WHAT CLASSIFIERS KNOW WHAT THEY DON’T?

Mohamed Ishmael Belghazi and David Lopez-Paz
{ishmaelb, dlp}@fb.com
Facebook AI Research, Paris, France

ABSTRACT

Being uncertain when facing the unknown is key to intelligent decision making. However, machine learning algorithms lack reliable estimates about their predictive uncertainty. This leads to wrong and overly-confident decisions when encountering classes unseen during training. Despite the importance of equipping classifiers with uncertainty estimates ready for the real world, prior work has focused on small datasets and little or no class discrepancy between training and testing data. To close this gap, we introduce UIMNET: a realistic, ImageNet-scale test-bed to evaluate predictive uncertainty estimates for deep image classifiers. Our benchmark provides implementations of eight state-of-the-art algorithms, six uncertainty measures, four in-domain metrics, three out-domain metrics, and a fully automated pipeline to train, calibrate, ensemble, select, and evaluate models. Our test-bed is open-source and all of our results are reproducible from a fixed commit in our repository. Adding new datasets, algorithms, measures, or metrics is a matter of a few lines of code—in so hoping that UIMNET becomes a stepping stone towards realistic, rigorous, and reproducible research in uncertainty estimation. Our results show that ensembles of ERM classifiers as well as single MIMO classifiers are the two best alternatives currently available to measure uncertainty about both in-domain and out-domain classes.

1 Introduction

I don’t think I’ve ever seen anything quite like this before
—HAL 9000 in 2001: A Space Odyssey

Deep image classifiers exceed at discriminating the set of in-domain classes observed during training. However, when confronting test examples from unseen out-domain classes, these classifiers can only predict in terms of their known in-domain categories, leading to wrong and overly-confident decisions (Hein et al., 2019; Ulmer and Cinà, 2020). In short, machine learning systems are unaware of their own limits, since “they do not know what they do not know”. Because of this reason, out-domain data cannot be safely identified and treated accordingly. Thus, it is reasonable to fear that, when deployed in-the-wild, the behavior of these classifiers becomes unpredictable and their performance crumbles by leaps and bounds (Ovadia et al., 2019).

The inability of machine learning systems to estimate their uncertainty and abstaining to classify out-domain classes is roadblock towards their implementation in critical applications. These include self-driving (Michelmore et al., 2018), medicine (Begoli et al., 2019), and the analysis of satellite imagery (Wadoux, 2019). Good uncertainty estimates are also a key ingredient in anomaly detection (Chalapathy and Chawla, 2019), active learning (Settles, 2009), safe reinforcement learning (Henaff et al., 2019), defending against adversarial examples (Goodfellow et al., 2014), and model interpretability (Alvarez-Melis and Jaakkola, 2017).

For an extensive literature review on uncertainty estimation and its applications, we refer the curious reader to the survey of Abdar et al. (2020) and the one of Ruff et al. (2021). Despite a research effort spanning multiple decades, machine learning systems still lack trustworthy estimates of their predictive uncertainty. In our view, one hindrance to this research program is the absence of realistic benchmarking and evaluation protocols. More specifically, prior attempts are limited in two fundamental ways. First, these experiment on small datasets such as SVHN and CIFAR-10 (van Amersfoort et al., 2021). Second, these do not provide a challenging set of out-domain data. Instead, they construct out-domain classes by using a second dataset (e.g., using MNIST in-domain versus FashionMNIST out-domain, cf. Van Amersfoort et al. (2020)) or by perturbing the in-domain classes using handcrafted transformations (such as Gaussian noise or blur,
What classifiers know what they don’t?

| Datasets       | Algorithms  | Uncertainty measures | In-Domain metrics | Out-Domain metrics | Ablations         |
|----------------|-------------|-----------------------|-------------------|--------------------|-------------------|
| ImageNot       | ERM         | Largest               | ACC@1             | AUC                | Calibration (y/n) |
|                | Mixup       | Gap                   | ACC@5             | InAsIn             | Spectral norm (y/n) |
|                | Soft-labeler| Entropy               | ECE               | InAsOut            | Model size (RN18 / RN50) |
|                | RBF         | Jacobian              | NLL               | OutAsIn            |                   |
|                | RND         | GMM                   |                   | OutAsOut           |                   |
|                | OCs         | Native                |                   |                    |                   |
|                | MC-Dropout  |                       |                   |                    |                   |
|                | MIMO        | (+ Ensembles)         |                   |                    |                   |

Table 1: The UIMNET test-bed suite for uncertainty estimation.

see ImageNet-C Hendrycks and Dietterich (2019)). Both approaches result in simplistic benchmarking, and little is learned about uncertainty estimation for the real world. The purpose of this work is to introduce an end-to-end benchmark and evaluation protocol bridging this disconnect. At the time of writing, UIMNET is the most exhaustive benchmark for uncertainty estimation in the literature.

