# Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable?

Anna-Kathrin Kopetzki *[1] Bertrand Charpentier *[1] Daniel Zügner [1] Sandhya Giri [1] Stephan Günnemann [1]

## Abstract

Dirichlet-based uncertainty (DBU) models are a recent and promising class of uncertainty-aware models. DBU models predict the parameters of a Dirichlet distribution to provide fast, high-quality uncertainty estimates alongside with class predictions. In this work, we present the first large-scale, in-depth study of the robustness of DBU models under adversarial attacks. Our results suggest that uncertainty estimates of DBU models are not robust w.r.t. three important tasks: **(1)** indicating correctly and wrongly classified samples; **(2)** detecting adversarial examples; and **(3)** distinguishing between in-distribution (ID) and out-of-distribution (OOD) data. Additionally, we explore the first approaches to make DBU models more robust. While adversarial training has a minor effect, our median smoothing based approach significantly increases robustness of DBU models.

## 1. Introduction

Neural networks achieve high predictive accuracy in many tasks, but they are known to have two substantial weaknesses: First, neural networks are not robust against adversarial perturbations, i.e., semantically meaningless input changes that lead to wrong predictions (Szegedy et al., 2014; Goodfellow et al., 2015). Second, standard neural networks are unable to identify samples that are different from the ones they were trained on and tend to make over-confident predictions at test time (Lakshminarayanan et al., 2017). These weaknesses make them impracticable in sensitive domains like financial, autonomous driving or medical areas which require trust in predictions.

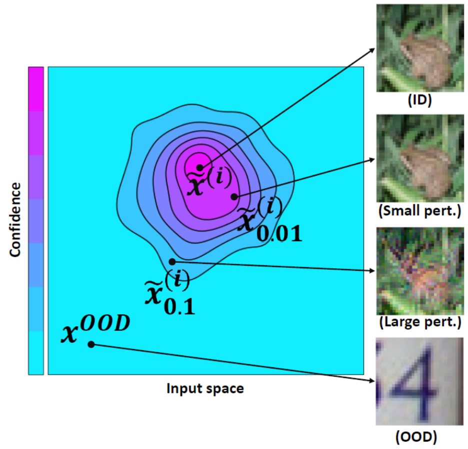


*Figure 1.* Visualization of the desired uncertainty estimates.

To increase trust in neural networks, models that provide predictions along with the corresponding uncertainty have been proposed. An important, quickly growing family of models is the Dirichlet-based uncertainty (DBU) family (Malinin & Gales, 2018a; 2019; Sensoy et al., 2018; Malinin et al., 2019; Charpentier et al., 2020; Zhao et al., 2020; Nandy et al., 2020; Shi et al., 2020; Sensoy et al., 2020). Contrary to other approaches such as Bayesian Neural Networks (Blundell et al., 2015; Osawa et al., 2019; Maddox et al., 2019), drop out (Gal & Ghahramani, 2016) or ensembles (Lakshminarayanan et al., 2017), DBU models provide efficient uncertainty estimates at test time in a single forward pass by directly predicting the parameters of a Dirichlet distribution over categorical probability distributions. DBU models have the advantage that they provide both, aleatoric uncertainty estimates resulting from irreducible uncertainty (e.g. class overlap or noise) and epistemic uncertainty estimates resulting from the lack of knowledge about unseen data (e.g. an unknown object is presented to the model). Both uncertainty types can be quantified from Dirichlet distributions using different uncertainty measures such as differential entropy, mutual information, or pseudo-counts. These uncertainty measures show outstanding performance in, e.g., the detection of OOD samples and thus are superior to softmax based confidence (Malinin & Gales, 2018a; 2019; Charpentier et al., 2020).

*Equal contribution [1]Technical University of Munich, Germany; Department of Informatics. Correspondence to: Anna-Kathrin Kopetzki <kopetzki@in.tum.de>, Bertrand Charpentier <charpent@in.tum.de>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

Neural networks from the families outlined above are expected to *know what they don't know*, i.e. they are supposed to notice when they are unsure about a prediction. This raises questions with regards to adversarial examples: should uncertainty estimation methods *detect* these corrupted samples by indicating a high uncertainty on them and abstain from making a prediction? Or should uncertainty estimation be *robust* to adversarial examples and assign the correct label even under perturbations? We argue that being robust to adversarial perturbations is the best option (see Figure 1) for two reasons. First, in image classification a human is usually not able to observe any difference between an adversarial example and an unperturbed image. Second, the size of the perturbation corresponding to a good adversarial example is typically small w.r.t. the $L_p$-norm and thus assumed to be semantically meaningless. Importantly, robustness should not only be required for the class predictions, but also for the uncertainty estimates. This means that DBU models should be able to distinguish robustly between ID and OOD data even if those are perturbed.

In this work, we focus on DBU models and analyze their robustness capacity w.r.t. class predictions as well as uncertainty predictions. In doing so, we go beyond simple softmax output confidence by investigating advanced uncertainty measures such as differential entropy. Specifically, we study the following questions:

1. *Is low uncertainty a reliable indicator of correct predictions?*
2. *Can we use uncertainty estimates to detect label attacks on the class prediction?*
3. *Are uncertainty estimates such as differential entropy a robust feature for OOD detection?*

In addressing these questions we place particular focus on adversarial perturbations of the input to evaluate the *worst case* performance of the models on increasing complex data sets and attacks. We evaluate robustness of DBU models w.r.t. to these three questions by comparing their performance on unperturbed and perturbed inputs. Perturbed inputs are obtained by computing *label attacks* and *uncertainty attacks*, which are a new type of attacks we propose. While label attacks aim at changing the class prediction, uncertainty attacks aim at changing the uncertainty estimate such that ID data is marked as OOD data and vice versa. In total, we performed more than $138,800$ attack settings to explore the robustness landscape of DBU models. Those settings cover different data sets, attack types, attack losses, attack radii, DBU model types and initialisation seeds. Finally, we propose and evaluate median smoothing and adversarial training based on label attacks and uncertainty attacks to make DBU models more robust. Our median smoothing approach provides certificates on epistemic uncertainty measures such as differential entropy and allows to certify uncertainty estimation. The code and further supplementary material is available online (`www.daml.in.tum.de/dbu-robustness`).

## 2. Related work

The existence of adversarial examples is a problematic property of neural networks (Szegedy et al., 2014; Goodfellow et al., 2015). Previous works have study this phenomena by proposing adversarial attacks (Carlini & Wagner, 2017; Brendel et al., 2018; Zügner et al., 2018), defenses (Cisse et al., 2017; Gu & Rigazio, 2015) and verification techniques (Wong & Kolter, 2018; Singh et al., 2019; Cohen et al., 2019; Bojchevski et al., 2020; Kopetzki & Günnemann, 2021). This includes the study of different settings such as i.i.d. inputs, sequential inputs and graphs (Zheng et al., 2016; Bojchevski & Günnemann, 2019; Cheng et al., 2020; Schuchardt et al., 2021).

In the context of uncertainty estimation, robustness of the class prediction has been studied in previous works for Bayesian Neural Networks (Blundell et al., 2015; Osawa et al., 2019; Maddox et al., 2019), drop out (Gal & Ghahramani, 2016) or ensembles (Lakshminarayanan et al., 2017) focusing on data set shifts (Ovadia et al., 2019) or adversarial attacks (Carbone et al., 2020; Cardelli et al., 2019; Wicker et al., 2020). Despite their efficient and high quality uncertainty estimates, the robustness of DBU models has not been investigated in detail yet — indeed only for one single DBU model, (Malinin & Gales, 2019) has briefly performed attacks aiming to change the label. In contrast, our work focuses on a large variety of DBU models and analyzes two robustness properties: robustness of the class prediction w.r.t. adversarial perturbations and robustness of uncertainty estimation w.r.t. our newly proposed attacks against uncertainty measures.

