BEYOND FIRST-ORDER UNCERTAINTY ESTIMATION WITH EVIDENTIAL MODELS FOR OPEN-WORLD RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

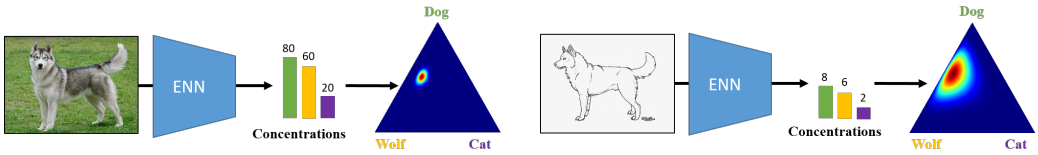
Reliable uncertainty estimation is crucial when deploying a classifier in the wild. In this paper, we tackle the challenge of jointly quantifying in-distribution and out-of-distribution (OOD) uncertainties. To this end, we leverage the second-order uncertainty representation provided by evidential models and we introduce *KLoS*, a KL-divergence criterion defined on the class-probability simplex. By keeping the full distributional information, KLoS captures more than existing first-order uncertainty measures such as the predictive entropy. In addition, KLoS acts as a class-wise density estimator and is effective to detect various types of OOD samples, in contrast with current second-order measures. We further design an auxiliary neural network, *KLoSNet*, to learn a refined criterion directly aligned with the evidential training objective. In the realistic context where no OOD data is available during training, our experiments show that KLoSNet outperforms first and second-order uncertainty metrics to simultaneously detect misclassifications and OOD samples. We also report comparisons in the presence of OOD training samples, which shed a new light on the impact of the vicinity of this data with OOD test data.

1 INTRODUCTION

Safety is a major concern in visual-recognition applications such as autonomous driving (McAllister et al., 2017) and medical imaging (Heckerman et al., 1992). However, modern neural networks (NNs) struggle to detect their own misclassifications (Hendrycks & Gimpel, 2017). In addition, when exposed to out-of-distribution (OOD) samples, NNs have been shown to provide overconfident predictions instead of abstaining (Hein et al., 2019; Nguyen et al., 2015). Obtaining reliable estimates of the predictive uncertainty is thus necessary to safely deploy models in open-world conditions (Bendale & Boult, 2015).

Notable progress has been made in NN uncertainty estimation with the renewal of Bayesian neural networks (Gal & Ghahramani, 2016; Maddox et al., 2019) and ensembling (Lakshminarayanan et al., 2017; Ovadia et al., 2019b). These techniques produce a probability density over the predictive categorical distribution obtained from sampling. A recent class of models, coined *evidential* (Sensoy et al., 2018; Joo et al., 2020; Malinin & Gales, 2019), proposes instead to explicitly learn the concentration parameters of a Dirichlet distribution over output probabilities. They have been shown to improve generalisation (Joo et al., 2020), OOD detection (Malinin & Gales, 2019; Nandy et al., 2020) and adversarial attack detection (Malinin & Gales, 2019).

Based on the subjective logic framework (Josang, 2016), evidential models capture different sources of uncertainty. First-order uncertainty relates to the expectation of the Dirichlet distribution and is caused by conflicting evidence, e.g., class confusion. Second-order uncertainty expresses the lack of evidence in a prediction (Shi et al., 2020), which is characterized by the spread of the Dirichlet distribution. First and second-order uncertainties are illustrated in Fig. 1 for a three-class classifier of *cat*, *dog* and *wolf* trained on natural images. It is first fed with the photo of a husky (Fig. 1a). Although being a breed of dog, huskies share lots of features with wolves, which leads to a large first-order uncertainty due to class confusion. In presence of a drawing of a husky (Fig. 1b), a similar



(a) Natural image of a husky
MCP $= 0.50$, entropy $= 0.97$, KLoS $= 97.85$

(b) Drawing of a husky
MCP $= 0.50$, entropy $= 0.97$, KLoS $= 104.71$

Figure 1: **Limitations of 1st-order uncertainty measures and their handling with KLoS**. (a) An in-distribution image with conflicting evidence (1st-order uncertainty) between *dog* and *wolf*. (b) An out-of-distribution (OOD) image with same class confusion is supposed to have larger 2nd-order uncertainty. An evidential neural network (ENN) outputs class-wise evidence information as concentration parameters of a Dirichlet density (visualized on the simplex) over 3-class distributions. Although this density is flatter for the second input, the predictive entropy and MCP, only based on 1st-order statistics, are equal for both inputs. In contrast, the proposed measure, KLoS, captures both 1st- and 2nd-order uncertainties, hence correctly reflecting the larger uncertainty for the OOD sample.

class confusion is expected, as well as a lower amount of evidence (second-order uncertainty) due to the distribution shift.

For open-world recognition, a model should be equipped with an uncertainty measure that accounts both for first-order and second-order uncertainties to detect misclassifications and OOD samples.

Surprisingly, previous works do not leverage the distribution over probabilities on the simplex to derive such a joint measure of the two sources of uncertainty. Some methods focus on OOD detection by characterizing only the distribution spread, e.g., using mutual information (Malinin & Gales, 2019; Nandy et al., 2020). Approaches targeting total uncertainty actually reduce probability distributions on the simplex to their expected value and compute first-order uncertainty measures (Joo et al., 2020; Sensoy et al., 2018), e.g., the predictive entropy or the *maximum class probability* (MCP). This causes a significant loss of information, as shown in Fig. 1: The resulting measures are invariant to the spread of the distribution, whereas uncertainty caused by class confusion and lack of evidence should be cumulative, a property naturally fulfilled by the predictive variance in Bayesian regression (Murphy, 2012).

In addition, some methods for evidential models (Malinin & Gales, 2018; 2019; Nandy et al., 2020) rely on using auxiliary OOD data during training to enforce higher distribution spread on OOD inputs. In contrast, when deprived access to OOD training data, the low-dispersion behavior is not guaranteed for all OOD examples (Charpentier et al., 2020; Sensoy et al., 2020) and second-order uncertainty measures struggle to discriminate them from in-distribution examples.