**Formal setup** Following conventional supervised learning, we learn a classifier $f$ using in-domain data from the distribution $P_{in}(x, y)$. After training, we endow the classifier with a real-valued uncertainty measure $u(f, x^\dagger)$. Given a test example $(x^\dagger, y^\dagger) \sim P$ with unobserved label $y^\dagger$, we declare $x^\dagger$ in-domain (hypothesizing $P = P_{in}$) if $u(f, x^\dagger)$ is small, whereas we declare $x^\dagger$ out-domain (hypothesizing $P \neq P_{in}$) if $u(f, x^\dagger)$ is large. Using these tools, our goal is to abstain from classifying out-domain test examples, and to classify with calibrated probabilities in-domain test examples. The sequel assumes that the difference between in- and out-domain resides in that the two groups of data concern disjoint classes.

**Contributions** We introduce UIMNET, a test-bed for large-scale, realistic evaluation of uncertainty estimates in deep image classifiers. We outline the components of the test-bed below, also summarized in Table 1.

(Sec. 2) We construct ImageNot, a perceptual partition of ImageNet into in-domain and out-domain classes. Unlike prior work focused on small datasets like SVHN and CIFAR-10, ImageNot provides a benchmark for uncertainty estimators at a much larger scale. Moreover, both in-domain and out-domain categories in ImageNot originate from the original ImageNet dataset. This provides realistic out-domain data, as opposed to prior work relying on a second dataset (e.g., MNIST as in-domain versus SVHN as out-domain), or handcrafted perturbations of in-domain classes (Gaussian noise or blur as out-domain).

(Sec. 3) We re-implement eight state-of-the-art algorithms from scratch, listed in Table 1. This allows a fair comparison under the exact same experimental conditions (training/validation splits, hyperparameter search, neural network architectures and random initializations). Furthermore, we also study ensembles of multiple training instances for each algorithm.

(Sec. 4) Each algorithm can be endowed with one out of six possible uncertainty measures, allowing an exhaustive study of what algorithms play well with what measures. Listed in Table 1, these are the largest softmax score, the gap between the two largest softmax scores, the softmax entropy, the norm of the Jacobian, a per-class Gaussian density model, and (for those available) an algorithm-specific measure.

(Sec. 5) For each classifier-measure pair, we study four in-domain metrics (top-1 and top-5 classification accuracy, log-likelihood, expected calibration error) and three out-domain metrics (the AUC at classifying in-domain versus out-domain samples using the selected uncertainty measure, as well as the confusion matrix at a fixed uncertainty threshold computed over an in-domain validation set).

(Sec. 6) We repeat our entire pipeline to accommodate three popular ablations to understand the impact of model calibration by temperature scaling, model size, and the use of spectral normalization.

2
(Sec. 7) UIMNET is entirely hands-off, since the pipeline from zero to \LaTeX{} tables is fully automated: this includes hyper-parameter search, model calibration, model ensembling, and the production of all the tables included in our experimental results.

(Sec. 8) Our experimental results illustrate that no classifier but those trained by using Multiple-Input and Multiple-Output (MIMO) systematically dominates calibrated classifiers trained by empirical risk minimization. Interestingly, a single MIMO classifier if \textit{given enough capacity} matches ensembles of ERMs in terms of AUC score and identification of in-domain instances and even edges them when identifying out-domain instances. Additionally, our result show that the use of spectral normalization does not significantly improve out of domain detection metrics while the effect of calibration remains mild at best.

UIMNET is open sourced at \url{https://github.com/facebookresearch/uimnet}. All of the tables presented in this paper are reproducible by running the main script in the repository at commit 0xSOON.

2 Constructing the ImageNot benchmark

The ImageNet dataset (Russakovsky et al., 2015), a gold standard to conduct research in supervised learning for computer vision, pertains the classification of images into 1000 different classes. Here we will use the ImageNet dataset to derive ImageNot, a large-scale and realistic benchmark for uncertainty estimation. In a nutshell, we will divide the 1000 classes of the original ImageNet dataset into in-domain classes (used to train and evaluate algorithms in-distribution) and out-domain classes (used to evaluate algorithms out-of-distribution).

To partition ImageNet into in-domain and out-domain, we featurize the entire dataset to understand the perceptual similarity between classes. To this end, we use a pre-trained ResNet-18 (He et al., 2016) to compute the average last-layer representation for each of the classes. Next, we use agglomerative hierarchical clustering Ward Jr (1963) to construct a tree describing the perceptual similarities between the 1000 average feature vectors. Such perceptual tree has 1000 leafs, each of them being a cluster containing one of the classes. During each step of the iterative agglomerative clustering algorithm the two closest clusters are merged, where the distance between two clusters is computed using the criterion of Ward Jr (1963). The algorithm halts when there are only two clusters left to merge, forming the root node of the tree.

At this point, we declare the 266 classes to the left of the root as in-domain, and the first 266 classes to the right of the root as out-domain. In the sequel, we call “training set” and “validation set” to a 90/10 random split from the original ImageNet “train” set. We call “testing set” to the original ImageNet “val” split. The exact in-/out-domain class partition as well as the considered train/validation splits are specified in Appendix C.

While inspired by the BREEDS dataset (Santurkar et al., 2020), our benchmark ImageNot is conceived to tackle a different problem. The aim of the BREEDS dataset is to classify ImageNet into a small number of super-classes, each of them containing a number of perceptually-similar sub-classes. The BREEDS training and testing distributions differ on the sub-class proportions contributing to their super-classes. Since the BREEDS task is to classify super-classes, the set of labels remains constant from training to testing conditions. This is in contrast to ImageNot, where the algorithm observes only in-domain classes during training, but both in-domain and out-domain classes during evaluation. While BREEDS studies the important problem of domain generalization (Gulrajani and Lopez-Paz, 2020), where there is always a right prediction to make within the in-domain classes during evaluation, here we focus on measuring uncertainty and abstaining from out-domain classes unseen during training.