This so called *uncertainty attack* directly targets uncertainty estimation and are different from traditional *label attacks*, which target the class prediction (Madry et al., 2018; Dang-Nhu et al., 2020). They allow us to jointly evaluate robustness of the class prediction and robustness of uncertainty estimation. This goes beyond previous attack defenses that were either focused on evaluating *robustness w.r.t. class predictions* (Carlini & Wagner, 2017; Weng et al., 2018) or detecting attacks against the class prediction (Carlini & Wagner, 2017).

Different models have been proposed to account for uncertainty while being robust. (Smith & Gal, 2018) and (Lee et al., 2018) have tried to improve label attack detection based on uncertainty using drop-out or density estimation. In addition to improving label attack detection for large unseen perturbations, (Stutz et al., 2020) aimed at improving robustness w.r.t. class label predictions on small input perturbations. They used adversarial training and soft labels for adversarial samples further from the original input. (Qin et al., 2020) suggested a similar adversarial training procedure, that softens labels depending on the input robustness. These previous works consider the aleatoric uncertainty that is contained in the predicted categorical probabilities, but in contrast to DBU models they are not capable of taking epistemic uncertainty into account.

Recently, four studies tried to obtain certificates on aleatoric uncertainty estimates. (Tagasovska & Lopez-Paz, 2019) and (Kumar et al., 2020) compute confidence intervals and certificates on softmax predictions. (Bitterwolf et al., 2020) uses interval bound propagation to compute bounds on softmax predictions within the $L_\infty$-ball around an OOD sample using ReLU networks. (Meinke & Hein, 2020) focuses on obtaining certifiably low confidence for OOD data. These four studies estimate confidence based on softmax predictions, which accounts for aleatoric uncertainty only. In this paper, we provide certificates which apply for all uncertainty measures. In particular, we use our certificates on epistemic uncertainty measures such as differential entropy which are well suited for OOD detection.

## 3. Dirichlet-based uncertainty models

Standard (softmax) neural networks predict the parameters of a categorical distribution $\boldsymbol{p}^{(i)} = [p_1^{(i)}, \dots, p_C^{(i)}]$ for a given input $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, where $C$ is the number of classes. Given the parameters of a categorical distribution, the *aleatoric uncertainty* can be evaluated. The aleatoric uncertainty is the uncertainty on the class label prediction $y^{(i)} \in \{1, \dots, C\}$. For example if we predict the outcome of an unbiased coin flip, the model is expected to have high aleatoric uncertainty and predict $p(\text{head}) = 0.5$.

In contrast to standard (softmax) neural networks, DBU models predict the parameters of a Dirichlet distribution – the natural prior of categorical distributions – given input $\boldsymbol{x}^{(i)}$ (i.e. $q^{(i)} = \text{Dir}(\boldsymbol{\alpha}^{(i)})$ where $f_\theta(\boldsymbol{x}^{(i)}) = \boldsymbol{\alpha}^{(i)} \in \mathbb{R}_+^C$). Hence, the *epistemic distribution* $q^{(i)}$ expresses the *epistemic* uncertainty on $\boldsymbol{x}^{(i)}$, i.e. the uncertainty on the categorical distribution prediction $\boldsymbol{p}^{(i)}$. From the epistemic distribution, follows an estimate of the *aleatoric distribution* of the class label prediction $\text{Cat}(\bar{\boldsymbol{p}}^{(i)})$ where $\mathbb{E}_{q^{(i)}}[\boldsymbol{p}^{(i)}] = \bar{\boldsymbol{p}}^{(i)}$. An advantage of DBU models is that one pass through the neural network is sufficient to compute epistemic distribution, aleatoric distribution, and predict the class label:

$$q^{(i)} = \text{Dir}(\boldsymbol{\alpha}^{(i)}), \;\; \bar{p}_c^{(i)} = \frac{\alpha_c^{(i)}}{\alpha_0^{(i)}}, \;\; y^{(i)} = \arg\max_c \, [\bar{p}_c^{(i)}] \tag{1}$$

where $\alpha_0^{(i)} = \sum_{c=1}^{C} \alpha_c^{(i)}$. This parametrization allows to compute classic uncertainty measures in closed-form such as the total pseudo-count $m_{\alpha_0}^{(i)} = \sum_c \alpha_c^{(i)}$, the differential entropy of the Dirichlet distribution $m_{\text{diffE}}^{(i)} = h(\text{Dir}(\boldsymbol{\alpha}^{(i)}))$ or the mutual information $m_{\text{MI}}^{(i)} = I(y^{(i)}, \boldsymbol{p}^{(i)})$ (App. 6.1, (Malinin & Gales, 2018a)). Hence, these measure can efficiently be used to assign high uncertainty to unknown data, which makes DBU models specifically suited for detection of OOD samples.

Several recently proposed models for uncertainty estimations belong to the family of DBU models, such as PriorNet, EvNet, DDNet and PostNet. These models differ in terms of their parametrization of the Dirichlet distribution, the training, and density estimation. An overview of theses differences is provided in Table 1. In our study we evaluate all recent versions of these models.

Contrary to the other models, Prior Networks **(PriorNet)** (Malinin & Gales, 2018a; 2019) require OOD data for training to “teach” the neural network the difference between ID and OOD data. PriorNet is trained with a loss function consisting of two KL-divergence terms. The fist term is designed to learn Dirichlet parameters for ID data, while the second one is used to learn a flat Dirichlet distribution for OOD data:

$$\begin{aligned} L_{\text{PriorNet}} = \frac{1}{N} \Bigg[ & \sum_{\boldsymbol{x}^{(i)} \in \text{ID data}} [\text{KL}[\text{Dir}(\alpha^{\text{ID}}) || q^{(i)}]] \\ & + \sum_{\boldsymbol{x}^{(i)} \in OODdata} [\text{KL}[\text{Dir}(\alpha^{\text{OOD}}) || q^{(i)}]] \Bigg] \end{aligned} \tag{2}$$

where $\alpha^{\text{ID}}$ and $\alpha^{\text{OOD}}$ are hyper-parameters. Usually $\alpha^{\text{ID}}$ is set to $1e^1$ for the correct class and $1$ for all other classes, while $\alpha^{\text{OOD}}$ is set to $\mathbf{1}$ for all classes. There a two variants of PriorNet. The first one is trained based on reverse KL-divergence (Malinin & Gales, 2019), while the second one is trained with KL-divergence (Malinin & Gales, 2018a). In our experiments, we include the most recent reverse version of PriorNet, as it shows superior performance (Malinin & Gales, 2019).

