perfectly respected, with some blatant counter-examples. The same paradoxes are consistently observed, even when calibration techniques are used like *label smoothing* [30] and *confidence penalty* [26]. This disagreement between expectation and reality thus raises the question of the true utility of epistemic uncertainty as estimated by these models.

A necessary condition to solve these inconsistencies is to ensure that epistemic uncertainty is maximal in the absence of data, a property that can only result from an appropriate choice of prior, or equivalently, of the regularizer in the loss function. Based on this observation, we designed an elementary regularization function for ensembles of deep classifiers, called *conflictual loss* for reasons that will become obvious later on. We emphasize the strengths of the resulting *Conflictual Deep Ensembles* by showing experimentally that it restores both properties of epistemic uncertainty, without sacrificing either the performance or the calibration of the deep ensembles. To summarize, our contributions are the following:

- A method for assessing the quality of epistemic uncertainty of a model based on two principles.
- The empirical demonstration, using this method, that common models and calibration techniques do not satisfy (and even sometimes contradict) these quality criteria.
- A theoretical argument that suggests that these inconsistencies are due to the poor posterior approximation and not to the metric itself.
- A new regularizer for deep ensembles, called the *conflictual loss function*, designed to ensure the data-related principle of epistemic uncertainty.
- Experimental results showing that this technique restores both quality criteria of epistemic uncertainty without degrading the other performance scores (accuracy, calibration, OOD detection).

The rest of the paper is structured as follows: Section 2 presents some previous works in the field of uncertainty, calibration, and prior. Section 3 formalizes the two fundamental properties of epistemic uncertainty and gives some theoretical insights about them. Section 4 describes the *conflictual loss* for deep ensembles. Section 5 presents experimental results and section 6 concludes.

2 Related Work

The calibration of a model reflects how its predictive distributions are consistent with its errors on a test dataset. In [23], the authors discussed existing calibration metrics such as the *Expected Calibration Error* (ECE) [25] and introduced new measures like *Static Calibration Error* (SCE) to better take multiclass problems into account. Post-hoc calibration techniques are also popular: histogram binning [34], isotonic regression [35] and temperature scaling [8] to name the most common. The latter performs generally well on in-domain data but falls short when the data undergoes a distributional shift or are out-of-distribution (OOD) [24]. Few works have studied calibration of posterior predictive [24,33].

When it comes to priors and regularization in deep learning, there is a considerable literature that can be classified into two main categories: parameter-based (regularizers L1, L2, etc.) and output-based (such as label smoothing [30], confidence penalty [26]). The latter techniques are introduced primarily to avoid peaky outputs, which are a sign of overfitting [20]. Priors have given rise to numerous works in general, with some specific to Bayesian deep learning (see survey [5] on this subject). To the best of our knowledge, there is no work focusing on calibrating epistemic uncertainty based on priors and without using an additional model to compute uncertainties.

To avoid the multiple evaluations of BNNs at inference time, some authors also propose to estimate the epistemic uncertainty more directly: the model predicts its uncertainty about the prediction. More precisely, the model produces a second-order distribution, that is a distribution over class distributions. Such “all in one” approaches (Evidential Deep Learning [28], Prior Networks [18], Information Aware Dirichlet [31], to name a few) require some specific training schemes. Indeed, [1] shows that training in a classical way such models by minimizing a second-order loss, does not entail well-calibrated uncertainty estimates.

The task of OOD detection is important in many applications. OOD samples are fundamentally different from the samples used during training [27,28]. In theory, these examples should yield a high epistemic uncertainty. The task of OOD detection is challenging as shown in [24]: most of the methods tested on different benchmarks resulted in high-confidence predictions on OOD samples. As shown in [24,21], DE and MC-Dropout are competitive benchmarks across different tasks including OOD detection. Additionally, some approaches to detect OOD data involve training the model with both in-domain and OOD samples [19,16,18]. However, they have some limitations such as relying on the choice of the OOD dataset and that epistemic uncertainty is shifted into aleatoric uncertainty as discussed in [12].

3 Principles of Epistemic Uncertainty

In this section, we introduce two principles that characterize an *idealized* measure of epistemic uncertainty. For each of them, we give a formal definition, we justify why it is a desirable feature and we give a first analysis on its practical validity.

In what follows, we consider a family of probabilistic models that estimate the distribution of some measurable output $Y \in \mathcal{Y}$ given some input vector $\mathbf{x} \in \mathcal{X}$, thanks to a parametric function $f_{\boldsymbol{\theta}}$ parameterized by a vector $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$ of parameters, *i.e.*, $p(y | \mathbf{x}, \boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\mathbf{x})$. As Bayesian inference requires the definition of a prior $p(\boldsymbol{\theta})$, the notation \mathcal{M} for a *formal model* refers hereafter to the pair $\mathcal{M} = (f_{\boldsymbol{\theta}}, p(\boldsymbol{\theta}))$. Given such a model \mathcal{M} conditioned on some training sample $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^*$, we consider a metric function $\mathcal{U}_{\mathcal{D}, \mathcal{M}}^e : \mathcal{X} \rightarrow \mathbb{R}^+$ that maps to an input \mathbf{x} , the measure of epistemic uncertainty conveyed by joint distribution $p(y, \boldsymbol{\theta} | \mathbf{x}, \mathcal{D})$. We assume that this metric grows with epistemic uncertainty and is non-negative. This is the case of common metrics like *mutual information*

$$\mathcal{U}_{\mathcal{D}, \mathcal{M}}^i(\mathbf{x}) = I(Y ; \Theta | \mathbf{x}, \mathcal{D}) = H(Y | \mathbf{x}, \mathcal{D}) - H(Y | \Theta, \mathbf{x}, \mathcal{D}),$$

where $H(\cdot | \cdot)$ denotes conditional entropy. In case of regression (*i.e.*, $Y \in \mathbb{R}$), another option is *difference of variances*

$$\mathcal{U}_{\mathcal{D}, \mathcal{M}}^v(\mathbf{x}) = \mathbb{V}(Y | \mathbf{x}, \mathcal{D}) - \mathbb{V}(Y | \Theta, \mathbf{x}, \mathcal{D}),$$

where $\mathbb{V}(Y | \mathbf{x}, \mathcal{D})$ refers to the variance of the output when it follows the posterior predictive $p(Y | \mathbf{x}, \mathcal{D})$ and $\mathbb{V}(Y | \Theta, \mathbf{x}, \mathcal{D})$ refers to the average of variances of individual models $p(Y | \boldsymbol{\theta}, \mathbf{x})$, *i.e.*,

$$\mathbb{V}(Y | \Theta, \mathbf{x}, \mathcal{D}) = \int \mathbb{V}(Y | \boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta}.$$

3.1 Data-related Principle of Epistemic Uncertainty

The first principle simply states that epistemic uncertainty reduces as more training samples become available.

Definition 1 (First principle). *An epistemic uncertainty metric $\mathcal{U}_{\mathcal{D}, \mathcal{M}}^e$ is (ideally) a non-increasing function of training sample \mathcal{D} , *i.e.*,*

$$\forall \mathcal{M}, \forall \mathbf{x}, \forall \mathcal{D}_1, \forall \mathcal{D}_2, \quad \mathcal{D}_1 \subseteq \mathcal{D}_2 \Rightarrow \mathcal{U}_{\mathcal{D}_1, \mathcal{M}}^e(\mathbf{x}) \geq \mathcal{U}_{\mathcal{D}_2, \mathcal{M}}^e(\mathbf{x}).$$

The reason why this property is desirable is illustrated by the next *thought experiment* in the context of *active learning*: suppose that a model \mathcal{M} has been trained on samples \mathcal{D}_1 so far. Now comes a new unlabelled sample \mathbf{x} . Since epistemic uncertainty is the ideal criterion for measuring the information that could be gained by labeling a new sample, the measure $\mathcal{U}_{\mathcal{D}_1, \mathcal{M}}^e(\mathbf{x})$ is compared to some decision threshold σ in order to decide whether the sample is worth being labeled by an expert. Assuming that this is not the case, sample \mathbf{x} is discarded. Later on, the train set has been enriched with more samples \mathcal{D}_2 . If it is possible that $\mathcal{U}_{\mathcal{D}_1 \cup \mathcal{D}_2, \mathcal{M}}^e(\mathbf{x}) > \mathcal{U}_{\mathcal{D}_1, \mathcal{M}}^e(\mathbf{x})$, then it is also possible that $\mathcal{U}_{\mathcal{D}_1 \cup \mathcal{D}_2, \mathcal{M}}^e(\mathbf{x}) \geq \sigma$ so that this time the system would have asked for the labeling of \mathbf{x} . This behavior would go against what we expect, *i.e.*, a model trained on more data has necessarily learned more information. The first principle bans such a scenario.