Also related, our benchmark ImageNot is similar to the ImageNet-O dataset of (Hendrycks et al., 2021). However, their out-domain classes are obtained from heterogeneous sources outsider of ImageNet, which have distinct statistics that can be easily picked up by the classifier. In contrast, the starting point for both in-domain and out-domain classes of our ImageNot is the same (the original ImageNet dataset) and thus should maximally overlap in terms of image statistics, leading to a more challenging, realistic benchmark.

3 Algorithms

We benchmark eight supervised learning algorithms that are commonly applied to tasks involving uncertainty estimation. Each algorithm consumes one in-domain training set of image-label pairs \( \{ (x_i, y_i) \}_{i=1}^n \) and
returns a predictor $f(x) = w(φ(x))$, composed by a featurizer $φ : \mathbb{R}^{3\times224\times224} \to \mathbb{R}^K$ and a classifier $w : \mathbb{R}^K \to \mathbb{R}^C$. We consider predictors implemented using deep convolutional neural networks (LeCun et al., 2015). Given an input image $x^t$, all predictors return a softmax vector $f(x^t) = (f(x^t)_c)^T_{c=1}$ over $C$ classes. The considered algorithms are:

- **Empirical Risk Minimization**, or ERM (Vapnik, 1992), or vanilla training.
- **Mixup** (Zhang et al., 2017) chooses a predictor minimizing the empirical risk on mixed examples $(\tilde{x}, \tilde{y})$, built as:

  \[ \begin{align*}
  \lambda &\sim \text{Beta}(\alpha, \alpha), \\
  \tilde{x} &\sim \lambda \cdot x_i + (1 - \lambda) \cdot x_j, \\
  \tilde{y} &\sim \lambda \cdot y_i + (1 - \lambda) \cdot y_j,
  \end{align*} \]

  where $\alpha$ is a mixing parameter, and $((x_i, y_i), (x_j, y_j))$ is a random pair of training examples. Mixup has been shown to improve both generalization performance (Zhang et al., 2017) and calibration error (Thulasidasan et al., 2019).
- **Random Network Distillation**, or RND (Burda et al., 2018), finds an ERM predictor $f(x) = w(φ(x))$, but also trains an auxiliary classifier $w_{\text{student}}$ to minimize

\[ \|w_{\text{student}}(φ(x)) - w_{\text{teacher}}(φ(x))\|_2^2, \]

where $w_{\text{teacher}}$ is a fixed classifier with random weights. RND has shown good performance as a tool for exploration in reinforcement learning.
- **Orthogonal Certificates**, or OC (Tagasovska and Lopez-Paz, 2018), is analogous to RND for $w_{\text{teacher}}(φ(x)) = \tilde{0}_K$ for all $x$. That is, the goal of $w_{\text{student}}$ is to map all the in-domain training examples to zero in $k$ different ways (or certificates). To ensure diverse and non-trivial certificates, we regularize each weight matrix $W$ of $w_{\text{student}}$ to be orthogonal by adding a regularization term $\|W^TW - I\|_2$. OCs have shown good performance at the task of estimating uncertainty across a variety of classification tasks.
- **MC-Dropout** (Gal and Ghahramani, 2016) uses ERM over a family of predictors with one or more dropout layers (Srivastava et al., 2014). These stochastic dropout layers remain active at test time, allowing the predictor to produce multiple softmax vectors $\{f(x^t, \text{dropout}_T)\}^3_{T=1}$ for each test example $x^t$. Here, dropout is a random dropout mask sampled anew. MCDropout is one of the most popular baselines to estimate uncertainty.
- **MIMO** (Havasi et al., 2021) is a variant of ERM over predictors accepting $T$ images and producing $T$ softmax vectors. For example, MIMO with $T = 3$ is trained to predict jointly the label vector $(y_1, y_2, y_3)$ using a predictor $h(x, x_1, x_2)$, where $((x_i, y_i))^{3}_{i=1}$ is a random triplet of training examples. Given a test point $x^t$, we form predictions by replicating and averaging, that is $f(x^t) = \frac{1}{3} \sum_{i=1}^{3} h(x^t, x_i, x^t)$. 
- **Radial Basis Function**, or RBF (Broomhead and Lowe, 1988), is a variant of ERM where we transform the logit vector $z \mapsto e^{-z^2}$ before passing them to the final softmax layer. In such a way, the logit norm $\|z\| \to \infty$, the predicted softmax vector tends to the maximum entropy solution $\frac{1}{C}C$, signaling high uncertainty far away from the training data. RBFs have been proposed as defense to adversarial examples (Goodfellow et al., 2014), but they remain under-explored given the difficulties involved in their training.
- **Soft labeler** (Hinton et al., 2015; Szegedy et al., 2016) is a variant of ERM where the one-hot vector labels $y_i$ are smoothed such that every zero becomes $\ell_{\text{min}} > 0$ and the solitary one becomes $\ell_{\text{max}} < 1$. Softening labels avoids saturation in the final softmax layer in neural network predictors, one of the main causes of overly-confident predictors. Using soft labels, we can identify softmax vectors with entries exceeding $\ell_{\text{max}}$ as “over-shoots”, regarding them as uncertain extrapolations.