Evidential Networks **(EvNet)** (Sensoy et al., 2018) are trained with a loss that computes the sum of squares between the on-hot encoded true label $\boldsymbol{y}*^{(i)}$ and the predicted categorical $\boldsymbol{p}^{(i)}$ under the Dirichlet distribution:

$$L_{\text{EvNet}} = \frac{1}{N} \sum_i \mathbb{E}_{\boldsymbol{p}^{(i)} \sim \text{Dir}(\boldsymbol{\alpha}^{(i)})} || \boldsymbol{y}*^{(i)} - \boldsymbol{p}^{(i)} ||^2 \tag{3}$$

Ensemble Distribution Distillation **(DDNet)** (Malinin et al., 2019) is trained in two steps. First, an ensemble of $M$ classic neural networks needs to be trained. Then, the soft-labels $\{\boldsymbol{p}_m^{(i)}\}_{m=1}^{M}$ provided by the ensemble of networks are

*Table 1.* Summary of DBU models. Further details on the loss functions are provided in the appendix.

| | $\alpha^{(i)}$**-parametrization** | **Loss** | **OOD training data** | **Ensemble training** | **Density estimation** |
|---|---|---|---|---|---|
| **PostNet** | $f_\theta(\boldsymbol{x}^{(i)}) = \boldsymbol{1} + \boldsymbol{\alpha}^{(i)}$ | Bayesian loss | No | No | Yes |
| **PriorNet** | $f_\theta(\boldsymbol{x}^{(i)}) = \boldsymbol{\alpha}^{(i)}$ | Reverse KL | Yes | No | No |
| **DDNet** | $f_\theta(\boldsymbol{x}^{(i)}) = \boldsymbol{\alpha}^{(i)}$ | Dir. Likelihood | No | Yes | No |
| **EvNet** | $f_\theta(\boldsymbol{x}^{(i)}) = \boldsymbol{1} + \boldsymbol{\alpha}^{(i)}$ | Expected MSE | No | No | No |

distilled into a Dirichlet-based network by fitting them with the maximum likelihood under the Dirichlet distribution:

$$L_{\text{DDNet}} = -\frac{1}{N}\sum_{i}\sum_{m=1}^{M}[\ln q^{(i)}(\pi^{im})] \tag{4}$$

where $\pi^{im}$ denotes the soft-label of $m$th neural network.

Posterior Network **(PostNet)** (Charpentier et al., 2020) performs density estimation for ID data with normalizing flows and uses a Bayesian loss formulation:

$$L_{\text{PostNet}} = \frac{1}{N}\sum_{i}\mathbb{E}_{q(p^{(i)})}[\text{CE}(p^{(i)}, y^{(i)})] - H(q^{(i)}) \tag{5}$$

where CE denotes the cross-entropy. All loss functions can be computed in closed-form. For more details please have a look at the original paper on PriorNet (Malinin & Gales, 2018a), PostNet (Charpentier et al., 2020), DDNet (Malinin & Gales, 2019) and EvNet (Sensoy et al., 2018). Note that EvNet and PostNet model the Dirichlet parameters as $f_\theta(\boldsymbol{x}^{(i)}) = 1 + \boldsymbol{\alpha}^{(i)}$ while PriorNet, RevPriorNet and DDNet compute them as $f_\theta(\boldsymbol{x}^{(i)}) = \boldsymbol{\alpha}^{(i)}$.

## 4. Robustness of Dirichlet-based uncertainty models

We analyze robustness of DBU models on tasks in connection with uncertainty estimation w.r.t. the following four aspects: *accuracy*, *confidence calibration*, *label attack detection* and *OOD detection*. Uncertainty is quantified by differential entropy, mutual information or pseudo counts. A formal definition of all uncertainty estimation measures is provided in the appendix (see Section 6.1).

Robustness of Dirichlet-based uncertainty models is evaluated based on *label attacks* and a newly proposed type of attacks called *uncertainty attacks*. While label attacks aim at changing the predicted class, uncertainty attacks aim at changing the uncertainty assigned to a prediction. All previous works are based on label attacks and focus on robustness w.r.t. the class prediction. Thus, we are the first to propose attacks targeting uncertainty estimates such as differential entropy and analyze desirable robustness properties of DBU models beyond the class prediction. Label attacks and uncertainty attacks both compute a perturbed input $\tilde{\boldsymbol{x}}^{(i)}$ close to the original input $\boldsymbol{x}^{(i)}$ i.e. $||\boldsymbol{x}^{(i)} - \tilde{\boldsymbol{x}}^{(i)}||_2 < r$ where $r$ is the attack radius. This perturbed input is obtained by optimizing a loss function $l(\boldsymbol{x})$ using Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD). Furthermore, we include a black box attack setting (Noise) which generates 10 noise samples from a Gaussian distribution, which is centered at the original input. From these 10 perturbed samples we choose the one with the greatest effect on the loss function and use it as attack. To complement attacks, we compute certificates on uncertainty estimates using median smoothing (yeh Chiang et al., 2020).

The following questions we address by our experiments have a common assessment metric and can be treated as binary classification problems: distinguishing between correctly and wrongly classified samples, discriminating between non-attacked input and attacked inputs or differentiating between ID data and OOD data. To quantify the performance of the models on these binary classification problems, we compute the area under the precision recall curve (AUC-PR).

Experiments are performed on two image data sets (MNIST (LeCun & Cortes, 2010) and CIFAR10 (Krizhevsky et al., 2009)), which contain bounded inputs and two tabular data sets (Segment (Dua & Graff, 2017) and Sensorless drive (Dua & Graff, 2017)), consisting of unbounded inputs. Note that unbounded inputs are challenging since it is impossible to describe the infinitely large OOD distribution. As PriorNet requires OOD training data, we use two further image data sets (FashionMNIST (Xiao et al., 2017) and CIFAR100 (Krizhevsky et al., 2009)) for training on MNIST and CIFAR10, respectively. All other models are trained without OOD data. To obtain OOD data for the tabular data sets, we remove classes from the ID data set (class window for the Segment data set and class 9 for Sensorless drive) and use them as the OOD data. Further details on the experimental setup are provided in the appendix (see Section 6.2).

### 4.1. Uncertainty estimation under label attacks

Label attacks aim at changing the predicted class. To obtain a perturbed input with a different label, we maximize the cross-entropy loss $\tilde{\boldsymbol{x}}^{(i)} \approx \arg\max_{\boldsymbol{x}} l(\boldsymbol{x}) = \text{CE}(\boldsymbol{p}^{(i)}, \boldsymbol{y}^{(i)})$ under the radius constraint. For the sake of completeness we additionally analyze label attacks w.r.t. to their performance of changing the class prediction and the accuracy of the neural network under label attacks constraint by different radii (see Appendix, Table 8). As expected and partially shown by previous works, none of the DBU models is robust against label attacks. However, we note that PriorNet is slightly more robust than the other DBU models. This might be explained by the use of OOD data during training, which can be seen as some kind of robust training. From now on, we switch to the core focus of this work and analyze robustness properties of uncertainty estimation.

**Is low uncertainty a reliable indicator of correct predictions?**

*Expected behavior:* Predictions with low uncertainty are more likely to be correct than high uncertainty predictions. *Assessment metric:* We distinguish between correctly classified samples (label 0) and wrongly classified ones (label 1) based on the differential entropy scores produced by the DBU models (Malinin & Gales, 2018a). Correctly classified samples are expected to have low differential entropy, reflecting the model's confidence, and analogously wrongly predicted samples are expected to have higher differential entropy. *Observed behavior:* Note that the positive and negative class are not balanced, thus, the use of AUC-PR scores (Saito & Rehmsmeier, 2015) are important to enable meaningful measures. While uncertainty estimates are indeed an indicator of correctly classified samples on unperturbed data, none of the models maintains its high performance on perturbed data computed by PGD, FGSM or Noise label attacks (see. Table 2, 14 and 15). Thus, using uncertainty estimates as indicator for correctly labeled inputs is not robust to adversarial perturbations. This result is notable, since the used attacks do not target uncertainty.