Next, we analyze the extent to which mutual information satisfies this principle. Although examples can be found such that the observation of a specific sample increases mutual information rather than reducing it, we can ask whether the first principle is satisfied when averaging over all possible observations, *i.e.*, in expectation. At first glance, we would be tempted to answer in the negative, as there exist random variables X , Y and Z such that $I(X; Y) < I(X; Y | Z)$. For making the answer positive, we need the assumption that samples are iid.

Theorem 1. *The mutual information metric satisfies the first principle in expectation with respect to new random iid samples \mathcal{D}_2 , *i.e.*,*

$$\forall \mathcal{M}, \forall \mathbf{x}, \forall \mathcal{D}_1, \quad \mathcal{U}_{\mathcal{D}_1, \mathcal{M}}^i(\mathbf{x}) \geq \mathbb{E}[\mathcal{U}_{\mathcal{D}_1 \cup \mathcal{D}_2, \mathcal{M}}^i(\mathbf{x})] .$$

Proof. For the sake of clarity and without loss of generality, we assume \mathcal{D}_2 is made of one single sample (\mathbf{x}', Y') . Also for conciseness, we denote κ the triplet $(\mathcal{D}_1, \mathbf{x}, \mathbf{x}')$ as these terms have no incidence on the proof below. Then considering a “test” input \mathbf{x} for which we want to estimate the epistemic uncertainty conveyed by its output Y , we consider the difference

$$\begin{aligned} \Delta I &= \mathcal{U}_{\mathcal{D}_1, \mathcal{M}}^i(\mathbf{x}) - \mathbb{E}[\mathcal{U}_{\mathcal{D}_1 \cup \mathcal{D}_2, \mathcal{M}}^i(\mathbf{x})] \\ &= I(Y; \Theta | \mathbf{x}, \mathcal{D}_1) - I(Y; \Theta | \mathbf{x}, \mathcal{D}_1, Y', \mathbf{x}') \\ &= I(Y; \Theta | \mathcal{D}_1, \mathbf{x}, \mathbf{x}') - I(Y; \Theta | Y', \mathcal{D}_1, \mathbf{x}, \mathbf{x}') \text{ as } Y, \Theta \perp\!\!\!\perp \mathbf{X}' | \mathbf{X} \\ &= H(Y | \kappa) - H(Y | \Theta, \kappa) - H(Y | Y', \kappa) + H(Y | \Theta, Y', \kappa) \\ &= H(Y | \kappa) - H(Y | Y', \kappa) - (H(Y | \Theta, \kappa) - H(Y | \Theta, Y', \kappa)) \\ &= I(Y; Y' | \kappa) - I(Y; Y' | \Theta, \kappa) . \end{aligned}$$

But Y and Y' are iid samples, *i.e.*, they are independent given Θ and κ , and thus $I(Y; Y' | \Theta, \kappa) = 0$. Hence, $\Delta I = I(Y; Y' | \kappa) \geq 0$, and the proof is now complete. \square

3.2 Model-related Principle of Epistemic Uncertainty

The second principle essentially expresses *overfitting*: given two models trained with the same set of samples, if one has more expressive power than the other, then it should have a larger measure of epistemic uncertainty since the choice of model candidates is wider, *i.e.*, its *posterior distribution* is more spread out.

While this principle seems intuitive, the formalization of the underlying notion of expressive power requires the use of complex theories of statistical learning (*e.g.*, VC dimension), which we avoid since it is unnecessary. Indeed, we can only consider models that are, by construction, ordered in increasing order of complexity, as defined below.

Definition 2 (Submodel). We say model $\mathcal{M}_a = (f_{\theta_a}^a, p_a(\theta_a))$ is a submodel of model $\mathcal{M}_b = (f_{\theta_b}^b, p_b(\theta_b))$, denote by $\mathcal{M}_a \leq_{\mathcal{M}} \mathcal{M}_b$, if θ_a is a subset of parameters θ_b so that $\theta_b = (\theta_a, \theta_{b'})$ and there exists a constant vector $\theta_{b'}^0 \in \Omega_{\theta_{b'}}$, such that

$$\forall \theta_a \in \Omega_{\theta_a}, \quad f_{\theta_a}^a = f_{(\theta_a, \theta_{b'}^0)}^b \quad \text{and} \quad p_a(\theta_a) = p_b(\theta_a \mid \Theta_{b'} = \theta_{b'}^0).$$

Moreover, when priors are chosen in such a way that individual parameters are independent, freezing $\theta_{b'}$ has no impact on the prior of θ_a , so that the condition on priors simplifies to $p_a(\theta_a) = p_b(\theta_a)$. Since the submodel relation is reflexive, transitive, and antisymmetric, it defines a partial ordering on the set of parameterized models, that can be used to state the second principle.

Definition 3 (Second principle). An epistemic uncertainty metric $\mathcal{U}_{\mathcal{D}, \mathcal{M}}^e$ should be a non-decreasing function over the set of parameterized models, i.e.,

$$\forall \mathcal{D}, \forall \mathbf{x}, \forall \mathcal{M}_1, \forall \mathcal{M}_2, \quad \mathcal{M}_1 \leq_{\mathcal{M}} \mathcal{M}_2 \Rightarrow \mathcal{U}_{\mathcal{D}, \mathcal{M}_1}^e(\mathbf{x}) \leq \mathcal{U}_{\mathcal{D}, \mathcal{M}_2}^e(\mathbf{x})$$

To see why this principle is desirable, consider the following example in the field of *explainability*. Given two black-box models \mathcal{M}_1 and \mathcal{M}_2 with comparable performance, let's assume that $\mathcal{U}_{\mathcal{D}, \mathcal{M}_1}^e$ is larger on average than $\mathcal{U}_{\mathcal{D}, \mathcal{M}_2}^e$ when estimated on a test set. A larger value of epistemic uncertainty for a given input \mathbf{x} means more diverse and thus inconsistent stochastic functions $p(y \mid \theta, \mathbf{x})$ as θ follows the posterior distribution. Therefore, model \mathcal{M}_2 provides on average more similar and consistent functions than \mathcal{M}_1 . Explaining the output $p(y \mid \mathbf{x}, \mathcal{D})$ of a model amounts to summarize in an understandable format these stochastic functions $p(y \mid \theta, \mathbf{x})$ taken as a whole. Model \mathcal{M}_2 is thus preferred since the explanation of its output is shorter on average. But if the second principle is broken, it is possible that \mathcal{M}_1 is formally a submodel of \mathcal{M}_2 . If so, \mathcal{M}_1 is by construction a restriction of \mathcal{M}_2 , providing systematically shorter explanations, a contradiction. In summary, without the second principle, measures of epistemic uncertainty could make inconsistent *Occam's razor principle* and all related concepts (sparsity, minimum description length, etc).

Focusing again on the mutual information metric, we see that the second principle is also verified in expectation, as it is always true that $I(Y; \Theta_1 \mid \mathcal{D}, \mathbf{x}) \leq I(Y; \Theta_1, \Theta_2 \mid \mathcal{D}, \mathbf{x})$. Indeed, the left term of this inequality can be understood as a weighted average of mutual information $I(Y; \Theta_1 \mid \theta_2^0, \mathcal{D}, \mathbf{x})$ of every submodel θ_2^0 . However, the weight of submodel θ_2^0 is given by the posterior $p(\theta_2 \mid \mathcal{D})$. The interpretation of this inequality in expectation is therefore difficult and of little interest in the context of *model selection* since, in practice, we want to compare a model with a given submodel, not with the whole distribution of submodels.