**Ensembles of predictors** We also consider ensembles of predictors trained by each of the algorithms above. Ensembles are commonly regarded as the state-of-the-art in uncertainty estimation (Lakshminarayanan et al., 2016). In particular, and for each algorithm, we construct bagging ensembles by (i) selecting the best $K \in \{1, 5\}$ predictors $\{f_k\}^K_{k=1}$ from all considered random initializations and hyper-parameters, and (ii) returning the average function $f(x^t) := \frac{1}{K} \sum_{k=1}^{M} f^k(x^t)$. 

\[1\]In the sequel, we denote by classifier the last few layers in the predictor that follow after the featurizer.
4 Uncertainty measures

Using UIMNET, we can equip a trained predictor \( f \) with six different uncertainty measures. An uncertainty measure is a real-valued function \( u(f, x^t) \) designed to return small values for in-domain instances \( x^t \), and large values for out-domain instances \( x^t \). To describe the different measures, let \( \{ s_1, \ldots, s_C \} \) be the softmax scores returned by \( f(x^t) \) sorted in decreasing order.

- **Largest** (Hendrycks and Gimpel, 2016) returns minus the largest softmax score, \(-s_{(1)}\)
- **Softmax gap** (Tagasovska and Lopez-Paz, 2018) returns \( s_{(2)} - s_{(1)} \).
- **Entropy** (Shannon, 1948) returns \(-\sum_{c=1}^{C} s_c \log s_c \).
- **Norm of the Jacobian** (Novak et al., 2018) returns \( \| \nabla_{x} f(x^t) \|_2^2 \).
- **GMM** (Mukhoti et al., 2021) estimates one Gaussian density \( \mathcal{N}(\phi(x); \mu_c, \Sigma_c) \) per-class, on top of the feature vectors \( \phi(x) \) collected from a in-domain validation set. Given a test example \( x^t \), return \(-\sum_{c=1}^{C} \lambda_c \cdot \mathcal{N}(\phi(x^t); \mu_c, \Sigma_c) \), where \( \lambda_c \) is the proportion of in-domain validation examples from class \( c \).
- **Test-time augmentation** (Ashukha et al., 2020) returns \(-\max_c (\frac{1}{A} \sum_{a=1}^{A} f(x^t_a)) \). This is the measure “Largest” about the average prediction over \( A \) random data augmentations \( \{ x^t_a \}_{a=1}^{A} \) of the test instance \( x^t \).

These uncertainty measures are applicable to all the algorithms considered in Section 3. Additionally, some algorithms provide their Native uncertainty measures, outlined below.

- For **Mixup**, we return \( \| \lambda \cdot f(x^t) + (1 - \lambda) \cdot \bar{y} - f(\lambda \cdot x^t + (1 - \lambda) \cdot \bar{x}) \|_2^2 \), where we recall that \( \lambda \sim \text{Beta}(\alpha, \alpha) \), and \((\bar{x}, \bar{y})\) is the average image and label from the training set. This measures if the test example \( x^t \) violates the Mixup criterion wrt the training dataset average.
- For **RND and OC**, we return \( \| w_{\text{student}}(\phi(x^t)) - w_{\text{teacher}}(\phi(x^t)) \|_2^2 \), that is, we consider a prediction uncertain if the outputs of the student and teacher disagree. We expect this disagreement to be related predictive uncertainty, as the student did not observe the behaviour of the teacher at out-domain instances \( x^t \).
- For **Soft labeler** we return \((s_{(1)} - \ell_{\text{max}})^2\). This measures the discrepancy between the largest softmax and the positive soft label target, able to signal overly-confident predictions.
- For **MC-Dropout** and **Ensembles**, and following (Lakshminarayanan et al., 2016), we return the Jensen-Shannon divergence between the \( K \) members (or stochastic forward passes) \( f^1, \ldots, f^K \) of the ensemble:

\[
H \left( \frac{1}{K} \sum_{k=1}^{K} f^k(x^t) \right) - \frac{1}{K} \sum_{k=1}^{K} H(f^k(x^t)).
\]

Note that the models ERM, OC, and RND are equivalent in all aspects except in comparisons involving the uncertainty measure Native. This is because the only difference between these three models is the training of an external student for RND and OC, used only in their Native uncertainty measures.

5 Evaluation metrics

For each algorithm-measure pair, we evaluate several metrics both in-domain and out-domain.

5.1 In-domain metrics

Following (Havasi et al., 2021), we implement four metrics to assess the performance and calibration of predictors when facing in-domain test examples.

- **Top-1 and Top-5** classification accuracy (Russakovsky et al., 2015).
- **Expected Calibration Error** or ECE (Guo et al., 2017):

\[
\frac{1}{B} \sum_{b=1}^{B} \left| \frac{B_b}{n} \cdot \text{acc}(f, B_b) - \text{conf}(f, B_b) \right|,
\]
where $B_b$ contains the examples where the algorithm predicts a softmax score of $b$. The functions $acc$ and $conf$ compute the average classification accuracy and largest softmax score of $f$ over $B_b$. In a nutshell, ECE is minimized when $f$ is calibrated, that is, $f$ is wrong $p\%$ of the times it predicts a largest softmax score $p$. Following (Guo et al., 2017), we discrete $b \in [0, 1]$ into 15 equally-spaced bins.