**Can uncertainty estimates be used to detect label attacks against the class prediction?**

*Expected behavior:* Adversarial examples are not from the natural data distribution. Therefore, DBU models are expected to detect them as OOD data by assigning them a higher uncertainty. We expect that perturbations computed based on a bigger attack radius $r$ are easier to detect as their distance from the data distribution is larger. *Assessment metric:* The goal of attack-detection is to distinguish between unperturbed samples (label 0) and perturbed samples (label 1). Uncertainty on samples is quantified by the differential uncertainty (Malinin & Gales, 2018a). Unperturbed samples are expected to have low differential entropy, because they are from the same distribution as the training data, while perturbed samples are expected to have a high differential entropy. *Observed behavior:* Table 8 shows that the accuracy of all models decreases significantly under PGD label attacks, but none of the models is able to provide an equivalently increasing attack detection rate (see Table 3). Even larger perturbations are hard to detect for DBU models.

Similar results are obtained when we use mutual information or the precision $\alpha_0$ to quantify uncertainty (see appendix Table 13 and 12). Although PGD label attacks do not explicitly consider uncertainty, they seem to generate adversarial examples with similar uncertainty as the original input. Such high-certainty adversarial examples are illustrated in Figure 2, where certainty is visualized based on the precision $\alpha_0$, which is supposed to be high for ID data and low for OOD data. While the original input (perturbation size 0.0) is correctly classified as frog and ID data, there exist adversarial examples that are classified as deer or bird. The certainty ($\alpha_0$-score) on the prediction of these adversarial examples has a similar or even higher value than on the prediction of the original input. Using the differential entropy to distinguish between ID and OOD data results in the same ID/OOD assignment since the differential entropy of the three right-most adversarial examples is similar or even smaller than on the unperturbed input.

Under the less powerful FGSM and Noise attacks (see Appendix), DBU models achieve mostly higher attack detection rates than under PGD attacks. This suggests that uncertainty estimation is able to detect weak attacks, which is consistent with the observations in (Malinin & Gales, 2018b) but fails under stronger PGD attacks.

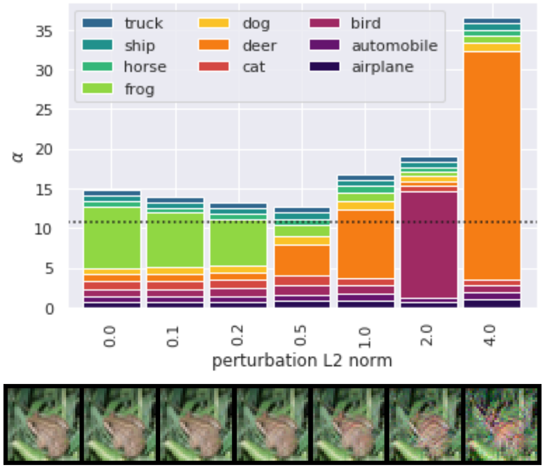


*Figure 2.* Input and predicted Dirichlet-parameters under label attacks (dotted line: threshold to distinguish ID and OOD data).

On tabular data sets, PostNet shows a better label attack detection rate for large perturbations. This observation might be explained by the fact that the density estimation of the ID samples has been shown to work better for tabular data sets (Charpentier et al., 2020). Overall, none of the DBU models provides a reliable indicator for adversarial inputs that target the class prediction.

## 4.2. Attacking uncertainty estimation

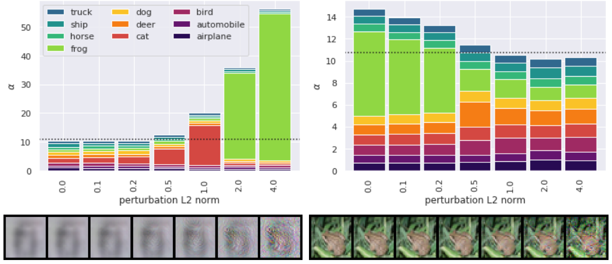
DBU models are designed to provide sophisticated uncertainty estimates (beyond softmax scores) alongside predic-

*Table 2.* Distinguishing between correctly predicted and wrongly predicted labels based on the differential entropy under PGD label attacks (metric: AUC-PR).

| | CIFAR10 | | | | | | Sensorless | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Att. Rad. | 0.0 | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 | 0.0 | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 |
| PostNet | **98.7** | 88.6 | 56.2 | 7.8 | 1.2 | 0.4 | 99.7 | 8.3 | 3.9 | 3.6 | **7.0** | **9.8** |
| PriorNet | 92.9 | 77.7 | 60.5 | **37.6** | **24.9** | **11.3** | 99.8 | 10.5 | 3.2 | 0.7 | 0.2 | 0.2 |
| DDNet | 97.6 | **91.8** | **78.3** | 18.1 | 0.8 | 0.0 | 99.7 | 11.9 | 1.6 | 0.4 | 0.2 | 0.1 |
| EvNet | 97.9 | 85.9 | 57.2 | 10.2 | 4.0 | 2.4 | **99.9** | **22.9** | **13.0** | **6.0** | 3.7 | 3.2 |

*Table 3.* Label Attack-Detection by normally trained DBU models based on differential entropy under PGD label attacks (AUC-PR).

| | CIFAR10 | | | | | Sensorless | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Att. Rad. | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 |
| PostNet | **63.4** | **66.9** | 42.1 | 32.9 | 31.6 | 47.7 | 42.3 | 36.9 | **48.5** | **85.0** |
| PriorNet | 53.3 | 56.0 | 55.6 | **49.2** | 42.2 | 38.8 | 33.6 | 31.4 | 33.1 | 40.9 |
| DDNet | 55.8 | 60.5 | **57.3** | 38.7 | 32.3 | **53.5** | 42.2 | 35.0 | 32.8 | 32.6 |
| EvNet | 48.4 | 46.9 | 46.3 | 46.3 | **44.5** | 48.2 | **42.6** | **38.2** | 36.0 | 37.2 |



(a) OOD uncertainty attack (b) ID uncertainty attack

*Figure 3.* ID and OOD input with corresponding Dirichlet-parameters under uncertainty attacks (dotted line: threshold to distinguish ID and OOD).

tions and use them to detect OOD samples. In this section, we propose and analyze a new attack type that targets these uncertainty estimates. DBU models enable us to compute uncertainty measures i.e. differential entropy, mutual information and precision $\alpha_0$ in closed from (see (Malinin & Gales, 2018a) for a derivation). Uncertainty attacks use this closed form solution as loss function for PGD, FGSM or Noise attacks. Since differential entropy is the most widely used metric for ID-OOD-differentiation, we present results based on the differential entropy loss function $\tilde{\boldsymbol{x}}^{(i)} \approx \arg\max_{\boldsymbol{x}} l(\boldsymbol{x}) = \text{Diff-E}(\text{Dir}(\alpha^{(i)}))$:

$$\begin{aligned}\text{Diff-E}(\text{Dir}(\alpha^{(i)})) = &\sum_{c}^{K} \ln\Gamma(\alpha_c^{(i)}) - \ln\Gamma(\alpha_0^{(i)}) \\ &- \sum_{c}^{K} (\alpha_c^{(i)} - 1)\cdot(\Psi(\alpha_c^{(i)}) - \Psi(\alpha_0^{(i)}))\end{aligned} \tag{6}$$

where $\alpha_0^{(i)} = \sum_c \alpha_c^{(i)}$. Result based on further uncertainty measures, loss functions and more details on attacks are provided in the appendix.