To illustrate, consider a *Bayesian linear regression model* with known homoskedastic variance and isotropic normal prior, i.e.,

$$Y = \sum_{i \in \mathcal{I}} \Theta_i \psi_i(\mathbf{x}) + \varepsilon \quad \text{with} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad \Theta_i \sim \mathcal{N}(0, \sigma_0^2).$$

where the regressor functions ψ_i are chosen in a large collection indexed by \mathcal{I} . After making some *variable selection*, we can force to zero some coefficients θ_i ,

only keeping regressors of the index in $\mathcal{I}' \subseteq \mathcal{I}$. As priors on coefficients are independent, the definition of the resulting submodel $\mathcal{M}_{\mathcal{I}'}$ is obtained just by replacing \mathcal{I} by \mathcal{I}' in the above definition. Now given such a submodel $\mathcal{M}_{\mathcal{I}'}$, we would like to ensure the measure of epistemic uncertainty is smaller for $\mathcal{M}_{\mathcal{I}'}$ than for $\mathcal{M}_{\mathcal{I}}$. For the sake of simplicity, suppose that regressor functions are decorrelated, *i.e.*, $\mathbb{E}[\psi(\mathbf{X})\psi(\mathbf{X})^T] = \text{Id}$. It can be shown that the epistemic uncertainty of submodel $\mathcal{M}_{\mathcal{I}'}$ as estimated by the *difference of variances*, is

$$\mathcal{U}_{\mathcal{D}, \mathcal{M}_{\mathcal{I}'}}^v(\mathbf{x}) = \mathbb{V}(Y | \mathbf{x}, \mathcal{D}) - \mathbb{V}(Y | \Theta, \mathbf{x}, \mathcal{D}) = \frac{|\mathcal{I}'|}{\sigma_0^{-2} + |\mathcal{D}| \sigma^{-2}}.$$

As this value increases with the number $|\mathcal{I}'|$ of parameters and decreases with the number $|\mathcal{D}|$ of examples, the first and second principles are always satisfied.

This result naturally raises the question of whether it can be generalized to models like deep neural networks. As there are no simple analytical answers for such complex models, we present an experimental study in section 5 to assess the extent to which the two principles are satisfied by various classification models. This includes the method of *Confictual Deep Ensembles*, which we introduce in the next section.

4 Confictual Deep Ensembles

As we shall see in Sect. 5, several classifiers present abnormally low levels of epistemic uncertainty in the low-data regime. This “hole” is contrary to the first principle which states that epistemic uncertainty should be maximal in the absence of training data. Why does this happen? By rewriting mutual information $I(Y; \Theta | \mathbf{x}, \mathcal{D})$ as

$$I(Y; \Theta | \mathbf{x}, \mathcal{D}) = \int p(\theta | \mathcal{D}) D_{\text{KL}}(p(y | \theta, \mathbf{x}) \| p(y | \mathbf{x}, \mathcal{D})) d\theta,$$

we can interpret it as a weighted average of divergence between predictions $p(y | \theta, \mathbf{x})$ of individual models and prediction $p(y | \mathbf{x}, \mathcal{D})$ of the averaged model. Therefore, the hole reflects the absence of diversity between output distributions $p(y | \theta, \mathbf{x})$ in the low-data regime. But in this regime, this lack of variability is mostly a consequence of the choice of the prior or, equivalently, of the regularization term in the loss function. This explains why the hole is particularly visible in experiments using *label smoothing*, since this regularization technique drives output distributions closer to the same uniform distribution.

This observation also suggests that designing a prior that favors diversity or, in other words, *discordance* between output distributions, could fill the hole of epistemic uncertainty. Such an objective can be achieved simply by constructing a so-called *confictual deep ensemble*, where each classifier in the ensemble slightly favors a class of its own. In the absence of data, these slight tendencies are enough to create discordance in the output distributions and therefore a high level of mutual information.

In practice, a *conflictual deep ensemble* of order k is implemented as an ensemble of $k \times C$ deep classifiers such that every class $c \in \{1, \dots, C\}$ is mapped to k models $\{\theta_i^c\}_{1 \leq i \leq k}$. Denoting by $P(y | \theta, \mathbf{x})$ the probability of class y as predicted for input \mathbf{x} by model θ , we define the *conflictual loss* for class c as

$$(1) \quad L_c(\theta) = - \sum_{(\mathbf{x}, y) \in \mathcal{D}} (\log P(y | \theta, \mathbf{x}) + \lambda \log P(c | \theta, \mathbf{x})),$$

where the first term is the log-likelihood and the second term is the bias that slightly favors class c . This bias term can be interpreted as if for each observed example (\mathbf{x}, y) in the train set, we add λ faked examples of selected class c . In practice λ has been empirically fixed to 0.05, meaning there is one faked example for 20 real examples. We then train every model of the ensemble independently, using for model θ_i^c the loss function L_c .

The Conflictual Loss both resembles and differs from Label Smoothing: like LS, the conflictual term $\log P(c | \theta, \mathbf{x})$ is not an independent regularization term, but factorized into the sum of the log-likelihoods. However, unlike LS which encourages classifiers to be more concordant by promoting the uniform distribution, the Conflictual Loss encourages classifiers to be contradictory.

Existing works [2,3] have sought to address the diversity of the models in the ensemble. Their focus was on an “anti-regularization” of the model’s weights resulting in weights with high magnitudes without sacrificing the model’s performance. Although, by construction, Conflictual loss aims at creating diversity in the ensemble outputs, we expect each model in the ensemble to adjust its weights accordingly.

5 Empirical Analysis

We conducted several experiments to assess to what extent both principles of epistemic uncertainty discussed in Sect. 3 are verified. We compared *Conflictual Deep Ensemble* to *MC-Dropout* [6], *Deep Ensemble* [14] and *EDL* [28]. In addition, we tested *Label Smoothing* (LS) [30] in combination with MC-Dropout. Moreover, we evaluated *Confidence Penalty* regularization combined with MC-Dropout and reported the results in Appendix 1.C due to the space limit.⁴

For the varying number of samples used to train the models, we considered, after a 20% validation-train split identical for all models, fractions of the entire training set that grow exponentially from 0.005 to 1. To evaluate the *data-related principle*, we made sure that by increasing the training set, new examples are added to the previous training set rather than randomly selecting a new independent subset from the entire training set. Additionally, for a fixed ratio, we emphasize that the models are trained on the same samples.

We used Multilayer Perceptron (MLP) models with two hidden layers for a straightforward control of its size: since layers are dense, they are invariant under the permutation of neurons. As a consequence the submodel relation defined in

⁴ Code available at: <https://github.com/fellajimed/Conflictual-Loss>

Sect. 3 is simplified: given two networks \mathcal{M}_1 and \mathcal{M}_2 of L dense layers whose sizes are respectively $(n_1^{[1]}, \dots, n_1^{[L]})$ and $(n_2^{[1]}, \dots, n_2^{[L]})$,

$$\mathcal{M}_1 \leq_{\mathcal{M}} \mathcal{M}_2 \iff \forall i, n_1^{[i]} \leq n_2^{[i]}$$

We thus consider for every method a chain of submodels whose sizes of the hidden layers grow exponentially, starting from (128, 64) neurons up to (2048, 1024). Epistemic uncertainty is estimated using 20 forward passes at inference time for MC-Dropout and 10 MLPs in the case of Deep Ensemble. We set the order $k = 1$ for the Conflictual DE so that it also contains $C = 10$ MLPs. Each hidden layer is followed by a Dropout layer ($p = 0.3$) and a ReLU activation function.