- **Negative Log Likelihood (NLL)** Also known as the cross-entropy loss, this is the objective minimized during the training process of the algorithms.

### 5.2 Out-domain metrics

After measuring the performance of a predictor in-domain, we equip it with an uncertainty measure. We assess the uncertainty estimates of each predictor-measure pair using three metrics:

- **Area Under the Curve**, or AUC (Tagasovska and Lopez-Paz, 2018), describes how well does the predictor-measure pair distinguish between in-domain and out-domain examples over all thresholds of the uncertainty measure.

- **Confusion matrix at fixed threshold**. To reject out-domain examples in real scenarios, one must fix a threshold $\theta$ for the selected uncertainty measure. We do so by computing the 95% quantile of the uncertainty measure, computed over an in-domain validation set. Then, at testing time, we declare one example out-domain if the uncertainty measure exceeds $\theta$. This strategy is equivalent to the statistical hypothesis test with null “H$_0$: the observed example is in-domain”. To understand where does the uncertainty measure hit or miss, we monitor two metrics: $\text{InAsIn}$ (percentage of in-domain examples classified as in-domain), $\text{OutAsOut}$ (percentage of in-domain examples classified as out-domain). Please note that From these two metrics, we can deduce: $\text{InAsOut}$ (percentage in-domain examples classified as out-domain, also known as false positives or type-I errors), $\text{OutAsIn}$ (percentage out-domain examples classified as in-domain, also known as false negatives or type-II errors).

### 6 Ablations

We execute our entire test-bed under three additional ablations often discussed in the literature of uncertainty estimation.

- We study the effect of calibration by temperature scaling (Platt et al., 1999). To this end, we introduce a temperature scaling $\tau > 0$ before the softmax layer, resulting in predictions $\text{Softmax}(\frac{z}{\tau})$ about the logit vector $z$. We estimate the optimal temperature $\hat{\tau}$ by minimizing the NLL of the predictor across an in-domain validation set. We evaluate all metrics for both the un-calibrated ($\tau = 1$) and calibrated ($\tau = \hat{\tau}$) predictors. According to previous literature (Guo et al., 2017), calibrated models provide better in-domain uncertainty estimates.

- We analyze the impact of spectral normalization applied to the featurizer $\phi$. Several recent works (Liu et al., 2020; Van Amersfoort et al., 2020; van Amersfoort et al., 2021; Mukhoti et al., 2021) have highlighted the importance of controlling both the smoothness and sensitivity of the feature extraction process to achieve high-quality uncertainty estimates. On the one hand, enforcing smoothness upper bounds the Lipschitz constant of $\phi$, limiting its reaction to changes in the input. Smoothness is often enforced by normalizing each weight matrix in $\phi$ by its spectral norm (Miyato et al., 2018). On the other hand, enforcing sensitivity lower bounds the Lipschitz constant of $\phi$, ensuring that the feature space reacts in some amount when the input changes. Sensitivity is often enforced by residual connections (He et al., 2016), present in the ResNet models that we will use throughout our experiments.

- We analyze the impact of the model size, running experiments with both ResNet-18 and ResNet-50 models (He et al., 2016).

### 7 Experimental protocol

We are now ready to conduct experiments on the ImageNet benchmark for uncertainty estimation (Section 2) for all combinations of algorithms (Section 3) and measures (Section 4), that we will evaluate under all metrics (Section 5) and ablations (Section 6).
What classifiers know what they don’t?

Table 2: Default hyper-parameters and random search grids for all algorithms.

| Algorithm    | Hyper-parameter       | Default value | Random search distribution |
|--------------|-----------------------|---------------|---------------------------|
| all          | learning rate         | 0.1           | $1^{\text{Uniform}}(-2, -0.3)$ |
|              | momentum              | 0.9           | $1^{\text{Choice}}([0.5, 0.9, 0.99])$ |
|              | weight decay          | $10^{-4}$     | $1^{\text{Uniform}}(-5, -3)$  |
| Mixup        | mixing parameter      | 0.3           | $1^{\text{Choice}}([0.1, 0.2, 0.3, 1, 2])$ |
| MC-Dropout   | dropout rate          | 0.05          | $1^{\text{Choice}}([0.05, 0.1, 0.2])$ |
|              | number of passes      | 10            | $1^{\text{Choice}}([10])$    |
| MIMO         | number of subnetworks | 2             | $1^{\text{RandInt}}(2, 5)$   |
|              | prob. input repetition| 0.6           | $1^{\text{Uniform}}(0, 1)$   |
|              | batch repetition      | 2             | $1^{\text{RandInt}}(1, 5)$   |
| RND, OC      | teacher width         | 128           | $1^{\text{Choice}}([64, 128, 256])$ |
|              | teacher depth         | 3             | $1^{\text{Choice}}([2, 3, 4])$ |
|              | regularization        | 0             | $1^{\text{Uniform}}(-2, 1)$  |
| Soft labeler | soft label value      | 128           | $1^{\text{Choice}}([0.7, 0.8, 0.9])$ |

Hyper-parameter search. We train each algorithm sixty times, arising from a combination of (i) ResNet-18 or ResNet-50 architectures (ii) using or not spectral normalization, (iii) five hyper-parameter trials, and (iv) three random train/validation splits of the in-domain data (data seeds). We opt for a random hyper-parameter search (Bergstra and Bengio, 2012), where the search grid for each algorithm is detailed in Table 2. More specifically, while the first trial uses the default hyper-parameter configuration suggested by the authors of each algorithm, the additional four trials explore random hyper-parameters.