We analyze the performance of DBU models under uncertainty attacks w.r.t. two tasks. First, uncertainty attacks are computed on ID data aiming to indicate it as OOD data, while OOD data is left non-attacked. Second, we attack OOD data aiming to indicate it as ID data, while ID data is not attacked. Hence, uncertainty attacks target at posing ID data as OOD data and vice versa.

**Are uncertainty estimates a robust feature for OOD detection?**
*Expected behavior:* We expect DBU models to be able to distinguish between ID and OOD data by providing reliable uncertainty estimates, even under small perturbations. Thus, we expect uncertainty estimates of DBU models to be robust under attacks. *Assessment metric:* We distinguish between ID data (label 0) and OOD data (label 1) based on the differential entropy as uncertainty scoring function (Malinin & Gales, 2018a). Differential entropy is expected to be small on ID samples and high on OOD samples. Experiments on further uncertainty measure and results on the AUROC metric are provided in the appendix. *Observed behavior:* OOD samples are perturbed as illustrated in Figure 3. Part (a) of the figure illustrates an OOD-samples, that is correctly identified as OOD. Adding adversarial perturbations $\geq 0.5$ changes the Dirichlet parameters such that the resulting images are identified as ID, based on precision or differential entropy as uncertainty measure. Perturbing an ID sample (part (b)) results in images that are marked as

*Table 4.* OOD detection based on differential entropy under PGD uncertainty attacks against differential entropy computed on ID data and OOD data (metric: AUC-PR).

| | ID-Attack (non-attacked OOD) | | | | | | OOD-Attack (non-attacked ID) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Att. Rad. | 0.0 | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 | 0.0 | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 |
| **CIFAR10 – SVHN** | | | | | | | | | | | | |
| PostNet | 81.8 | 64.3 | 47.2 | 22.4 | 17.6 | **16.9** | 81.8 | 60.5 | 40.7 | 23.3 | 21.8 | 19.8 |
| PriorNet | 54.4 | 40.1 | 30.0 | 17.9 | 15.6 | 15.4 | 54.4 | 40.7 | 30.7 | 19.5 | 16.5 | 15.7 |
| DDNet | **82.8** | **71.4** | **59.2** | **28.9** | 16.0 | 15.4 | **82.8** | **72.0** | **57.2** | 20.8 | 15.6 | 15.4 |
| EvNet | 80.3 | 62.4 | 45.4 | 21.7 | **17.9** | 16.5 | 80.3 | 58.2 | 46.5 | **34.6** | **28.0** | **23.9** |
| **Sens. – Sens. class 10, 11** | | | | | | | | | | | | |
| PostNet | **74.5** | **39.8** | **36.1** | **36.0** | **45.9** | **46.0** | **74.5** | **43.3** | **42.0** | **32.1** | **35.1** | **82.6** |
| PriorNet | 32.3 | 26.6 | 26.5 | 26.5 | 26.6 | 28.3 | 32.3 | 26.7 | 26.6 | 26.6 | 27.0 | 30.4 |
| DDNet | 31.7 | 26.8 | 26.6 | 26.5 | 26.6 | 27.1 | 31.7 | 27.1 | 26.7 | 26.7 | 26.8 | 26.9 |
| EvNet | 66.5 | 30.5 | 28.2 | 27.1 | 28.1 | 31.8 | 66.5 | 38.7 | 36.1 | 30.2 | 28.2 | 28.8 |

OOD samples. OOD detection performance of all DBU models rapidly decreases with the size of the perturbation, regardless of whether attacks are computed on ID or OOD data (see Table 4). This performance decrease is also observed with AUROC as metric, attacks based on FGSM, Noise, when we use mutual information or precision $\alpha_0$ to distinguish between ID samples and OOD samples (see appendix Table 21 - 28). Thus, using uncertainty estimation to distinguish between ID and OOD data is not robust.

### 4.3. How to make DBU models more robust?

Our robustness analysis based on label attacks and uncertainty attacks shows that predictions, uncertainty estimation and the differentiation between ID and OOD data are not robust. Next, we explore approaches to improve robustness properties of DBU models w.r.t. these tasks based on randomized smoothing and adversarial training.

**Randomized smoothing** was originally proposed for certification of classifiers (Cohen et al., 2019). The core idea is to draw multiple samples $\boldsymbol{x}_s^{(i)} \sim \mathcal{N}(\boldsymbol{x}^{(i)}, \sigma)$ around the input data $\boldsymbol{x}^{(i)}$, to feed all these samples through the neural network, and to aggregate the resulting set of predictions (e.g. by taking their mean), to get a smoothed prediction. Besides allowing certification, as a side effect, the smoothed model is more robust. Our idea is to use randomized smoothing to improve robustness of DBU models, particularly w.r.t. uncertainty estimation. In contrast to discrete class predictions, however, certifying uncertainty estimates such as differential entropy scores requires a smoothing approach that is able to handle continuous values as in regression tasks. So far, only few works for randomized smoothing for regression models have been proposed (Kumar et al., 2020; yeh Chiang et al., 2020). We choose median smoothing (yeh Chiang et al., 2020), because it is applicable to unbounded domains as required for the uncertainty estimates covered in this work. In simple words: The set of uncertainty scores obtained from the $\boldsymbol{x}_s^{(i)} \sim \mathcal{N}(\boldsymbol{x}^{(i)}, \sigma)$ is aggregated by taking their median.

In the following experiments we focus on differential entropy as the uncertainty score. We denote the resulting smoothed differential entropy, i.e. the median output, as $m(\boldsymbol{x}^{(i)})$. Intuitively, we expect that the random sampling around a data point as well as the outlier-insensitivity of the median to improve the robustness of the uncertainty estimates w.r.t. adversarial examples.

To measure the performance and robustness of our smoothed DBU models, we apply median smoothing on the same tasks as in the previous sections, i.e., distinguishing between correctly and wrongly labeled inputs, attack detection, OOD detection and compute the corresponding AUC-PR score under label attacks and uncertainty attacks. The bold, middle part of the columns in Tables 5, 6, and 7 show the AUC-PR scores on CIFAR10, which we call *empirical performance* of the smoothed models. To facilitate the comparison with the base model of Section 4, we highlight the AUC-PR scores in blue in cases where the smooth model is more robust. The highlighting clearly shows that randomized smoothing increases the robustness of the empirical performance on OOD detection. OOD detection under strong PGD attacks (attack radius $\geq 0.5$) performs comparable to random guessing (i.e. AUC-PR scores around 50% whith 50% ID and 50% OOD data). This shows that DBU models are not reliably efficient w.r.t. this task. In attack detection and distinguishing between correctly and wrongly predicted