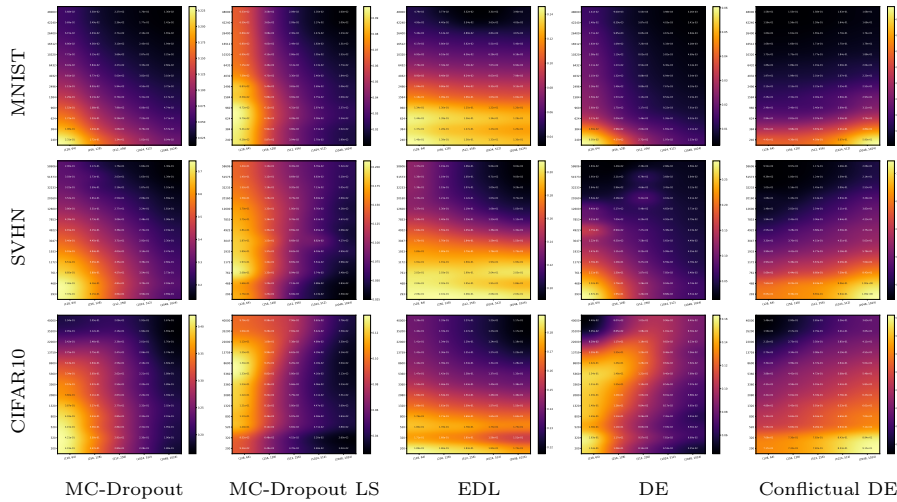


Fig. 1: Heatmaps of epistemic uncertainty (mutual information) on MNIST, SVHN, and CIFAR10 datasets; for MC-Dropout, label smoothing combined with MC-Dropout (MC-Dropout LS), EDL, Deep Ensembles (DE), and Conflictual DE. For each heatmap, the x-axis gives the sizes of the hidden layers and the y-axis gives the number of training samples. Both have logarithmic scales. Color scales are different. Epistemic uncertainty should decrease along the y-axis (data-related principal) and increase along the x-axis (model-related principal).

These models are tested on three datasets: MNIST, SVHN and CIFAR10. For the former, we apply the MLPs to the raw images, whereas for the latter, image embeddings are computed using a pre-trained ResNet34 model and used as inputs for the MLPs. In all the heatmaps, we represent the sizes of the hidden layers on the x-axis, while the y-axis corresponds to the length of the training set. Empirically, we set the value of λ in Eq. 1 to 0.05. The hyperparameters of label smoothing, confidence penalty, and EDL are taken from their respective

papers [30,18,26] and are set to 0.1 (for the first two) and 0.01 for EDL. We refer the reader to Appendix 1.A for more details.

		MC-Dropout	MC-Dropout LS	EDL	DE	Conflictual DE
E.U.	MNIST	0.067	0.040	0.087	0.014	0.245
	SVHN	0.348	0.099	0.152	0.079	0.427
	CIFAR10	0.258	0.068	0.149	0.102	0.507
Acc. \uparrow	MNIST	93.30%	93.90%	91.46%	93.52%	93.86%
	SVHN	68.10%	68.58%	63.46%	69.00%	69.32%
	CIFAR10	60.71%	60.56%	60.93%	61.62%	61.84%
Brier score \downarrow	MNIST	0.101	0.216	0.164	0.102	0.094
	SVHN	0.457	0.498	0.491	0.470	0.419
	CIFAR10	0.576	0.572	0.540	0.589	0.528
SCE \downarrow	MNIST	0.0056	0.0661	0.0335	0.0076	0.0120
	SVHN	0.0289	0.0485	0.0276	0.0345	0.0158
	CIFAR10	0.0377	0.0435	0.0235	0.0444	0.0194
OOD \uparrow	MNIST	82.72%	58.70%	81.85%	88.33%	86.18%
	SVHN	78.59%	85.09%	60.39%	77.20%	78.22%
	CIFAR10	68.81%	63.22%	70.29%	69.09%	74.20%
Mis. \uparrow	MNIST	92.54%	73.78%	87.83%	93.44%	93.16%
	SVHN	80.11%	64.24%	78.00%	82.63%	78.21%
	CIFAR10	73.51%	58.13%	72.36%	74.79%	74.09%

Table 1: Average value per heatmap for different metrics: E.U. (Epistemic uncertainty, Fig 1), Acc. (accuracy, Fig 4), Brier score (Fig 5), SCE (Fig 3), OOD detection (Fig 2) and Mis. (misclassification, Fig 7). *Warning:* if the values of accuracy seem low, it is because they are average values computed over the entire accuracy heatmap, including very small datasets and models.

Figure 1 represents by heatmaps the average of mutual information estimated on the test set as a function of data and model sizes, while table 2 summarizes these heatmaps, indicating the frequencies with which methods comply with the two principles. As aforementioned, the mutual information validates the two principals in expectation. Therefore, we report the average of the test set as a Monte Carlo estimate of this expectation. We see that all methods but *Conflictual DE* are in complete contradiction with the *model-related principle*: the larger the model, the smaller the epistemic uncertainty when precisely the opposite is expected. The *data-related principle* seems better respected, but not perfectly, especially on CIFAR10. In line with comments of Sect. 4, *MC-Dropout LS* is the method that most often breaks the first principle, even if none of these methods achieves a perfect score. Although the exact cause of these phenomena remains unknown, the partial violation of the first principle allows us to circumscribe the origin of the problem. According to property 1, we know

		MC-Dropout	MC-Dropout LS	EDL	DE	Conflictual DE
1 st principle	MNIST	100%	78.33%	91.67%	100%	100%
	SVHN	91.67%	55%	88.33%	88.33%	96.67%
	CIFAR10	76.67%	38.33%	91.67%	41.67%	98.33%
2 nd principle	MNIST	0%	0%	5.77%	0%	96.15%
	SVHN	0%	0%	15.38%	17.30%	98.08%
	CIFAR10	11.54%	0%	5.77%	15.38%	100%

Table 2: Aggregation of the heatmaps in Fig. 1 showing to which degree the two principles are satisfied. The percentages for the first (resp. second) principle represent the frequencies with which jumping to the next larger dataset (resp. to the next smaller model) results in a decrease of mutual information.

mutual information necessarily satisfies the data-related property in expectation when estimated on the exact posterior distribution. Since the violation of this principle is not related to the metric, the only possible cause is a poor quality of the posterior approximation, due to *procedural variability*, *i.e.*, convergence randomness during stochastic gradient descent.

In comparison, *Conflictual DE* obtains almost perfect scores for both principles, proving the conflictual loss’s regularizing effect. While the method was specifically designed to satisfy the first principle, compliance with the second principle was not expected at such level. It seems that the expressive power of networks serves as an amplifier for the discord sown between classifiers by the conflictual loss.

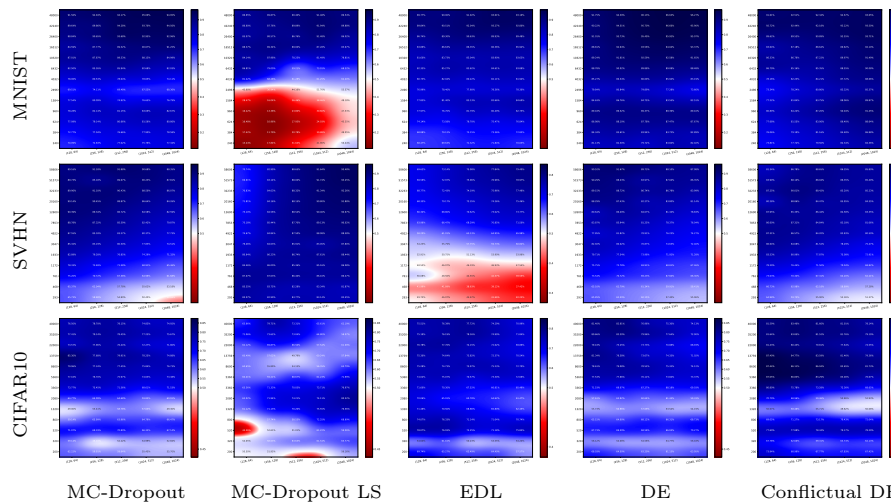


Fig. 2: Heatmaps of AUROC for OOD detection based on epistemic uncertainty. Same representation as Fig. 1 but with the same color scale per dataset.

As shown in Appendix 1.B and summarized in Tab. 1, the calibrated epistemic uncertainty resulting from Conflictual DE does not come at the expense of the performance of the model. In fact, Conflictual DE yields comparable or even superior performance compared to DE, as it performs the best overall on CIFAR10 in terms of accuracy (Fig. 4) and has comparable results to the best method on MNIST. When taking into account the accuracy of the output probabilities with the Brier score, Conflictual DE appears to be the best model on both datasets.