Contributions. In this paper, we propose a measure that accounts for both in-distribution and out-of-distribution sources of uncertainty and that does no require the access to auxiliary OOD data at train time. Our contributions are threefold: (1) Leveraging the second-order uncertainty representation that evidential models provide, we introduce *KLoS*, a KL-divergence measure on the simplex based on the Dirichlet distribution; KLoS provides richer estimates than standard first-order measures by considering both 1st-order and 2nd-order uncertainties. (2) Noting that KLoS naturally reflects the training objective used in evidential models, we propose to learn an auxiliary model, named *KLoSNet*, to regress the values of this objective for training samples and to improve uncertainty estimation. (3) To assess the quality of uncertainty estimates in open-world recognition, we design the new task of simultaneous detection of misclassifications and OOD samples. Extensive experiments show the benefits of KLoSNet on various image datasets and model architectures. In particular, we highlight the class-wise density estimator behaviour of KLoSNet, a crucial property in the absence of OOD training data. We also shed a new light on the impact of the type of OOD training samples. Finally, we show that KLoS can be successfully combined with ensembling to improve performance.

2 CAPTURING FIRST-ORDER AND SECOND-ORDER UNCERTAINTIES

Section 2.2 presents our measure to capture first-order and second-order uncertainties with evidential models. We propose a confidence learning approach to enhance first-order uncertainty estimation in Section 2.3. Beforehand, we review evidential models and their learning in Section 2.1 to put in perspective the benefits of the proposed approach.

2.1 BACKGROUND: EVIDENTIAL NEURAL NETWORKS

Preliminaries. Let us first consider the problem of inferring from a set $\mathcal{D}$ of N rolls the probability that a dice with C sides comes up as face c (Murphy, 2012). We denote $\boldsymbol{\pi} = (\pi_1 \cdots \pi_C)$ the random variable over categorical probabilities, where $\sum_{c=1}^{C} \pi_c$ = 1, and which lives on the (C-1)-dimensional simplex $\triangle^{C-1}$. Assuming *i.i.d.* data, its likelihood reads $p(\mathcal{D}|\boldsymbol{\pi}) = \prod_{c=1}^{C} \pi_c^{N_c}$ where N_c is the count of class c among the N draws. For its conjugate properties with the categorical distribution, the prior $p(\boldsymbol{\pi})$ can be modeled as a Dirichlet distribution with concentration parameters $\boldsymbol{\beta}$. Then, the posterior $p(\boldsymbol{\pi}|\mathcal{D})$ is also a Dirichlet distribution with parameters (β_1+N_1,..., β_C+N_C) and the posterior predictive distribution for a single multinoulli trial has the closed form $P(c|\mathcal{D}) = \mathbb{E}\big[\pi_c|\mathcal{D}\big] = \frac{\beta_c+N_c}{\sum_c \beta_c+N}$.

Let us extend the Bayesian treatment of a single categorical distribution to classification, i.e., the goal is to predict the class label y from a categorical distribution that depends on input $\boldsymbol{x}$. The training dataset $\mathcal{D}$ consists of N *i.i.d.* samples $(\boldsymbol{x}, y)$ drawn from an unknown joint distribution $p(\mathbf{x}, \mathbf{y})$. Obviously, for a test sample $\boldsymbol{x}^*$, its label frequency count is now unknown and we are not able to estimate the posterior predictive distribution $P(\mathbf{y} = c|\boldsymbol{x}^*, \mathcal{D})$. Bayesian models and ensembling approximate the posterior predictive distribution by marginalizing over the network's parameters thanks to sampling. But this comes at the cost of multiple forward passes. Evidential Neural Networks (ENN) propose instead to model explicitly the posterior distribution over categorical probabilities $p(\boldsymbol{\pi}|\boldsymbol{x}, y)$ by a variational Dirichlet distribution,

$$q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x}) = \text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{\theta})\big) = \frac{\Gamma(\alpha_0(\boldsymbol{x}, \boldsymbol{\theta}))}{\prod_{c=1}^{C} \Gamma(\alpha_c(\boldsymbol{x}, \boldsymbol{\theta}))} \prod_{c=1}^{C} \pi_c^{\alpha_c(\boldsymbol{x}, \boldsymbol{\theta})-1}, \tag{1}$$

whose concentration parameters $\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{\theta}) = \exp f(\boldsymbol{x}, \boldsymbol{\theta})$ are output by a network f with parameters $\boldsymbol{\theta}$; Γ is the Gamma function and $\alpha_0(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{c=1}^{C} \alpha_c(\boldsymbol{x}, \boldsymbol{\theta})$. Precision α_0 controls the sharpness of the density with more mass concentrating around the mean as α_0 grows. By conjugate property, the predictive distribution for a new point $\boldsymbol{x}^*$ is $P(\mathbf{y} = c|\boldsymbol{x}^*, \mathcal{D}) \approx \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x}^*)}[\pi_c] = \frac{\exp f_c(\boldsymbol{x}^*, \boldsymbol{\theta})}{\sum_{k=1}^{C} \exp f_k(\boldsymbol{x}^*, \boldsymbol{\theta})}$, which is the usual output of a network f with softmax activation.

The concentration parameters $\boldsymbol{\alpha}$ can be interpreted as pseudo-counts representing the amount of evidence in each class. For instance, in Fig. 1a, the $\boldsymbol{\alpha}$'s output by the ENN indicates that the image is almost equally likely to be classified as *wolf* or as *dog*. More interestingly, it also distinguishes this in-distribution sample from the OOD one in Fig. 1b via the total amount of evidence α_0.

Training Objective. The ENN training (Joo et al., 2020) is formulated as a variational approximation to minimize the KL divergence between the distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x})$ and the true posterior distribution $p(\boldsymbol{\pi}|\boldsymbol{x}, y)$:

$$\mathcal{L}_{\text{var}}(\boldsymbol{\theta}; \mathcal{D}) = \mathbb{E}_{(\boldsymbol{x}, y)\sim p(\mathbf{x}, \mathbf{y})} \big[\text{KL}\big(q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x}) \parallel p(\boldsymbol{\pi}|\boldsymbol{x}, y)\big)\big] \tag{2}$$

$$\propto \frac{1}{N} \sum_{(\boldsymbol{x}, y)\in\mathcal{D}} \Big(\psi(\alpha_y(\boldsymbol{x}, \boldsymbol{\theta})) - \psi(\alpha_0(\boldsymbol{x}, \boldsymbol{\theta})) + \text{KL}(q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x}) \parallel p(\boldsymbol{\pi}|\boldsymbol{x}))\Big). \tag{3}$$

where ψ is the digamma function. Following Joo et al. (2020), we use the non-informative uniform prior $p(\boldsymbol{\pi}|\boldsymbol{x}) = \text{Dir}\big(\boldsymbol{\pi}|\mathbf{1}\big)$, where $\mathbf{1}$ is the all-one vector, and weigh the KL-divergence term with $\lambda > 0$[1]:

$$\mathcal{L}_{\text{var}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{N} \sum_{(\boldsymbol{x}, y)\in\mathcal{D}} \Big(\psi(\alpha_y) - \psi(\alpha_0) + \lambda \text{KL}\big(\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \parallel \text{Dir}(\boldsymbol{\pi}|\mathbf{1})\big)\Big). \tag{4}$$