Model selection. After training all instances of a given algorithm, we report the average and standard deviation (over data seeds) for all metrics of two chosen models. On the one hand, we report metrics for the best model ($k = 1$), which is the one minimizing the average (over data seeds) log-likelihood in the in-domain validation set. On the other hand, we report metrics also for the ensemble model formed by the five different hyper-parameter trials. Finally, we report all results of both the best and the ensemble model with and without calibration and for all model sizes (Section 6).

Software and hardware. All algorithms are implemented in PyTorch (Paszke et al., 2019) and train ResNet-18 or ResNet-50 backbones (He et al., 2016) initialized by following (Glorot and Bengio, 2010). In addition, the Gaussian process in DUE is implemented in GPyTorch (Gardner et al., 2018). These models are trained by stochastic gradient descent (Robbins and Monro, 1951; Bottou, 2012) for 100 epochs, with mini-batches of 256 examples distributed over 8 NVIDIA Tesla V100 GPUs (yielding an effective mini-batch of size 32 per GPU). During training, we decay the learning rate by a factor of 10 every 30 epochs.

8 Conclusions

Table 3 summarizes our results for in-domain metrics (Section 5.1), and Table 4 summarizes our results for out-domain metrics (Section 5.2) and the best performing measure (Entropy, see Section 4). These tables contain the results for all algorithms (Section 3), ablations (use of spectral normalization, use of temperature calibration; see Section 6), and ensembling (whether bagging the best $k = 1$ or $k = 5$ models) for the largest model (ResNet50) considered. Appendix A contains the in-domain results for ResNet18, and Appendix B contains the out-domain result tables for all the other measures. From the in-domain results summarized in Table 3, we identify the following key takeaways:

- No algorithm significantly outperforms ERM in any in-domain metric.
- Ensembling multiple models ($k = 5$) improves all in-domain metrics.
- Temperature calibration helps, decreasing the expected calibration error by up to 30%.
- Spectral normalization has a neutral and marginal effect on in-domain metrics.
- Most algorithms are able to achieve similar performances, as their hyper-parameter searches allow them to behave like ERM.
Next, from the **out-domain** results summarized in Table 4, we identify the following key takeaways:

- A single MIMO is the only classifier outperforming ensembles of ERM classifiers on the OutAsOut metric, or the rate of correct identification of out-domain instances.
- No other algorithm significantly outperforms ERM in any out-domain metric.
- Ensembling models ($k = 5$) improves all out-domain metrics for all classifiers but MIMO.
- Temperature calibration has a neutral and marginal effect on out-domain metrics.
- Spectral normalization has a marginal negative effect on out-domain metrics.
- Most algorithms are able to achieve similar performances, as their hyper-parameter searches allow them to behave like ERM.

Appendix B, summarizing the **out-domain** performance of all other uncertainty measures, provides us with additional takeaways, in particular:

- The best performing uncertainty measures, are Entropy, Largest, and Gap, with Entropy taking the lead with a generous gap.
- The uncertainty measures Augmentations and Jacobian exhibit a poor performance.
- Increasing model size leads to large improvements in the out-domain metric OutAsOut.
- Given enough capacity, a single MIMO classifier surpasses ensembles of 5 ERM classifiers on the out-domain metric OutAsOut.
- The Native uncertainty measures (those specific measures provided by some algorithms) do not exhibit a good performance.

From all of these results, our recommendation is to use **calibrated ensembles of ERM classifiers** or a single **MIMO classifier**, with no need for spectral normalization, and allowing the largest model size. When computational resources do not allow large model sizes or the use of ensembles, we recommend to use a single **calibrated ERM models**.