*Table 5.* Distinguishing between correctly and wrongly predicted labels based on differential entropy under PGD label attacks. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

| | **Att. Rad.** | 0.0 | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 |
|---|---|---|---|---|---|---|---|
| Smoothed models | **PostNet** | 80.5 · **91.5** · 94.5 | 52.8 · **71.6** · 95.2 | 31.9 · **51.0** · 96.8 | 5.6 · **11.7** · 100.0 | 0.3 · **0.6** · 100.0 | 0.0 · **0.0** · 100.0 |
| | **PriorNet** | 81.9 · **86.8** · 88.0 | 69.6 · **78.0** · 90.1 | 50.9 · **65.8** · 89.4 | 36.5 · **59.9** · 97.0 | 24.3 · **39.3** · 100.0 | 9.2 · **17.9** · 100.0 |
| | **DDNet** | 65.9 · **81.2** · 83.0 | 55.8 · **70.5** · 87.2 | 37.8 · **56.8** · 88.1 | 10.1 · **21.9** · 94.3 | 0.9 · **1.6** · 99.6 | 0.0 · **0.0** · 100.0 |
| | **EvNet** | 76.3 · **90.2** · 91.7 | 54.7 · **74.3** · 95.7 | 31.6 · **51.5** · 94.5 | 5.8 · **11.9** · 86.9 | 1.9 · **7.0** · 100.0 | 1.1 · **4.0** · 100.0 |
| Smoothed + adv. w. label attacks | **PostNet** | - | 52.1 · **71.8** · 95.6 | 31.2 · **47.9** · 96.1 | 7.8 · **14.7** · 98.6 | 1.8 · **4.4** · 100.0 | 0.3 · **0.5** · 100.0 |
| | **PriorNet** | - | 57.6 · **71.7** · 88.9 | 46.1 · **64.5** · 90.1 | 38.1 · **59.3** · 99.5 | 32.3 · **51.7** · 100.0 | 22.1 · **41.6** · 97.4 |
| | **DDNet** | - | 58.6 · **78.4** · 92.2 | 49.4 · **66.0** · 90.5 | 12.0 · **21.4** · 98.1 | 0.8 · **1.0** · 96.6 | 0.0 · **0.0** · 100.0 |
| | **EvNet** | - | 24.3 · **34.2** · 51.8 | 32.6 · **49.5** · 95.5 | 5.9 · **13.0** · 100.0 | 2.6 · **5.2** · 99.9 | 2.9 · **5.9** · 100.0 |
| Smoothed + adv. w. uncert. attacks | **PostNet** | - | 52.8 · **74.2** · 94.6 | 33.0 · **49.4** · 87.5 | 7.7 · **14.2** · 99.0 | 0.6 · **1.2** · 100.0 | 0.7 · **1.1** · 100.0 |
| | **PriorNet** | - | 50.6 · **68.1** · 88.6 | 44.4 · **66.1** · 96.0 | 35.1 · **57.4** · 98.4 | 18.4 · **32.2** · 100.0 | 15.2 · **29.3** · 100.0 |
| | **DDNet** | - | 68.8 · **84.4** · 93.2 | 45.1 · **60.8** · 86.8 | 12.3 · **22.0** · 91.0 | 0.8 · **1.7** · 87.0 | 0.0 · **0.0** · 100.0 |
| | **EvNet** | - | 54.2 · **73.7** · 96.1 | 30.5 · **50.0** · 99.5 | 7.1 · **13.9** · 100.0 | 3.7 · **8.7** · 75.2 | 3.3 · **5.8** · 100.0 |

*Table 6.* Attack detection (PGD label attacks) based on differential entropy. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

| | **Att. Rad.** | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 |
|---|---|---|---|---|---|---|
| Smoothed models | **PostNet** | 33.1 · **50.4** · 89.9 | 31.0 · **50.2** · 96.9 | 30.7 · **50.2** · 100.0 | 30.7 · **50.0** · 100.0 | 30.7 · **50.2** · 100.0 |
| | **PriorNet** | 35.9 · **50.6** · 74.5 | 33.0 · **50.3** · 82.8 | 31.2 · **50.0** · 95.7 | 30.7 · **50.4** · 99.9 | 30.7 · **50.4** · 100.0 |
| | **DDNet** | 36.3 · **50.3** · 76.4 | 32.8 · **49.9** · 84.6 | 30.8 · **50.1** · 98.0 | 30.7 · **50.2** · 100.0 | 30.7 · **50.2** · 100.0 |
| | **EvNet** | 32.9 · **50.4** · 89.8 | 31.4 · **50.1** · 94.0 | 30.8 · **50.0** · 98.0 | 30.7 · **50.3** · 100.0 | 30.7 · **49.6** · 100.0 |
| Smoothed + adv. w. label attacks | **PostNet** | 32.7 · **50.1** · 90.4 | 31.1 · **50.2** · 96.5 | 30.7 · **50.2** · 99.7 | 30.7 · **50.3** · 100.0 | 30.7 · **50.2** · 100.0 |
| | **PriorNet** | 35.2 · **51.8** · 78.6 | 32.8 · **51.1** · 84.4 | 30.8 · **50.2** · 98.7 | 30.7 · **50.5** · 100.0 | 30.8 · **50.1** · 98.2 |
| | **DDNet** | 35.5 · **50.6** · 79.2 | 33.4 · **50.3** · 84.1 | 30.8 · **50.1** · 99.2 | 30.7 · **50.0** · 100.0 | 30.7 · **50.5** · 100.0 |
| | **EvNet** | 40.3 · **50.4** · 66.8 | 31.4 · **50.3** · 95.8 | 30.7 · **50.3** · 100.0 | 30.7 · **50.1** · 100.0 | 30.7 · **50.0** · 100.0 |
| Smoothed + adv. w. uncert. attacks | **PostNet** | 33.3 · **50.6** · 88.7 | 32.5 · **50.1** · 87.9 | 30.7 · **49.9** · 99.8 | 30.7 · **50.1** · 100.0 | 30.7 · **50.0** · 100.0 |
| | **PriorNet** | 34.5 · **51.0** · 80.1 | 31.4 · **50.6** · 92.8 | 30.9 · **50.0** · 97.7 | 30.7 · **50.1** · 100.0 | 30.7 · **50.0** · 100.0 |
| | **DDNet** | 37.4 · **50.8** · 74.5 | 33.4 · **50.2** · 83.0 | 30.9 · **50.1** · 96.8 | 30.8 · **49.9** · 98.1 | 30.7 · **49.9** · 100.0 |
| | **EvNet** | 32.8 · **50.1** · 92.0 | 30.8 · **50.0** · 99.6 | 30.7 · **50.1** · 100.0 | 31.2 · **50.2** · 96.1 | 31.0 · **50.0** · 100.0 |

labels the smoothed DBU model are mostly more robust than the base models for attack radii $\geq 0.5$.

**Certified performance.** Using the median based on smoothing improves the empirical robustness, but it does not provide formal guarantees how low/high the performance might actually get under perturbed data (since any attack is only a heuristic). Here, we propose novel guarantees by exploiting the individual certificates we obtain via randomized smoothing. Note that the certification procedure (yeh Chiang et al., 2020) enables us to derive lower and upper bounds $\underline{m}(\boldsymbol{x}^{(i)}) \leq m(\boldsymbol{x}^{(i)}) \leq \overline{m}(\boldsymbol{x}^{(i)})$ which hold with high probability and indicate how much the median might change in the worst-case when $\boldsymbol{x}^{(i)}$ gets perturbed subject to a specific (attack) radius.

These bounds allow us to compute certificates that bound the performance of the smooth models, which we refer to as the *guaranteed lowest performance* and *guaranteed highest performance*. More precisely, for the guaranteed lowest performance of the model we take the pessimistic view that all ID data points realize their individual upper bounds $\overline{m}(\boldsymbol{x}^{(i)})$, i.e. have their highest possible uncertainty (worst case). On the other hand, we assume all OOD samples realize their lower bounds $\underline{m}(\boldsymbol{x}_s^{(i)})$. Using these values as the uncertainty scores for all data points we obtain the guaranteed lowest performance of the model. A guaranteed lowest performance of e.g. 35.0 means that even under the worst case conditions an attack is not able to decrease the performance below 35.0. Analogously, we can take the optimistic view to obtain the guaranteed highest performance of the smoothed models. Tables 5, 6 and 7 show the guaranteed lowest/highest performance (non-bold, left/right of the empirical performance). Our results show that the difference between guaranteed highest and guaranteed lowest performance increases with the attack radius, which might be explained by the underlying lower/upper bounds on the median being tighter for smaller perturbations.