We also checked the quality of epistemic uncertainty to discriminate between OOD and in-distribution (ID) examples, since epistemic uncertainty is expected to be higher for OOD samples than for the ID samples. We use FashionMNIST, SVHN, and CIFAR10 as OOD datasets for the models trained on MNIST, CIFAR10, and SVHN respectively. As shown in Fig. 2, Conflictual DE yields high AUROC overall on CIFAR10 and in the low-data regime on MNIST. MC-Dropout LS performs the worst on MNIST and CIFAR10 at the task of OOD detection as it sometimes results in lower epistemic uncertainty in OOD samples than ID samples. To some extent, we notice the same results in the task of misclassification detection which consists in distinguishing between the correctly classified samples and the misclassified samples based on epistemic uncertainty (see Appendix 1.B: Fig. 7).

Finally, we take a look at the calibration of the models with different methods. As shown in Fig. 3, the SCE with Conflictual DE is the most consistent and the lowest, yielding calibrated models even with small training sets.

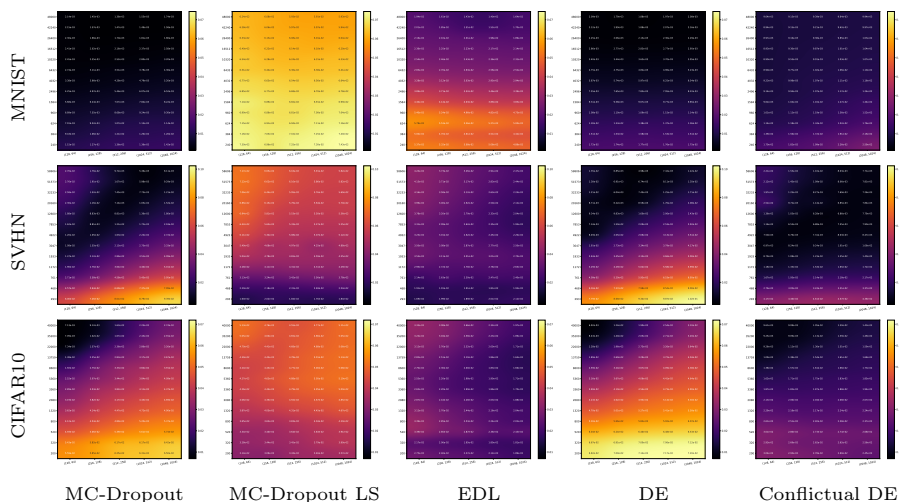


Fig. 3: Heatmaps of the SCE. Same representation as Fig. 2. Lower is better.

As the number of hyperparameters to be set should be kept to a minimum, we wondered whether the hyperparameter λ introduced by Conflictual DE could re-

place advantageously that of weight-decay. Therefore we carried out experiments without weight-decay to study its effect on performance (Appendix 1.D). While some methods see their performance deteriorate, the results with Conflictual DE are more or less the same. This robustness suggests that Conflictual DE can be used without weight-decay, resulting in a zero balance for the number of hyperparameters. To conclude this section, we provide a qualitative and comparative summary of methods on Tab. 3.

	MC-Dropout	MC-Dropout LS	EDL	DE	Conflictual DE
1 st principle	✓	✗	✓	✗	✓
2 nd principle	✗	✗	✗	✗	✓
Accuracy	++	++	+	++	++
Brier score	mixed	mixed	+	mixed	++
Calibration	mixed	–	+	mixed	++
OOD	+	mixed	mixed	++	++
Misclassification	+	–	+	++	++

Table 3: A comparative summary of the performance of the tested methods. Heatmaps of accuracy, Brier score, and misclassification can be found in Fig. 4, Fig. 5 and Fig. 7, respectively (see Appendix).

6 Conclusion

We have shown in this paper that, contrary to expectations, epistemic uncertainty as produced by state-of-the-art models, does not decrease steadily as the training data increases or as the model complexity decreases. We then introduced *conflictual deep ensembles* and showed that they restore not only the first but both principles of epistemic uncertainty, without compromising performance.

Still, this work raises several questions and many perspectives of research: The exact reasons why epistemic uncertainty paradoxically collapses when network complexity grows, have yet to be found. While conflictual deep ensembles have been specially designed to satisfy the first principle, it is surprising how well this technique solves the second principle as well. Why is this so? Although the conflictual loss naturally suits deep ensembles, nothing is preventing it from being applied to BNNs, provided that the samples of the prior can be efficiently partitioned into class-specific subsets. The validity of such an approach remains to be demonstrated. The question remains whether the observed phenomena generalize to models more complex than MLPs or to other problems like regression. We hope that these and other perspectives will entail further studies on this subject.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bengs, V., Hüllermeier, E., Waegeman, W.: Pitfalls of epistemic uncertainty quantification through loss minimisation. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) *Advances in Neural Information Processing Systems* (2022), https://openreview.net/forum?id=epjxT_ARZW5
2. de Mathelin, A., Deheeger, F., Mougeot, M., Vayatis, N.: Deep Anti-Regularized Ensembles provide reliable out-of-distribution uncertainty quantification (Apr 2023)
3. de Mathelin, A., Deheeger, F., Mougeot, M., Vayatis, N.: Maximum Weight Entropy (Sep 2023)
4. Depeweg, S., Hernandez-Lobato, J.M., Doshi-Velez, F., Udluft, S.: Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In: *Proceedings of the 35th International Conference on Machine Learning*. pp. 1184–1193. PMLR (Jul 2018), <https://proceedings.mlr.press/v80/depeweg18a.html>, iISSN: 2640-3498
5. Fortuin, V.: Priors in bayesian deep learning: A review. *International Statistical Review* **90**(3), 563–591 (2022). <https://doi.org/https://doi.org/10.1111/insr.12502>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12502>
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *International Conference on Machine Learning*. pp. 1050–1059 (Jun 2016), <http://proceedings.mlr.press/v48/gal16.html>
7. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian Active Learning with Image Data. In: *Proceedings of the 34th International Conference on Machine Learning*. pp. 1183–1192. PMLR (Jul 2017), <https://proceedings.mlr.press/v70/gal17a.html>, iISSN: 2640-3498
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *International conference on machine learning*. pp. 1321–1330. PMLR (2017)
9. Hinton, G.E., Camp, D.v.: Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In: Pitt, L. (ed.) *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory, COLT 1993*, Santa Cruz, CA, USA, July 26-28, 1993. pp. 5–13. ACM (1993). <https://doi.org/10.1145/168304.168306>, <https://doi.org/10.1145/168304.168306>
10. Huang, Z., Lam, H., Zhang, H.: Efficient uncertainty quantification and reduction for over-parameterized neural networks. In: *Thirty-seventh Conference on Neural Information Processing Systems* (2023), <https://openreview.net/forum?id=6vnwhzRinw>
11. Jeffreys, H.: An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **186**(1007), 453–461 (1946), <http://www.jstor.org/stable/97883>, publisher: The Royal Society
12. Kirsch, A., Mukhoti, J., van Amersfoort, J., Torr, P.H.S., Gal, Y.: On pitfalls in ood detection: Entropy considered harmful (2021), *uncertainty & Robustness in Deep Learning Workshop, ICML*