**Other results** Despite our best efforts, we were unable to obtain competitive performances when using the algorithms RBF, (Broomhead and Lowe, 1988) or the uncertainty measure GMM (Mukhoti et al., 2021). We believe that training RBFs at this large scale are challenging optimization problems that deserve further study in our community. Furthermore, we believe that the large number of classes in our study (266 ImageNET classes instead of the 10 CIFAR-10 classes often considered) makes for a difficult problem for density-based uncertainty measures such as GMM.
| algorithm | spectral | calibration | k | ACC@1  | ACC@5  | ECE    | NLL    |
|-----------|----------|-------------|---|--------|--------|--------|--------|
| ERM       | False    | initial     | 1.0 | 0.762 ± 0.004 | 0.859 ± 0.001 | 0.050 ± 0.003 | 0.952 ± 0.013 |
|           |          | learned     | 5.0 | 0.782 ± 0.001 | 0.871 ± 0.002 | 0.030 ± 0.001 | 0.855 ± 0.001 |
|           | True     | initial     | 1.0 | 0.763 ± 0.001 | 0.859 ± 0.004 | 0.046 ± 0.001 | 0.945 ± 0.009 |
|           |          | learned     | 5.0 | 0.782 ± 0.001 | 0.871 ± 0.002 | 0.034 ± 0.007 | 0.864 ± 0.016 |
| MCDropout | False    | initial     | 1.0 | 0.764 ± 0.002 | 0.859 ± 0.001 | 0.054 ± 0.002 | 0.957 ± 0.009 |
|           |          | learned     | 5.0 | 0.782 ± 0.001 | 0.872 ± 0.002 | 0.031 ± 0.001 | 0.853 ± 0.003 |
| MIMO      | False    | initial     | 1.0 | 0.764 ± 0.002 | 0.859 ± 0.001 | 0.045 ± 0.001 | 0.947 ± 0.011 |
|           |          | learned     | 5.0 | 0.782 ± 0.001 | 0.872 ± 0.002 | 0.033 ± 0.005 | 0.861 ± 0.011 |
| Mixup     | False    | initial     | 1.0 | 0.768 ± 0.005 | 0.862 ± 0.003 | 0.050 ± 0.002 | 0.932 ± 0.017 |
|           |          | learned     | 5.0 | 0.769 ± 0.001 | 0.862 ± 0.001 | 0.057 ± 0.004 | 0.925 ± 0.005 |
| OC        | False    | initial     | 1.0 | 0.767 ± 0.003 | 0.863 ± 0.002 | 0.054 ± 0.009 | 0.949 ± 0.050 |
|           |          | learned     | 5.0 | 0.769 ± 0.002 | 0.863 ± 0.002 | 0.052 ± 0.002 | 0.937 ± 0.005 |
| RND       | False    | initial     | 1.0 | 0.765 ± 0.004 | 0.860 ± 0.002 | 0.054 ± 0.003 | 0.956 ± 0.016 |
|           |          | learned     | 5.0 | 0.781 ± 0.001 | 0.871 ± 0.001 | 0.029 ± 0.001 | 0.855 ± 0.004 |
| SoftLabeler| False   | initial     | 1.0 | 0.763 ± 0.002 | 0.857 ± 0.001 | 0.035 ± 0.004 | 1.001 ± 0.005 |
|           |          | learned     | 5.0 | 0.781 ± 0.001 | 0.874 ± 0.002 | 0.131 ± 0.003 | 1.003 ± 0.002 |
|           | True     | initial     | 1.0 | 0.761 ± 0.003 | 0.858 ± 0.002 | 0.092 ± 0.002 | 1.043 ± 0.006 |

Table 3: In-domain results for backbone ResNet50.
| Algorithm | Spectral | Calibration | k | AUC      | InAsIn  | OutAsOut |
|-----------|----------|-------------|---|----------|---------|-----------|
| ERM       | False    | initial     | 1.0 | 0.869 ± 0.003 | 0.941 ± 0.001 | 0.382 ± 0.012 |
|           |          | learned     | 5.0 | 0.874 ± 0.001 | 0.941 ± 0.002 | 0.375 ± 0.008 |
|           | True     | initial     | 1.0 | 0.864 ± 0.004 | 0.942 ± 0.001 | 0.359 ± 0.009 |
|           |          | learned     | 5.0 | 0.864 ± 0.000 | 0.940 ± 0.002 | 0.366 ± 0.006 |
| MCDropout | False    | initial     | 1.0 | 0.858 ± 0.007 | 0.943 ± 0.001 | 0.354 ± 0.022 |
|           |          | learned     | 5.0 | 0.872 ± 0.002 | 0.941 ± 0.002 | 0.371 ± 0.007 |
|           | True     | initial     | 1.0 | 0.860 ± 0.008 | 0.943 ± 0.001 | 0.358 ± 0.023 |
|           |          | learned     | 5.0 | 0.862 ± 0.001 | 0.940 ± 0.002 | 0.361 ± 0.009 |
| MIMO      | False    | initial     | 1.0 | 0.859 ± 0.004 | 0.942 ± 0.003 | 0.357 ± 0.005 |
|           |          | learned     | 5.0 | 0.872 ± 0.002 | 0.940 ± 0.001 | 0.373 ± 0.008 |
|           | True     | initial     | 1.0 | 0.861 ± 0.004 | 0.943 ± 0.003 | 0.359 ± 0.006 |
|           |          | learned     | 5.0 | 0.862 ± 0.001 | 0.940 ± 0.001 | 0.366 ± 0.003 |
| Mixup     | False    | initial     | 1.0 | 0.871 ± 0.002 | 0.942 ± 0.001 | 0.403 ± 0.013 |
|           |          | learned     | 5.0 | 0.858 ± 0.003 | 0.939 ± 0.002 | 0.336 ± 0.015 |
|           | True     | initial     | 1.0 | 0.869 ± 0.007 | 0.942 ± 0.001 | 0.400 ± 0.003 |
|           |          | learned     | 5.0 | 0.846 ± 0.003 | 0.938 ± 0.002 | 0.333 ± 0.014 |
| OC        | False    | initial     | 1.0 | 0.859 ± 0.006 | 0.942 ± 0.003 | 0.359 ± 0.020 |
|           |          | learned     | 5.0 | 0.872 ± 0.001 | 0.939 ± 0.003 | 0.374 ± 0.009 |
|           | True     | initial     | 1.0 | 0.845 ± 0.005 | 0.942 ± 0.002 | 0.341 ± 0.011 |
|           |          | learned     | 5.0 | 0.855 ± 0.002 | 0.941 ± 0.002 | 0.354 ± 0.008 |
| RND       | False    | initial     | 1.0 | 0.860 ± 0.008 | 0.939 ± 0.002 | 0.364 ± 0.018 |
|           |          | learned     | 5.0 | 0.870 ± 0.001 | 0.938 ± 0.003 | 0.350 ± 0.007 |
|           | True     | initial     | 1.0 | 0.839 ± 0.009 | 0.941 ± 0.002 | 0.332 ± 0.026 |
|           |          | learned     | 5.0 | 0.852 ± 0.003 | 0.938 ± 0.004 | 0.351 ± 0.011 |
| SoftLabeler | False  | initial    | 1.0 | 0.859 ± 0.003 | 0.941 ± 0.002 | 0.365 ± 0.003 |
|           |          | learned     | 5.0 | 0.871 ± 0.001 | 0.940 ± 0.002 | 0.371 ± 0.013 |
|           | True     | initial     | 1.0 | 0.861 ± 0.003 | 0.941 ± 0.002 | 0.359 ± 0.014 |
|           |          | learned     | 5.0 | 0.861 ± 0.002 | 0.940 ± 0.002 | 0.363 ± 0.015 |