[1]For conciseness, we denote $\alpha_c = \alpha_c(\boldsymbol{x}, \boldsymbol{\theta}), \forall c$ hereafter.

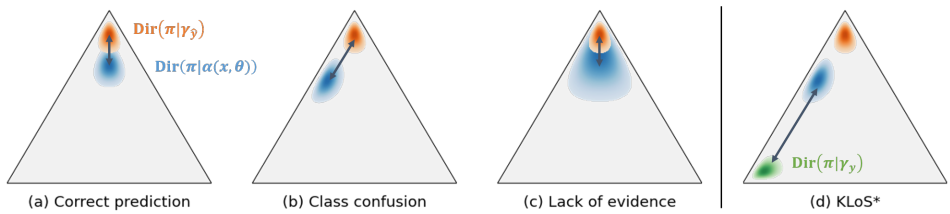


Figure 2: **KLoS and KLoS* on the probability simplex**. Given the input sample, the blue region represents the distribution predicted by the evidential model and the orange region represents the sharp Dirichlet distribution with parameters $\boldsymbol{\gamma}_{\hat{y}} = (1, \cdots, 1, \lambda^{-1}, 1, \ldots, 1)$ focused on the predicted class $\hat{y}$ (see Eq. (5)). (*Left*) Illustration of the behavior of KLoS in absence of uncertainty (a), in case of class confusion (b) and in case of lack of evidence (c); (*Right*) Different from KLoS, KLoS* measures the distance to a Dirichlet distribution centered on the true class y, represented in green.

In particular, minimizing loss Eq. (4) enforce training sample's precision α_0 to remain close to $C + 1/\lambda$ value (Malinin & Gales, 2019).

2.2 A KL-Divergence Measure on the Simplex

By explicitly learning a distribution of the categorical probabilities $\boldsymbol{\pi}$, evidential models can distinguish first-order from second-order uncertainty on the simplex. Inputs with large first-order uncertainty due to class confusion will have a distribution closer to the simplex center; conversely, inputs with large second-order uncertainty are expected to present flat distributions, reflecting the lack of evidence of the model on these points. To encompass both types of uncertainty, we need to design a measure that encapsulates the sharpness of the distribution and its location on the simplex.

While originally used to measure the total uncertainty (Malinin & Gales, 2019), the predictive entropy $\mathcal{H}[y|\boldsymbol{x}, \boldsymbol{\theta}]$ and the maximum class probability $\text{MCP}(\boldsymbol{x}, \boldsymbol{\theta}) = \max_c P(\mathbf{y} = c|\boldsymbol{x}, \boldsymbol{\theta})$ only account for the position on the simplex. These measures are invariant to the dispersion of the Dirichlet distribution that generates the categorical probabilities. This can be problematic, as illustrated in Fig. 1.

To capture first-order and second-order uncertainties, we introduce a novel measure, named *KLoS* for "KL on Simplex", that measures the KL divergence between the model's output and a sharp Dirichlet distribution with concentrations $\boldsymbol{\gamma}_{\hat{y}}$ focused on the *predicted* class $\hat{y}$:

$$\text{KLoS}(\boldsymbol{x}) \triangleq \text{KL}\Big(\text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{\theta})\big) \parallel \text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\gamma}_{\hat{y}}\big)\Big), \tag{5}$$

where $\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{\theta}) = \exp f(\boldsymbol{x}, \boldsymbol{\theta})$ are model's output and $\boldsymbol{\gamma}_{\hat{y}} = (1, \ldots, 1, \tau, 1, \ldots, 1)$ are the uniform concentration parameters except for the predicted class with $\tau = 1/\lambda + 1$.

The lower KLoS is, the more certain the prediction is. Correct predictions will have Dirichlet distributions close to the posterior distribution and will thus be associated with a low uncertainty measure (Fig. 2a). Samples with high class confusion will present concentration parameters closer to simplex's center than the target Dirichlet objective, resulting in a higher KLoS measure (Fig. 2b). KLoS also penalizes samples having a different precision α_0 than the precision $\alpha_0^* = \tau + C - 1$ of the target $\boldsymbol{\gamma}_{\hat{y}}$. For instance, samples with lacking evidence, i.e. having smaller precision than α_0^* and thus a wider spread on the simplex (Fig. 2c), receive a larger KLoS score. Since in-distribution samples are enforced to have precision close to α_0^* during training, KLoS will be effective to detect various types of OOD samples whose precision is far from α_0^*, and acts as a class-wise density estimator, a behavior also exhibited in Section 4.1. In contrast, second-order uncertainty measures e.g., mutual information, assume that OOD samples have smaller α_0, a property difficult to fulfill for models trained only with in-distribution samples (see Section 4.3).

Decomposition of KLoS. By approximating the digamma function ψ (see Supplementary E), KLoS can be decomposed as:

$$\text{KLoS}(\boldsymbol{x}) \approx (\tau - 1) \log \big(\frac{\alpha_{\hat{y}}}{\alpha_0}\big) + \Big((\tau - 1)\big(\frac{1}{2\alpha_0} - \frac{1}{2\alpha_{\hat{y}}}\big) + \text{KL}\big(\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \parallel \text{Dir}(\boldsymbol{\pi}|\mathbf{1})\big)\Big). \tag{6}$$

The first term is the standard log-likelihood and relates only to expected probabilities, hence to the first-order uncertainty. The ratio $\alpha_{\hat{y}}/\alpha_0$ makes it invariant to any scaling of the concentration parameters vector $\boldsymbol{\alpha}$. The other terms take into account the spread of the distribution by measuring how close α_0 is to $(\tau+C-1)$, and encode second-order uncertainty.