*Table 7.* OOD detection based on differential entropy under PGD uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

| | **Att. Rad.** | 0.0 | 0.1 | 0.2 | 0.5 | 1.0 | 2.0 |
|---|---|---|---|---|---|---|---|
| | | | | **ID-Attack** | | | |
| Smoothed models | **PostNet** | 72.1 · **82.7** · 88.0 | 35.0 · **56.6** · 97.4 | 31.9 · **65.6** · 99.8 | 30.7 · **50.6** · 100.0 | 30.7 · **46.9** · 100.0 | 30.7 · **51.6** · 100.0 |
| | **PriorNet** | 50.2 · **53.1** · 55.9 | 33.5 · **43.3** · 65.3 | 31.3 · **39.7** · 69.1 | 31.3 · **48.3** · 98.2 | 30.7 · **44.4** · 99.9 | 30.7 · **45.4** · 100.0 |
| | **DDNet** | 72.0 · **75.8** · 79.8 | 35.6 · **46.2** · 69.8 | 32.9 · **50.3** · 87.1 | 31.1 · **58.7** · 98.6 | 30.7 · **59.3** · 100.0 | 30.7 · **44.5** · 100.0 |
| | **EvNet** | 79.5 · **87.1** · 92.8 | 34.1 · **58.6** · 95.1 | 32.5 · **61.2** · 96.9 | 31.7 · **60.6** · 98.7 | 30.7 · **62.4** · 100.0 | 30.7 · **57.3** · 100.0 |
| Smoothed + adv. w. label attacks | **PostNet** | - | 35.0 · **58.5** · 97.7 | 31.2 · **46.6** · 97.4 | 30.8 · **57.7** · 99.7 | 30.7 · **49.8** · 100.0 | 30.7 · **50.9** · 100.0 |
| | **PriorNet** | - | 31.5 · **36.7** · 57.2 | 33.1 · **51.8** · 84.8 | 30.7 · **57.7** · 98.7 | 30.7 · **40.0** · 99.9 | 30.9 · **53.6** · 96.7 |
| | **DDNet** | - | 36.2 · **50.0** · 78.6 | 32.1 · **41.3** · 70.2 | 30.8 · **56.4** · 100.0 | 30.7 · **49.4** · 100.0 | 30.7 · **54.8** · 100.0 |
| | **EvNet** | - | 46.8 · **61.0** · 79.7 | 32.3 · **58.9** · 99.1 | 30.7 · **45.0** · 100.0 | 30.7 · **63.3** · 100.0 | 30.8 · **38.1** · 100.0 |
| Smoothed + adv. w. uncert. attacks | **PostNet** | - | 35.2 · **55.9** · 96.0 | 34.5 · **59.2** · 94.9 | 30.7 · **47.0** · 100.0 | 30.7 · **58.2** · 100.0 | 30.7 · **42.9** · 100.0 |
| | **PriorNet** | - | 31.8 · **38.9** · 64.1 | 31.0 · **41.8** · 87.9 | 30.7 · **42.9** · 99.2 | 30.7 · **48.6** · 100.0 | 30.7 · **46.6** · 100.0 |
| | **DDNet** | - | 39.7 · **52.1** · 75.7 | 36.4 · **56.8** · 83.8 | 31.0 · **51.5** · 97.4 | 31.0 · **56.8** · 97.8 | 30.7 · **49.1** · 100.0 |
| | **EvNet** | - | 34.8 · **64.9** · 99.6 | 30.8 · **48.9** · 99.8 | 30.7 · **66.8** · 100.0 | 30.9 · **41.5** · 93.8 | 31.1 · **55.1** · 100.0 |
| | | | | **OOD-Attack** | | | |
| Smoothed models | **PostNet** | 72.0 · **82.7** · 88.0 | 35.1 · **56.8** · 97.3 | 32.0 · **65.8** · 99.8 | 30.7 · **50.7** · 100.0 | 30.7 · **46.5** · 100.0 | 30.7 · **51.7** · 100.0 |
| | **PriorNet** | 50.3 · **53.1** · 55.9 | 33.6 · **43.7** · 65.9 | 31.3 · **39.8** · 69.4 | 31.3 · **48.3** · 98.2 | 30.7 · **44.5** · 99.9 | 30.7 · **46.4** · 100.0 |
| | **DDNet** | 72.0 · **75.8** · 79.8 | 35.6 · **46.2** · 70.0 | 32.9 · **50.1** · 86.7 | 31.1 · **58.8** · 98.6 | 30.7 · **59.3** · 100.0 | 30.7 · **44.6** · 100.0 |
| | **EvNet** | 79.5 · **87.1** · 92.8 | 34.1 · **58.8** · 95.2 | 32.6 · **61.2** · 96.9 | 31.7 · **60.5** · 98.7 | 30.7 · **62.4** · 100.0 | 30.7 · **57.6** · 100.0 |
| Smoothed + adv. w. label attacks | **PostNet** | - | 35.0 · **58.5** · 97.8 | 31.2 · **46.6** · 97.2 | 30.8 · **57.7** · 99.7 | 30.7 · **50.2** · 100.0 | 30.7 · **51.5** · 100.0 |
| | **PriorNet** | - | 31.6 · **37.3** · 59.3 | 33.2 · **52.7** · 85.8 | 30.7 · **57.8** · 98.7 | 30.7 · **40.1** · 99.9 | 30.9 · **53.8** · 96.8 |
| | **DDNet** | - | 36.4 · **50.2** · 78.9 | 32.1 · **41.5** · 70.4 | 30.9 · **56.2** · 100.0 | 30.7 · **49.3** · 100.0 | 30.7 · **55.1** · 100.0 |
| | **EvNet** | - | 47.2 · **61.1** · 80.0 | 32.4 · **59.1** · 99.1 | 30.7 · **45.0** · 100.0 | 30.7 · **63.2** · 100.0 | 30.8 · **38.0** · 100.0 |
| Smoothed + adv. w. uncert. attacks | **PostNet** | - | 35.3 · **56.4** · 96.1 | 34.5 · **59.0** · 94.9 | 30.7 · **46.8** · 100.0 | 30.7 · **57.8** · 100.0 | 30.7 · **43.2** · 100.0 |
| | **PriorNet** | - | 31.9 · **39.4** · 65.5 | 31.0 · **42.0** · 88.6 | 30.7 · **42.9** · 99.2 | 30.7 · **48.4** · 100.0 | 30.7 · **47.1** · 100.0 |
| | **DDNet** | - | 40.2 · **52.9** · 76.5 | 36.4 · **56.9** · 83.9 | 31.1 · **51.5** · 97.3 | 31.0 · **57.0** · 97.8 | 30.7 · **49.1** · 100.0 |
| | **EvNet** | - | 34.9 · **64.8** · 99.6 | 30.8 · **48.8** · 99.8 | 30.7 · **66.1** · 100.0 | 30.9 · **41.6** · 93.6 | 31.1 · **54.7** · 100.0 |

**Adversarial training.** Randomized smoothing improves robustness of DBU models and allows us to compute performance guarantees. However, an open question is whether it is possible to increase robustness even further by combining it with adversarial training. To obtain adversarially trained models we augment the data set using perturbed samples that are computed by PGD attacks against the cross-entropy loss (label attacks) or the differential entropy (uncertainty attacks). These perturbed samples $\tilde{\boldsymbol{x}}^{(i)}$ are computed during each epoch of the training based on inputs $\boldsymbol{x}^{(i)}$ and added to the training data (with the label $y^{(i)}$ of the original input). Tables 5, 6, and 7 illustrate the results. We choose the attack radius used during training and the $\sigma$ used for smoothing to be equal. To facilitate comparison, we highlight the empirical performance of the adversarially trained models in blue if it is better than the performance of the base model. Our results show that the additional use of adversarial training has a minor effect on the robustness and does not result in a significant further increase of the robustness.