13. Kuleshov, V., Fenner, N., Ermon, S.: Accurate Uncertainties for Deep Learning Using Calibrated Regression. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2796–2804. PMLR (Jul 2018), <https://proceedings.mlr.press/v80/kuleshov18a.html>, iSSN: 2640-3498
14. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), <https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html>
15. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), <https://openreview.net/forum?id=ryiAv2xAZ>
16. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), <https://openreview.net/forum?id=ryiAv2xAZ>
17. MacKay, D.J.C.: A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation* 4(3), 448–472 (May 1992). <https://doi.org/10.1162/neco.1992.4.3.448>, <https://direct.mit.edu/neco/article/4/3/448-472/5654>
18. Malinin, A., Gales, M.: Predictive Uncertainty Estimation via Prior Networks. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), <https://proceedings.neurips.cc/paper/2018/hash/3ea2db50e62ceefceaf70a9d9a56a6f4-Abstract.html>
19. Malinin, A., Ragni, A., Knill, K., Gales, M.: Incorporating uncertainty into deep learning for spoken language assessment. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 45–50. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). <https://doi.org/10.18653/v1/P17-2008>, <https://aclanthology.org/P17-2008>
20. Meister, C., Salesky, E., Cotterell, R.: Generalized entropy regularization or: There’s nothing special about label smoothing. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6870–6886. Association for Computational Linguistics, Online (Jul 2020). <https://doi.org/10.18653/v1/2020.acl-main.615>, <https://aclanthology.org/2020.acl-main.615>
21. Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks (2024)
22. Neal, R.M.: Bayesian Learning for Neural Networks, Lecture Notes in Statistics, vol. 118. Springer New York, New York, NY (1996). <https://doi.org/10.1007/978-1-4612-0745-0>, <http://link.springer.com/10.1007/978-1-4612-0745-0>
23. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
24. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA (2019)

25. Pakdaman Naeini, M., Cooper, G., Hauskrecht, M.: Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence* **29**(1) (Feb 2015). <https://doi.org/10.1609/aaai.v29i1.9602>, <https://ojs.aaai.org/index.php/AAAI/article/view/9602>
26. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G.E.: Regularizing neural networks by penalizing confident output distributions. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net (2017), <https://openreview.net/forum?id=HyhbYrGYe>
27. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., DePristo, M.A., Dillon, J.V., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, December 8-14, 2019, Vancouver, BC, Canada. pp. 14680–14691 (2019)
28. Sensoy, M., Kaplan, L.M., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, December 3-8, 2018, Montréal, Canada. pp. 3183–3193 (2018), <https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html>
29. Silva, F.L.D., Hernandez-Leal, P., Kartal, B., Taylor, M.E.: Uncertainty-Aware Action Advising for Deep Reinforcement Learning Agents. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(04), 5792–5799 (Apr 2020). <https://doi.org/10.1609/aaai.v34i04.6036>, <https://ojs.aaai.org/index.php/AAAI/article/view/6036>, number: 04
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2818–2826 (2016)
31. Tsiligkaridis, T.: Information aware max-norm dirichlet networks for predictive uncertainty estimation. *Neural Networks* **135**, 105–114 (2021)
32. Wimmer, L., Sale, Y., Hofman, P., Bischl, B., Hüllermeier, E.: Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. pp. 2282–2292. PMLR (Jul 2023), <https://proceedings.mlr.press/v216/wimmer23a.html>, iISSN: 2640-3498
33. Yao, J., Pan, W., Ghosh, S., Doshi-Velez, F.: Quality of Uncertainty Quantification for Bayesian Neural Network Inference (Jun 2019). <https://doi.org/10.48550/arXiv.1906.09686>, <http://arxiv.org/abs/1906.09686>, arXiv:1906.09686 [cs, stat]
34. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. p. 609–616. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
35. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p. 694–699. KDD '02, Association for Computing Machinery, New York, NY, USA (2002). <https://doi.org/10.1145/775047.775151>, <https://doi.org/10.1145/775047.775151>

Appendix 1.A Implementation details

Datasets. As detailed in the paper, the presented models were trained on MNIST and CIFAR10. A 20% validation-train split is first applied and then subsets were taken from the train sets (a total size of 48000 for MNIST, 58606 for SVHN, and 40000 for CIFAR10) for training. We made sure that the subsets used were balanced. CIFAR10 is encoded using a pre-trained ResNet34 model which is equivalent to the training of a ResNet34 model where the feature blocks are fixed and only the classification part is learned.

Data transformations. We apply a standard normalization (mean of 0 and standard deviation of 1) to the datasets. The same transformation is applied then to the test samples (whether there are ID or OOD samples). Only the training samples are used to train the models and no data augmentation is applied.

Training. The models were trained for 500 epochs on MNIST, 600 epochs on SVHN and 700 on CIFAR10. We used the SGD optimizer with weight decay, parameterized with (learning rate, momentum): (0.01, 0.95) for MNIST, (0.02, 0.95) for SVHN, and (0.04, 0.9) for CIFAR10. Each ensemble was trained on a single GPU and the best model (based on the validation loss) was tracked during training and used for early stopping and the learning rate scheduler.

Duration. The experiments took a total of 311, 186, and 70 GPU hours on MNIST, SVHN, and CIFAR10 respectively. The training was done on a cluster with several GPUs. See Tab. 4 for more details. The difference is mainly due to how CIFAR10 experiments are implemented: we first compute the embeddings of dimension 512 (once) using a pre-trained ResNet34 which are stored on disk.

	MC-Dropout	MC-Dropout LS	EDL	DE	Conflictual DE
MNIST	21.75	20.19	41.35	111	116.73
SVHN	12.36	13.37	14	68.33	76.40
CIFAR10	6.83	6.64	7.01	23.85	26

Table 4: Training duration for each method and each dataset in GPU hours.

NOTE: the MNIST experiments took longer than the experiments of SVHN and CIFAR10 due to 2 main reasons:

- We don't take into account the time needed to compute and save the embeddings for CIFAR10 and which is done only once.
- Most importantly, the format of the files for the CIFAR10' embeddings was optimized and thus it is faster. We further apply the same optimization to the MNIST dataset (by using an *identity* "embedding") and MNIST experiments should run faster with the new changes. This format was applied to SVHN which explains the performance gains. Refer to the code for more details.

The execution time for MNIST, after changes, should be less than the training time of SVHN.

Implementation. All experiments were implemented with PyTorch. Code available at: <https://github.com/fellajimed/Conflictual-Loss>

Appendix 1.B Additional results

We report additional metrics for the experiments of Sec 5 and we use the same representation: for each heatmap, the hidden layers on the x-axis and the number of samples used for training on the y-axis. They both have logarithmic scales. Results are on the same datasets and methods.

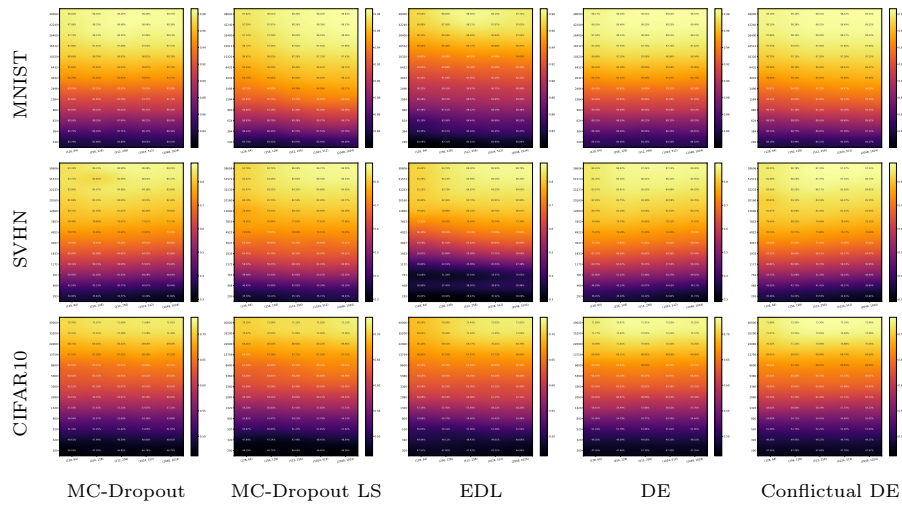


Fig. 4: Heatmaps of the accuracy. Color scales are the same per dataset.