2.3 Improving First-Order Uncertainty Representation with Confidence Learning

When the model misclassifies an example, i.e., the predicted class $\hat{y}$ differs from the ground truth y, KLoS measures the distance between the ENN's output and the wrongly estimated posterior $p(\boldsymbol{\pi}|\boldsymbol{x}, \hat{y})$. This may result in an arbitrarily high confidence / low KL-divergence value. Measuring instead the distance to the true posterior distribution $p(\pi|\boldsymbol{x}, y)$ (green region in Fig. 2d) would more likely yield a greater value, reflecting the fact that the classifier made an error. Thus, a better measure for misclassification detection would be:

$$\text{KLoS}^*(\boldsymbol{x}, y) \triangleq \text{KL}\Big(\text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{\theta})\big) \parallel \text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\gamma}_y\big)\Big), \tag{7}$$

where $\boldsymbol{\gamma}_y$ corresponds to the uniform concentrations except for the *true* class y with $\tau = 1/\lambda + 1$.

Connecting KLoS* with Evidential Training Objective. Choosing such value for τ results in KLoS* matching the objective function in Eq. (4). This means that KLoS* is explicitly minimized by the evidential model during training for in-distribution samples. By mimicking the evidential training objective, we reflect the fact that the model is confident about its prediction if KLoS* is close to zero. In addition, minimizing the KL-divergence between the variational distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x})$ and the posterior $p(\boldsymbol{\pi}|\boldsymbol{x}, y)$ is equivalent to maximizing the evidence lower bound (ELBO). Hence, a small KLoS* value corresponds to a high ELBO, which is coherent with the common assumption in variational inference that higher ELBO corresponds to "better" models (Gal, 2016).

Obviously, the true class of an output is not available when estimating confidence on test samples. We propose to learn KLoS* by introducing an auxiliary confidence neural network, termed KLoS-Net, with parameters $\boldsymbol{\omega}$, which outputs a confidence prediction $C(\boldsymbol{x}, \boldsymbol{\omega})$. KLoSNet consists of a small decoder, composed of several dense layers attached to the penultimate layer of the original classification network. During training, we seek $\boldsymbol{\omega}$ such that $C(\boldsymbol{x}, \boldsymbol{\omega})$ is close to KLoS$^*(\boldsymbol{x}, y)$, by minimizing

$$\mathcal{L}_{\text{KLoSNet}}(\boldsymbol{\omega}; \mathcal{D}) = \frac{1}{N} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} \big\| C(\boldsymbol{x}, \boldsymbol{\omega}) - \text{KLoS}^*(\boldsymbol{x}, y) \big\|^2. \tag{8}$$

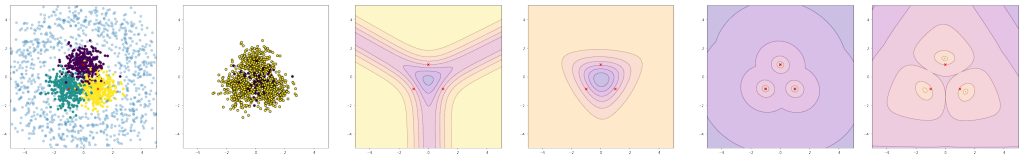
KLoSNet can be further improved by endowing it with its own feature extractor. Initialized with the encoder of the classification network, which must remain untouched for not affecting its performance, the encoder of KLoSNet can be finetuned along with its regression head. This amounts to minimizing Eq. (8) w.r.t. to both sets of parameters.

The training set for confidence learning is the one used for classification training. In the experiments, we observe a slight performance drop when using a validation set instead. Indeed, when dealing with models with high predictive performance and small validation sets, we end up with fewer misclassification examples than in the train set.

At test time, we now directly use KLoSNet's scalar output $C(\boldsymbol{x}, \boldsymbol{\omega}')$ as our uncertainty estimate. As previously, the lower the output value, the more confident the prediction.

3 Related Work

Misclassification Detection. Several works (Jiang et al., 2018; Corbière et al., 2019; Moon et al., 2020) aim to improve the standard MCP baseline (Hendrycks & Gimpel, 2017) in misclassification detection with NNs. In particular, Corbière et al. (2019) points out that the probability associated to the true class (TCP) offers a better uncertainty measure than MCP, provided it can be well estimated. The authors design an auxiliary NN to predict TCP on training points. They observe that using a held-out validation set improves the performance only when dealing with large-scale datasets or small models. We adapt their high-level idea for evidential models to tackle jointly the detection of misclassified in-distribution samples and of OOD samples.



(a) Toy dataset (b) Corr./Err. (c) Entropy (d) Mut. Inf. (e) Mahalanobis (f) KLoS

Figure 3: **Comparison of various uncertainty measures for a given evidential classifier on a toy dataset.** (a) Training samples from 3 input Gaussian distributions with large overlap (hence class confusion) and OOD test samples (blue); (b) Correct (yellow) and erroneous (red) class predictions on in-domain test samples; (c-f) Visualisation of different uncertainty measures derived from the evidential model trained on the toy dataset. Yellow (resp. purple) indicates high (resp. low) certainty.

Out-of-Distribution Detection. Bayesian neural networks (Gal & Ghahramani, 2016; Maddox et al., 2019; Neal, 1996) and ensembling (Lakshminarayanan et al., 2017; Rame & Cord, 2021) offer a principled approach for uncertainty estimation in which a second-order measure can be derived by measuring the dispersion between individual probabilities vectors. But this comes at the expense of an increased computational cost. Evidential models emulate an ensemble of models using a single network, but usually require OOD samples during training (Malinin & Gales, 2018; 2019; Nandy et al., 2020), which may be unrealistic in many applications (Koh et al., 2020). KLoS alleviates this constraint and remains effective without OOD in training by its class-density estimator behaviour. Another range of methods proposes to improve OOD detection on any pre-trained model. Among them, ODIN (Liang et al., 2018) mitigates over-confidence by post-processing logits with temperature scaling and by adding inverse adversarial perturbations. Lee et al. (2018) proposes a confidence score based on the class-conditional Mahalanobis distance, with the assumption of tied covariance. The authors also add inverse adversarial perturbations to improve their method. Although effective, both approaches need OOD data to tune hyperparameters, which might not generalize to other OOD datasets (Shafaei et al., 2019).