We conclude that median smoothing is a promising technique to increase robustness w.r.t. distinguishing between correctly labeled samples and wrongly labeled samples, attack detection and differentiation between in-distribution data and out-of-distribution data of all Dirichlet-based uncertainty models, while additional adversarial training has a minor positive effect on robustness.

## 5. Conclusion

This work analyzes robustness of uncertainty estimation by DBU models and answers multiple questions in this context. Our results show: (1) While uncertainty estimates are a good indicator to identify correctly classified samples on unperturbed data, performance decrease drastically on perturbed data-points. (2) None of the Dirichlet-based uncertainty models is able to detect PGD label attacks against the class prediction by uncertainty estimation, regardless of the used uncertainty measure. (3) Detecting OOD samples and distinguishing between ID-data and OOD-data is not robust. (4) Applying median smoothing to uncertainty estimates increases robustness of DBU models w.r.t. all analyzed tasks, while adversarial training based on label or uncertainty attacks resulted in minor improvements.

## Acknowledgments

This research was supported by BMW AG.

## References

Bitterwolf, J., Meinke, A., and Hein, M. Provable worst case guarantees for the detection of out-of-distribution data. *Neural Information Processing Systems*, 2020.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. *International Conference on Machine Learning*, 2015.

Bojchevski, A. and Günnemann, S. Certifiable robustness to graph perturbations. *Neural Information Processing Systems*, 2019.

Bojchevski, A., Klicpera, J., and Günnemann, S. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. *International Conference on Machine Learning*, 2020.

Brendel, W., Rauber, J., and Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *International Conference on Learning Representations*, 2018.

Carbone, G., Wicker, M., Laurenti, L., Patane, A., Bortolussi, L., and Sanguinetti, G. Robustness of bayesian neural networks to gradient-based attacks. *Neural Information Processing Systems*, 2020.

Cardelli, L., Kwiatkowska, M., Laurenti, L., Paoletti, N., Patane, A., and Wicker, M. Statistical guarantees for the robustness of bayesian neural networks. *Conference on Artificial Intelligence*, 2019.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. *ACM Workshop on Artificial Intelligence and Security*, 2017.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. *Symposium on Security and Privacy*, 2017.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Neural Information Processing Systems*, 2020.

Cheng, M., Yi, J., Chen, P.-Y., Zhang, H., and Hsieh, C.-J. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *AAAI Conference on Artificial Intelligence*, 2020.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. *International Conference on Machine Learning*, 2017.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *Neural Information Processing Systems, Machine Learning for Creativity and Design Workshop*, 2018.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning*, 2019.

Dang-Nhu, R., Singh, G., Bielik, P., and Vechev, M. Adversarial attacks on probabilistic autoregressive forecasting models. *International Conference on Machine Learning*, 2020.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. *International Conference on Machine Learning*, 2006.

Dua, D. and Graff, C. UCI machine learning repository. *University of California*, 2017.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 2016.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. *International Conference on Learning Representations*, 2015.

Kopetzki, A.-K. and Günnemann, S. Reachable sets of classifiers and regression models: (non-)robustness analysis and robust training. *Machine Learning Journal*, 2021.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10. *Canadian Institute for Advanced Research*, 2009.

Kumar, A., Levine, A., Feizi, S., and Goldstein, T. Certifying confidence via randomized smoothing. *Neural Information Processing Systems*, 2020.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Neural Information Processing Systems*, 2017.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. *National Institute of Standards and Technology*, 2010.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Neural Information Processing Systems*, 2018.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Neural Information Processing Systems*, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. *Neural Information Processing Systems*, 2018a.

Malinin, A. and Gales, M. Prior networks for detection of adversarial attacks. *arXiv:1812.02575 [stat.ML]*, 2018b.

Malinin, A. and Gales, M. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Neural Information Processing Systems*, 2019.

Malinin, A., Mlodozeniec, B., and Gales, M. Ensemble distribution distillation. *International Conference on Learning Representations*, 2019.

Meinke, A. and Hein, M. Towards neural networks that provably know when they don't know. *International Conference on Learning Representations*, 2020.

Nandy, J., Hsu, W., and Lee, M. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *International Conference on Machine Learning, Uncertainty and Robustness in Deep Learning Workshop*, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. *Neural Information Processing Systems, Deep Learning and Unsupervised Feature Learning Workshop*, 2011.

Osawa, K., Swaroop, S., Khan, M. E. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. Practical deep learning with bayesian principles. *Neural Information Processing Systems*, 2019.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Neural Information Processing Systems*, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. *Neural Information Processing Systems*, 2019.

Qin, Y., Wang, X., Beutel, A., and Chi, E. H. Improving uncertainty estimates through the relationship with adversarial robustness. *arXiv:2006.16375 [cs.LG]*, 2020.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 2015.

Schuchardt, J., Bojchevski, A., Klicpera, J., and Günnemann, S. Collective robustness certificates: Exploiting interdependence in graph neural networks. *International Conference on Learning Representations*, 2021.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Neural Information Processing Systems*, 2018.

Sensoy, M., Kaplan, L., Cerutti, F., and Saleki, M. Uncertainty-aware deep classifiers using generative models. *AAAI Conference on Artificial Intelligence*, 2020.

Shi, W., Zhao, X., Chen, F., and Yu, Q. Multifaceted uncertainty estimation for label-efficient deep learning. *Neural Information Processing Systems*, 2020.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

Singh, G., Ganvir, R., Püschel, M., and Vechev, M. Beyond the single neuron convex barrier for neural network certification. *Neural Information Processing Systems*, 2019.

Smith, L. and Gal, Y. Understanding measures of uncertainty for adversarial example detection. *Uncertainty in Artificial Intelligence*, 2018.

Stutz, D., Hein, M., and Schiele, B. Confidence-calibrated adversarial training: Generalizing to unseen attacks. *International Conference on Machine Learning*, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

Tagasovska, N. and Lopez-Paz, D. Single-model uncertainties for deep learning. *Neural Information Processing Systems*, 2019.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. *International Conference on Learning Representations*, 2018.

Wicker, M., Laurenti, L., Patane, A., and Kwiatkowska, M. Probabilistic safety for bayesian neural networks. *Uncertainty in Artificial Intelligence*, 2020.

Wong, E. and Kolter, J. Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. *International Conference on Machine Learning*, 2018.

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

yeh Chiang, P., Curry, M. J., Abdelkader, A., Kumar, A., Dickerson, J., and Goldstein, T. Detection as regression: Certified object detection by median smoothing. *Neural Information Processing Systems*, 2020.

Zhao, X., Chen, F., Hu, S., and Cho, J.-H. Uncertainty aware semi-supervised learning on graph data. *Neural Information Processing Systems*, 2020.

Zheng, S., Song, Y., Leung, T., and Goodfellow, I. Improving the robustness of deep neural networks via stability training. *Conference on Computer Vision and Pattern Recognition*, 2016.

Zügner, D., Akbarnejad, A., and Günnemann, S. Adversarial attacks on neural networks for graph data. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2018.