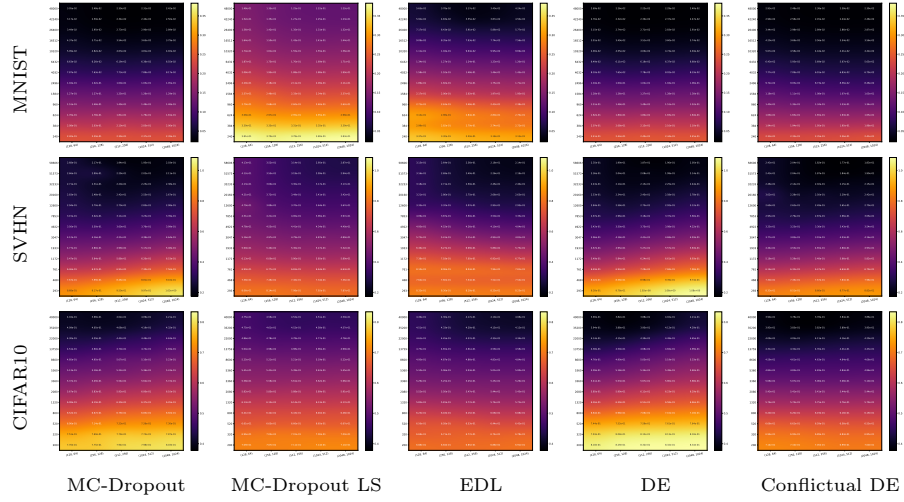


Fig. 5: Heatmaps of the Brier score. Color scales are the same per dataset.

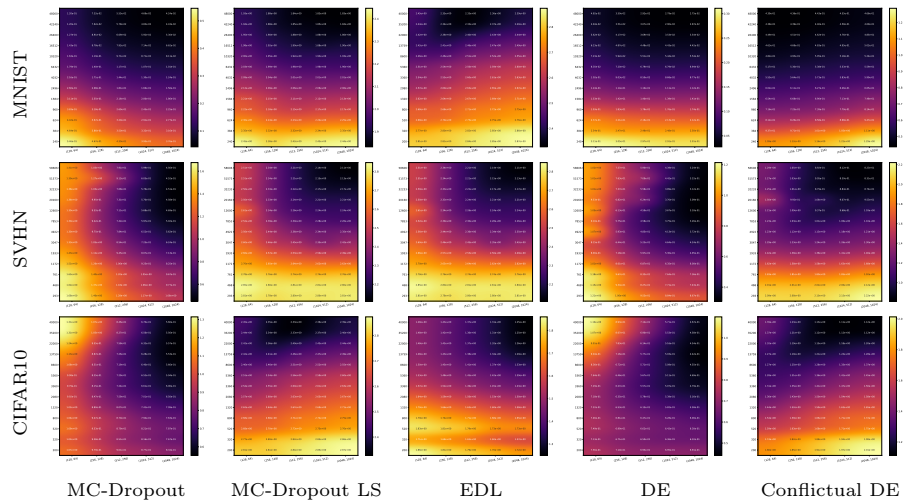


Fig. 6: Heatmaps of total uncertainty. Color scales are different for each heatmap.

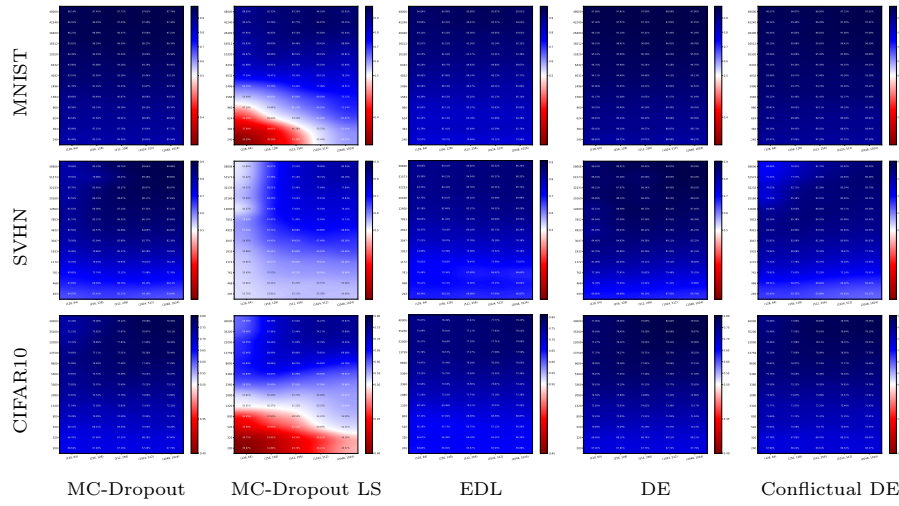
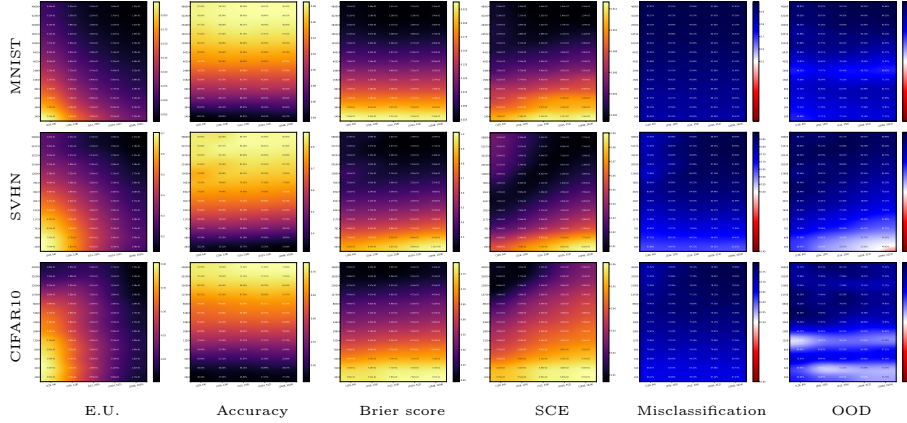


Fig. 7: Heatmaps of AUROC based on epistemic uncertainty for misclassification detection. Color scales are the same per dataset.

Appendix 1.C Results for Confidence Penalty

In this section, we report the results of Sect. 5 in the case of MC-Dropout with confidence penalty (MC-Dropout CP). We notice that, for both MNIST and CIFAR10, the results are comparable to those on MC-Dropout trained only with cross-entropy loss. The color scales are set per heatmap.



	E.U.	Accuracy	Brier score	SCE	Misclassification	OOD
MNIST	0.065	93.34%	0.099	0.0052	92.50%	82.62%
CIFAR10	0.241	60.71%	0.569	0.0354	73.28%	69.29%

Table 5: Aggregation of the results with MC-Dropout CP. Same metrics as in Table 1.

Appendix 1.D Models without weight decay

We also tested the same setup presented in Sect. 5 but without weight decay. We use the same representation as in Sect. 5 and Appendix 1.B. We notice that Conflictual DE stays consistent in all metrics: accuracy, calibration, Brier score, OOD detection, and misclassification detection. In addition, and most importantly, the 2 principles are verified whereas, for example, both are uncorroborated in the case of DE trained on CIFAR10. Furthermore, the model-related principle is not verified in any other method.

Overall, the results with weight decay are relatively better.

		MC-Dropout	EDL	DE	Conflictual DE
E.U.	MNIST	0.068	0.086	0.014	0.245
	CIFAR10	0.299	0.137	0.156	0.544
Acc. \uparrow	MNIST	93.31%	91.45%	93.52%	93.86%
	CIFAR10	60.34%	60.89%	61.61%	61.82%
Brier score \downarrow	MNIST	0.101	0.164	0.102	0.094
	CIFAR10	0.589	0.539	0.598	0.532
SCE \downarrow	MNIST	0.0057	0.0329	0.0076	0.0119
	CIFAR10	0.0418	0.0187	0.0474	0.0207
Mis. \uparrow	MNIST	92.55%	87.86%	93.43%	93.09%
	CIFAR10	73.75%	72.98%	74.84%	74.13%
OOD \uparrow	MNIST	82.85%	81.85%	88.19%	86.50%
	CIFAR10	67.49%	69.00%	67.68%	73.06%

Table 6: Average value per heatmap for different metrics without weight decay. Same metrics as in Table 1.