4 EXPERIMENTS

In this section, we assess our method against existing baseline uncertainty measures on synthetic data and we conduct extensive experiments across various image datasets, architectures and settings.

Experimental Setup. Uncertainty measures are derived from an evidential model (Eq. (4)) with $\lambda = 10^{-2}$ as in Joo et al. (2020); Malinin & Gales (2019), except for second-order metrics where we found that setting $\lambda = 10^{-3}$ improves performance. We rely on the learned classifier to train our auxiliary confidence model KLoSNet, using the same training set and following loss (Eq. (8)) with $\tau = 1/\lambda$. All details about architectures, training algorithms and datasets are available in Supplementary A.

Baselines. We evaluate our approach against: first-order uncertainty metrics (Maximum Class probability (*MCP*) and predictive entropy (*Entropy*)), second-order metrics (mutual information (*Mut. Inf.*), differential entropy (*Diff. Ent.*) and expected pairwise KL-divergence (*EPKL*)), post-training methods for OOD detection (*ODIN* and *Mahalanobis*) and for misclassification detection (*ConfidNet*). Except in Section 4.3, we consider setups where no OOD data is available for training. Consequently, the results reported for ODIN and Mahalanobis are obtained without adversarial perturbations, which is also the best configuration for the considered tasks. We indeed show in Supplementary B.3 that these perturbations degrade misclassification detection.

4.1 SYNTHETIC EXPERIMENT

We analyse the behavior of the KLoS measure and the limitations of existing first- and second-order uncertainty metrics on a 2D synthetic dataset composed of three Gaussian-distributed classes with equidistant means and identical isotropic variance (Fig. 3). This constitutes a scenario with high first-order uncertainty due to class overlap. OOD samples are drawn from a ring around the in-distribution dataset and are only used for evaluation. Fig. 3c shows that Entropy correctly assigns large uncertainty along decision boundaries, which is convenient to detect misclassifications, but yields low uncertainty for points far from the distribution. Surprisingly, Mut. Inf. (Fig. 3d) decreases

Table 1: **Comparative experiments on CIFAR-10 and CIFAR-100.** Misclassification (Mis.), out-of-distribution (OOD) and simultaneous (Mis+OOD) detection results (mean % AUROC and std. over 5 runs). Bold type indicates significantly best performance ($p < 0.05$) according to paired t-test

	Method	Mis.	**LSUN** OOD	**LSUN** Mis+OOD	**TinyImageNet** OOD	**TinyImageNet** Mis+OOD	**STL-10** OOD	**STL-10** Mis+OOD
CIFAR-10 VGG-16	MCP (Hendrycks & Gimpel, 2017)	87.6 ±1.6	79.7 ±1.1	84.9 ±1.1	80.3 ±1.5	85.2 ±1.5	60.3 ±1.2	75.2 ±1.4
	Entropy (Sensoy et al., 2018)	83.5 ±2.4	83.8 ±0.3	87.9 ±0.2	82.3 ±0.4	87.2 ±0.4	60.1 ±1.2	75.0 ±1.4
	ConfidNet (Corbière et al., 2019)	90.2 ±0.8	82.1 ±1.5	87.6 ±1.1	83.5 ±0.6	88.3 ±0.7	61.5 ±1.6	77.2 ±1.1
	Mut. Inf. (Malinin & Gales, 2019)	84.1 ±1.5	84.6 ±0.6	85.1 ±1.0	80.6 ±0.8	83.4 ±1.1	61.3 ±0.8	65.0 ±2.5
	Diff. Ent. (Malinin & Gales, 2018)	86.8 ±1.0	85.6 ±0.5	87.2 ±0.7	82.7 ±0.7	85.8 ±0.8	62.0 ±1.0	75.4 ±1.3
	EPKL (Malinin, 2019)	83.9 ±1.5	84.5 ±0.7	85.1 ±1.0	80.4 ±0.8	83.2 ±1.2	61.3 ±0.8	73.8 ±1.1
	ODIN (Liang et al., 2018)	86.0 ±2.0	79.5 ±1.2	83.8 ±1.5	79.6 ±1.9	84.0 ±2.0	54.7 ±1.5	65.0 ±2.6
	Mahalanobis (Lee et al., 2018)	91.2 ±0.3	**88.9** ±0.2	**91.3** ±0.1	**86.4** ±0.2	90.2 ±0.1	63.4 ±0.2	78.8 ±0.3
	KLoSNet (Ours)	**92.5** ±0.6	87.6 ±0.9	**91.7** ±0.9	**86.6** ±0.9	**91.2** ±0.8	**67.7** ±1.4	**81.8** ±0.9
CIFAR-10 ResNet-18	MCP (Hendrycks & Gimpel, 2017)	84.9 ±0.8	79.6 ±1.0	83.0 ±0.9	77.2 ±0.7	81.8 ±0.7	58.5 ±1.2	72.5 ±0.4
	Entropy (Sensoy et al., 2018)	84.6 ±0.8	79.6 ±1.1	82.8 ±0.9	77.2 ±0.7	81.6 ±0.7	58.4 ±1.2	72.2 ±0.4
	ConfidNet (Corbière et al., 2019)	90.7 ±0.4	84.6 ±1.1	88.6 ±0.6	83.5 ±1.1	88.0 ±0.6	63.2 ±1.2	77.9 ±0.5
	Mut. Inf (Malinin & Gales, 2019)	80.6 ±0.6	77.0 ±1.2	79.4 ±0.9	74.3 ±0.8	78.0 ±0.7	56.4 ±1.0	69.1 ±0.2
	Diff. Ent (Malinin & Gales, 2018)	82.7 ±0.6	78.3 ±1.2	81.1 ±0.9	75.9 ±0.8	79.9 ±0.7	57.5 ±1.1	70.8 ±0.3
	EPKL (Malinin, 2019)	80.2 ±0.6	76.8 ±1.3	79.0 ±0.9	74.1 ±0.8	77.7 ±0.7	56.2 ±1.0	68.9 ±0.3
	ODIN (Liang et al., 2018)	83.7 ±0.7	78.9 ±1.0	81.9 ±0.9	76.5 ±0.7	80.7 ±0.7	57.9 ±1.2	71.5 ±0.4
	Mahalanobis (Lee et al., 2018)	91.2 ±0.4	90.7 ±0.4	91.8 ±0.3	87.6 ±0.4	90.3 ±0.4	66.8 ±0.5	80.0 ±0.3
	KLoSNet (Ours)	**93.9** ±0.4	**93.1** ±1.1	**94.4** ±0.3	**90.6** ±0.6	**93.2** ±0.2	**68.5** ±0.3	**82.3** ±0.2
CIFAR-100 VGG-16	MCP (Hendrycks & Gimpel, 2017)	82.9 ±0.8	62.8 ±1.3	77.6 ±0.9	72.0 ±0.5	81.8 ±0.7	69.7 ±0.7	80.9 ±0.7
	Entropy (Sensoy et al., 2018)	82.2 ±0.8	63.2 ±1.4	77.2 ±1.0	72.5 ±0.6	81.5 ±0.8	70.1 ±0.8	80.6 ±0.7
	ConfidNet (Corbière et al., 2019)	84.4 ±0.6	65.3 ±2.0	80.0 ±1.3	73.8 ±0.6	83.7 ±0.7	71.5 ±0.6	82.7 ±0.3
	Mut. Inf. (Malinin & Gales, 2019)	78.9 ±0.8	65.6 ±0.7	76.2 ±0.9	71.8 ±0.2	79.1 ±0.4	70.1 ±0.6	78.5 ±0.6
	Diff. Ent. (Malinin & Gales, 2018)	80.2 ±0.8	65.6 ±0.9	77.2 ±0.8	72.7 ±0.3	80.4 ±0.4	71.0 ±0.5	79.7 ±0.5
	EPKL (Malinin, 2019)	78.8 ±0.8	65.2 ±1.0	76.1 ±0.9	71.6 ±0.2	78.9 ±0.4	70.0 ±0.6	78.3 ±0.6
	ODIN (Liang et al., 2018)	82.1 ±0.8	62.9 ±1.4	77.1 ±1.0	71.9 ±0.6	81.3 ±0.8	69.6 ±0.8	80.3 ±0.7
	Mahalanobis (Lee et al., 2018)	84.0 ±0.2	**71.1** ±1.0	82.4 ±0.5	**77.0** ±0.5	84.9 ±0.3	**75.4** ±0.3	84.3 ±0.5
	KLoSNet (Ours)	**86.7** ±0.4	68.4 ±1.1	**83.0** ±0.6	76.4 ±0.4	**86.4** ±0.4	**75.0** ±0.5	**86.0** ±0.4
CIFAR-100 ResNet-18	MCP (Hendrycks & Gimpel, 2017)	84.0 ±0.4	70.4 ±0.9	81.0 ±0.3	76.6 ±0.5	83.6 ±0.4	75.4 ±0.5	83.1 ±0.2
	Entropy (Sensoy et al., 2018)	83.7 ±0.4	70.4 ±0.9	80.8 ±0.3	76.9 ±0.5	83.5 ±0.3	75.7 ±0.5	83.0 ±0.3
	ConfidNet (Corbière et al., 2019)	**87.1** ±0.2	73.0 ±1.4	**84.5** ±0.6	79.1 ±0.3	86.8 ±0.3	**78.5** ±0.8	**86.6** ±0.5
	Mut. Inf (Malinin & Gales, 2019)	82.6 ±0.4	70.2 ±1.1	80.0 ±0.4	76.4 ±0.6	82.6 ±0.3	75.1 ±0.5	82.1 ±0.3
	Diff. Ent (Malinin & Gales, 2018)	83.0 ±0.4	70.1 ±1.1	80.2 ±0.4	76.8 ±0.5	83.0 ±0.3	75.6 ±0.5	82.5 ±0.3
	EPKL (Malinin, 2019)	82.5 ±0.4	70.2 ±1.1	80.0 ±0.4	76.3 ±0.6	82.5 ±0.3	75.0 ±0.5	82.0 ±0.2
	ODIN (Liang et al., 2018)	83.7 ±0.4	70.3 ±0.9	80.8 ±0.3	76.6 ±0.5	83.5 ±0.3	75.4 ±0.5	83.0 ±0.3
	Mahalanobis (Lee et al., 2018)	85.9 ±0.4	**75.2** ±0.6	**84.5** ±0.1	78.4 ±0.5	85.9 ±0.3	77.5 ±0.4	85.6 ±0.3
	KLoSNet (Ours)	86.9 ±0.3	73.1 ±0.4	**84.4** ±0.1	**80.8** ±0.2	**87.3** ±0.2	**79.0** ±0.2	**86.7** ±0.3

when moving away from the training data. This behavior is due to the linear nature of the toy dataset where models assign higher concentration parameters far from decision boundaries, hence smaller spread on the simplex, as also noted in (Charpentier et al., 2020). Additionally, Mut. Inf. does not reflect the uncertainty caused by class confusion along decision boundaries. Neither Entropy nor Mut. Inf. is suitable to detect OOD samples in this synthetic experiment. In contrast, KLoS allows discriminating both misclassifications and OOD samples from correct predictions as uncertainty increases far from in-distribution samples for each class (Fig. 3f). KLoS measures a distance between the model's output and a class-wise target distribution. Here, we can observe that it acts as a density estimator for each class.

We extend the comparison to include Mahalanobis (Fig. 3e), which also boils down to a density estimator by assuming Gaussian class conditionals on latent representations, here in the input space. However, Mahalanobis does not discriminate points close to the decision boundaries from points with a similar distance to the origin. Hence, it may have trouble detecting misclassifications. Additionally, KLoS does not assume Gaussian distributions in the latent space nor tied covariance, which may be a strong assumption when dealing with high-dimension latent space. In Supplementary B.1, a complementary quantitative evaluation on this toy problem confirms our findings regarding the inadequacy of first-order uncertainty measures such as MCP and Entropy, and the improvement provided by KLoS over Mahalanobis on misclassification detection.

Table 2: **Impact of confidence learning.** Comparison of detection performances between KLoS and KLoSNet for CIFAR-10 and CIFAR-100 experiments with VGG-16 architecture.