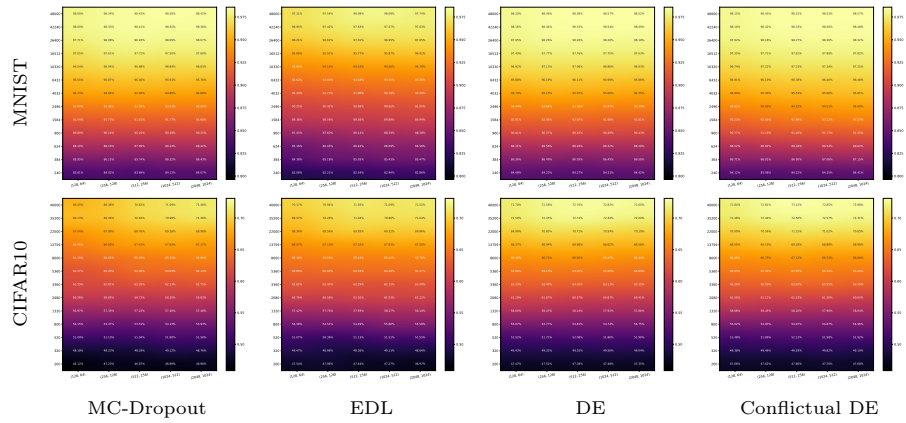


Fig. 8: Heatmaps of Accuracy. Color scales are the same per dataset.

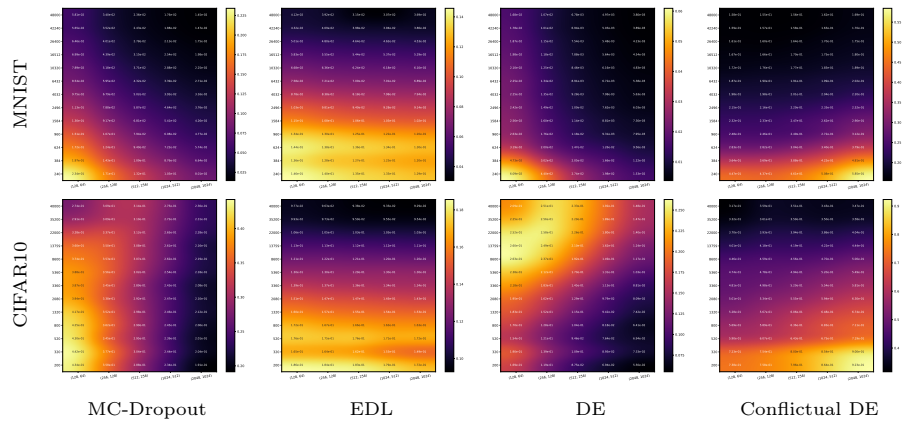


Fig. 9: Heatmaps of epistemic uncertainty (mutual information). Color scales are different for each heatmap.

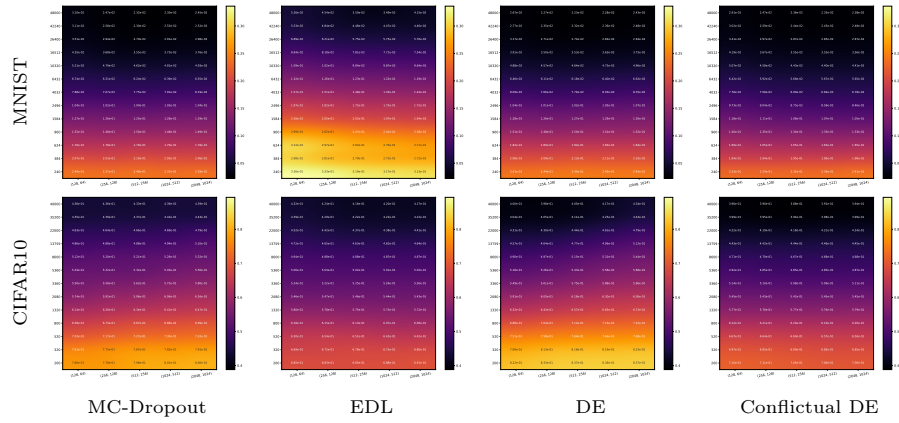


Fig. 10: Heatmaps of Brier score Color scales are the same per dataset.

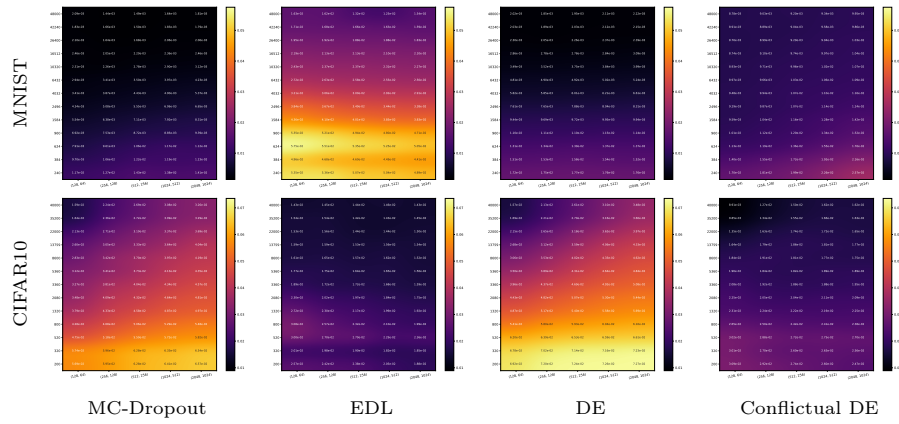


Fig. 11: Heatmaps of SCE. Color scales are the same per dataset.

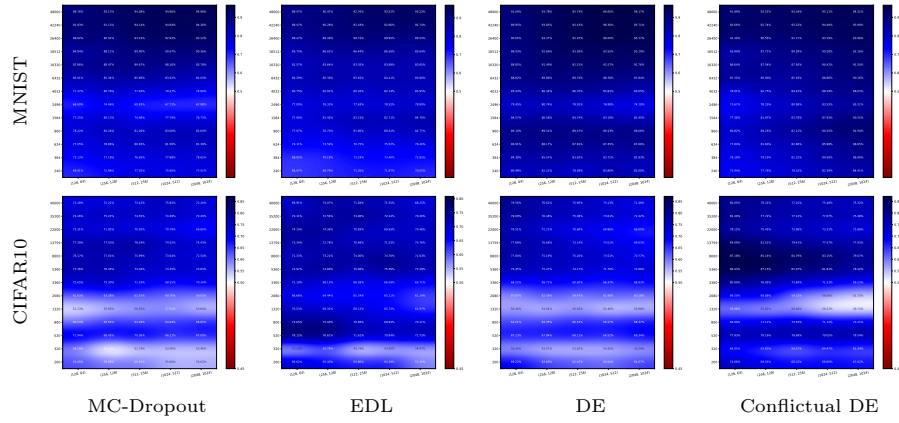


Fig. 12: Heatmaps of AUROC based on epistemic uncertainty for OOD detection. Color scales are the same per dataset.

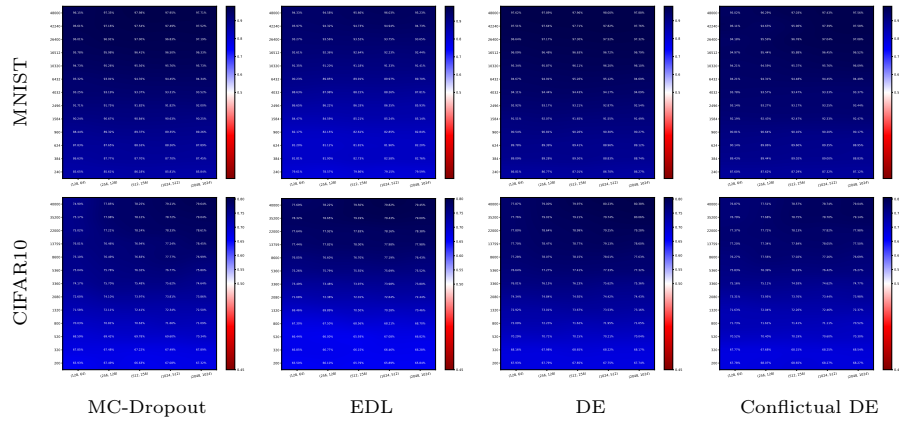


Fig. 13: Heatmaps of AUROC based on epistemic uncertainty for misclassification detection. Color scales are the same per dataset.