	Method	Mis.	**LSUN** OOD	**LSUN** Mis+OOD	**TinyImageNet** OOD	**TinyImageNet** Mis+OOD	**STL-10** OOD	**STL-10** Mis+OOD
CIFAR-10 VGG-16	KLoS	92.1 ±0.3	86.5 ±0.3	91.2 ±0.2	85.4 ±0.3	90.4 ±0.2	64.1 ±0.3	79.6 ±0.3
	KLoSNet	**92.5** ±0.6	**87.6** ±0.9	**91.7** ±0.9	**86.6** ±0.9	**91.2** ±0.8	**67.7** ±1.4	**81.8** ±0.9
CIFAR-100 VGG-16	KLoS	85.4 ±0.2	65.1 ±1.1	81.3 ±0.6	74.5 ±0.4	85.4 ±0.4	72.7 ±0.3	84.8 ±0.4
	KLoSNet	**86.7** ±0.4	**68.4** ±1.1	**83.0** ±0.6	**76.4** ±0.4	**86.4** ±0.4	**75.0** ±0.5	**86.0** ±0.4

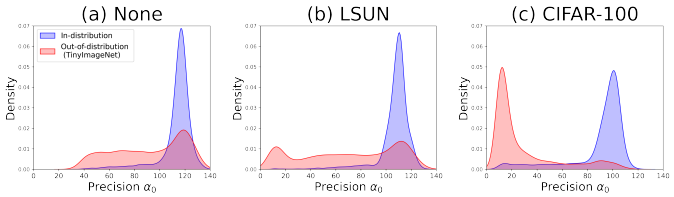


Figure 4: **Effect of OOD training data on precision α_0.** Density plots for CIFAR-10/TinyImageNet benchmark: (a) without OOD training data (None); (b) with inappropriate OOD samples (LSUN); (c) with close OOD samples (CIFAR-100).

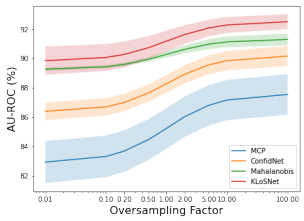


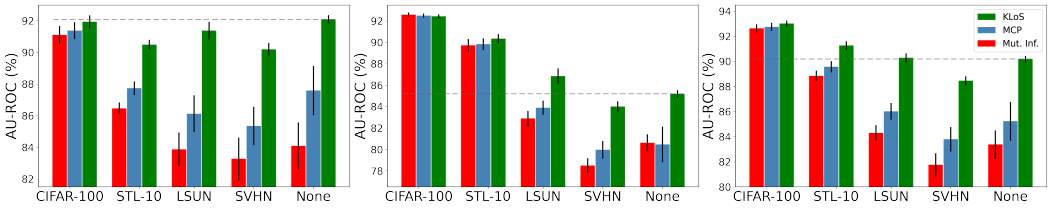
Figure 5: **Impact of the oversampling factor κ** on CIFAR-10/TinyImageNet benchmark.

4.2 Comparative Experiments

The task of detecting both in-distribution misclassifications and OOD samples gives the opportunity to jointly evaluate first-order and second-order uncertainty representations of a method. In this binary classification problem, correct predictions are considered as positive samples while misclassified inputs and OOD examples constitute negative samples. Following standard practices (Hendrycks & Gimpel, 2017), we use the area under the ROC curve (AUROC) to evaluate threshold-independent performance. Results for other relevant metrics are available in Supplementary B.4.

The models used in the experiments present high predictive performances. Most often, there are much fewer misclassifications in the test set than considered OOD samples. Hence, joint detection performances might be dominated by the evaluation of the quality of OOD detection. To mitigate this unbalance, we propose to consider the following scheme based on oversampling. Let $\mathcal{A}_{\mathrm{M}}$ be the subset of in-distribution test examples that are misclassified by the observed model and $\mathcal{A}_{\mathrm{O}}$ the set of OOD test samples. We randomly sample $\kappa|\mathcal{A}_{\mathrm{O}}|$ points in $\mathcal{A}_{\mathrm{M}}$, with $\kappa = 1$. Supposing $|\mathcal{A}_{\mathrm{O}}| \geq |\mathcal{A}_{\mathrm{M}}|$, this corresponds to oversampling the set of misclassifications. This over-sampled set is then added to the OOD set to form the negative examples for detection training. The set of correct predictions remains the same. We observed that the variance in AUROC due to this sampling is negligible and we report only the mean hereafter.

Experiments are conducted with VGG-16 (Simonyan & Zisserman, 2015) and ResNet-18 (He et al., 2016) architectures on CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009). The OOD datasets used for evaluation are LSUN (Yu et al., 2015), TinyImageNet and STL-10 (Coates et al., 2011). Along with simultaneous detection results, we also provide separate results for misclassifications detection and OOD detection respectively in Table 1. On OOD detection, density estimation-based methods such as Mahalanobis and KLoSNet outperform other methods, including second-order measures. ODIN also fails to deliver here as logits are small due to regularization in the evidential training objective. Indeed, for settings where OOD training data is not available, there is no guarantee that every OOD sample will result in lower predicted concentration parameters as shown by the density plot of precision α_0 in Fig. 4a. This stresses the importance of class-wise density estimation. While Mahalanobis may sometimes be slightly better than KLoSNet for OOD detection, it performs significantly less well in misclassification detection, in line with the behavior shown in synthetic experiments. As a result, KLoSNet appears to be the best measure in every simultaneous detection benchmark. For instance, for CIFAR-10/STL-10 with VGG-16, KLoSNet achieves 81.8% AUROC while the second best, Mahalanobis, scores 78.8%. We also observe that KLoS-



(a) Mis. detection (b) OOD detection (c) Mis.+OOD detection

Figure 6: **Comparative detection results with different OOD training datasets.** While using OOD samples in training improves performance in general, the gain value varies widely, sometimes even being negative for inappropriate OOD samples e.g., SVHN. KLoS remains the best measure in every setting. Experiment with VGG-16 architecture on CIFAR-10 dataset.

Net improves significantly misclassification detection, even compared to dedicated methods such as ConfidNet. In Supplementary C, we evaluate our approach on the task of selective classification when inputs are subject to corruptions (which can be seen as OOD samples that are close to the input distribution, hence particularly challenging), and we observe similar results.

Impact of Confidence Learning. To evaluate the effect of the uncertainty measure KLoS and of the auxiliary confidence network KLoSNet, we report a detailed ablation study in Table 2. We can notice that KLoSNet improves misclassification over KLoS but also OOD detection in every benchmark. We intuit that learning to improve misclassification detection also helps to spot some OOD inputs that share similar characteristics.

Oversampling Factor. When deploying a model in the wild, it is difficult to know beforehand the proportions of misclassifications and OOD samples it will have to handle. We vary the oversampling factor κ in $[0.01; 100]$ for CIFAR10/TinyImageNet in Fig. 5 to assess the robustness of our approach. KLoSNet consistently outperforms all other measures, with a larger gain when κ increases.

Combining KLoS with Ensembling. Aggregating predictions from an ensemble of neural networks not only improves generalization (Lakshminarayanan et al., 2017; Rame & Cord, 2021) but also uncertainty estimation (Ovadia et al., 2019a). We train an ensemble of ten evidential model on CIFAR-10 and evaluate the performance of KLoS obtained from averaged concentration parameters. On CIFAR-10/TinyImageNet benchmark, misclassification and OOD detection performances are improved by respectively $+1.8$ points and $+1.9$ points, resulting in a $+1.1$ points gain on the joint detection task.

4.3 Effect of Training with Out-of-Distribution Samples

Previous results demonstrate that simultaneous detection of misclassifications and OOD samples can be significantly improved by KLoSNet in settings without OOD training data. We now investigate settings where OOD samples are available. We train an evidential model to minimize the reverse KL-divergence (Malinin & Gales, 2019) between the model output and a sharp Dirichlet distribution focused on the predictive class for in-distribution samples, and between the model output and a uniform Dirichlet distribution for OOD samples. This loss induces low concentration parameters for OOD data and improves second-order uncertainty measures such as Mut. Inf, as observed in Figs. 4(b-c).

The literature on evidential models only deals with an OOD training set somewhat related to the in-distribution dataset, e.g. CIFAR-100 for models trained on CIFAR-10. Despite semantic differences, CIFAR-10 and CIFAR-100 images were collected the same way, which might explain the generalisation to other OOD samples in evaluation. Contrarily, CIFAR-10 objects and SVHN street-view numbers differ more for instance. In Fig. 6, we vary the OOD training set and compare the uncertainty metrics taken from the resulting models. As expected, using CIFAR-100 as training OOD data improves performance for every measure (MCP, Mut. Inf. and KLoS). However, the boost provided by training with OOD samples depends highly on the chosen dataset: The performance of Mut. Inf. decreases from 92.6% AUC with CIFAR-100 to 82.9% when switching to LSUN, and

even becomes worse with SVHN (78.5%) compared to using no OOD data (80.6%). We also note that KLoS outperforms or is on par with MCP and Mut. Inf. in every setting. These results confirm the adequateness of the KLoS metric for simultaneous detection and extend our findings to settings where OOD data is available at train time. Most importantly, using KLoS on models without OOD training data yields better detection performance than other measures taken from models trained with inappropriate OOD samples, here being every OOD dataset other than CIFAR-100.

5 DISCUSSION

We propose KLoSNet, an auxiliary model to estimate the uncertainty of a classifier for both in-domain and out-of-domain inputs. Based on evidential models, KLoSNet is trained to predict the KLoS* value of a prediction. By design, KLoSNet encompasses both first-order and second-order uncertainties, which is necessary for open-world recognition. Our experiments extensively demonstrate its effectiveness across various architectures, datasets and configurations, and reveal its class-wise density estimation behavior. Many real-world applications, e.g., based on active learning or domain adaptation, necessitate to correctly detect both types of uncertainties. It will be interesting to apply KLoSNet in these contexts. We also show that, far from being the panacea, using training OOD samples depends critically on the choice of these samples for existing uncertainty measures. KLoS, on the other hand, is more robust to this choice and can alleviate their use altogether. While finding suitable OOD samples may be easy for some academic datasets, it may turn more problematic in real-world applications, with the risk of degrading performance with an inappropriate choice. How to build suitable OOD training sets is an important open problem to attack.

REFERENCES

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 1

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In *Advances in Neural Information Processing Systems*, 2020. 2, 7

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011. 8

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, 2019. 5, 7

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016. 5

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, 2016. 1, 6

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 8

David Heckerman, Eric Horvitz, and Bharat Nathwani. Toward normative expert systems: Part i the pathfinder project. *Methods of information in medicine*, 31, 07 1992. doi: 10.1055/s-0038-1634867. 1

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017. 1, 5, 7, 8

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, 2018. 5

Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being bayesian about categorical probability. In *Proceedings of the International Conference on Machine Learning*, 2020. 1, 2, 3, 6

Audun Josang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 2016. 1

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. *arxiv*, 2020. 6

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 8

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. 1, 6, 9

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018. 6, 7

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of International Conference on Learning Representations*, 2018. 6, 7

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems 32*, 2019. URL `http://papers.nips.cc/paper/9472-a-simple-baseline-for-bayesian-uncertainty-in-deep-learning.pdf`. 1, 6

Andrey Malinin. *Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment*. PhD thesis, University of Cambridge, 2019. 7

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018. 2, 6, 7

Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019. 1, 2, 4, 6, 7, 9

Rowan McAllister, Yarin Gal, Alex Kendall, Mark van der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017. 1

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *Proceedings of the International Conference on Machine Learning*, 2020. 5

Kevin P. Murphy. *Machine learning : A Probabilistic Perspective*. MIT Press, 2012. 2, 3

Jay Nandy, Wynne Hsu, and Mong-Li Lee. Towards maximizing the representation gap between in-domain and out-of-distribution examples. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 6

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996. 6

A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 1

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019a. 9

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019b. 1

Alexandre Rame and Matthieu Cord. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *International Conference on Learning Representations*, 2021. 6, 9

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 2018. 1, 2, 7

Murat Sensoy, Lance Kaplan, Federico Cerutti, and Maryam Saleki. Uncertainty-aware deep classifiers using generative models. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020. 2

Alireza Shafaei, Mark Schmidt, and James J. Little. A less biased evaluation of out-of-distribution sample detectors. In *BMVC*, 2019. 6

W. Shi, Xujiang Zhao, Feng Chen, and Qi Yu. Multifaceted uncertainty estimation for label-efficient deep learning. In *Advances in Neural Information Processing Systems*, 2020. 1

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015. 8

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 8

A Appendix

You may include other additional sections here.