Table 4: Out-domain results for measure Entropy and backbone ResNet50.
References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Abbas Khosravi, U Rajendra Acharya, Vladimir Makarenkov, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *arXiv*, 2020.

David Alvarez-Melis and Tommi S Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv*, 2017.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv*, 2020.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 2019.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv*, 2018.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *arXiv preprint arXiv:1809.11165*, 2018.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv*, 2014.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv*, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *ICLR*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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 *CVPR*, 2019.

Mikael Henaff, Alfredo Canziani, and Yann LeCun. Model-predictive policy learning with uncertainty regularization for driving in dense traffic. *arXiv*, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv*, 2019.
What classifiers know what they don’t?

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv*, 2016.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 2015.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv*, 2020.

Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv*, 2018.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv*, 2021.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv*, 2018.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv*, 2019.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 1999.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv*, 2020.

Burr Settles. Active learning literature survey. 2009.

Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 1948.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.
Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *arXiv*, 2018.

Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *arXiv*, 2019.

Dennis Ulmer and Giovanni Cianà. Know your limits: Monotonicity & softmax make neural classifiers overconfident on ood data. *arXiv*, 2020.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. Improving deterministic uncertainty estimation in deep learning for classification and regression. *arXiv*, 2021.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *NeurIPS*, 1992.

Alexandre MJ-C Wadoux. Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma*, 2019.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 1963.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv*, 2017.
## A In-domain results

| algorithm | spectral | calibration | k       | ACC@1  | ACC@5  | ECE    | NLL     |
|-----------|----------|-------------|---------|--------|--------|--------|---------|
| ERM       | False    | initial     | 1.0     | 0.718 ± 0.001 | 0.821 ± 0.001 | 0.044 ± 0.001 | 1.139 ± 0.002 |
|           |          | learned     | 5.0     | 0.745 ± 0.001 | 0.843 ± 0.002 | 0.044 ± 0.001 | 1.017 ± 0.001 |
|           | True     | initial     | 5.0     | 0.745 ± 0.001 | 0.843 ± 0.002 | 0.049 ± 0.003 | 1.048 ± 0.004 |
|           |          | learned     | 1.0     | 0.720 ± 0.001 | 0.822 ± 0.001 | 0.046 ± 0.004 | 1.144 ± 0.004 |
|           |          | learned     | 5.0     | 0.744 ± 0.001 | 0.843 ± 0.001 | 0.045 ± 0.000 | 1.020 ± 0.003 |
| MCDropout | False    | initial     | 1.0     | 0.720 ± 0.002 | 0.825 ± 0.003 | 0.051 ± 0.003 | 1.142 ± 0.007 |
|           |          | learned     | 5.0     | 0.747 ± 0.001 | 0.843 ± 0.001 | 0.045 ± 0.001 | 1.017 ± 0.004 |
|           | True     | initial     | 5.0     | 0.747 ± 0.001 | 0.843 ± 0.001 | 0.049 ± 0.004 | 1.050 ± 0.007 |
|           |          | learned     | 1.0     | 0.719 ± 0.003 | 0.822 ± 0.004 | 0.047 ± 0.004 | 1.149 ± 0.013 |
|           |          | learned     | 5.0     | 0.746 ± 0.001 | 0.842 ± 0.001 | 0.047 ± 0.002 | 1.020 ± 0.005 |
| MIMO      | False    | initial     | 1.0     | 0.722 ± 0.002 | 0.825 ± 0.000 | 0.046 ± 0.002 | 1.123 ± 0.005 |
|           |          | learned     | 5.0     | 0.723 ± 0.002 | 0.823 ± 0.003 | 0.090 ± 0.002 | 1.161 ± 0.005 |
|           | True     | initial     | 5.0     | 0.723 ± 0.002 | 0.823 ± 0.003 | 0.050 ± 0.004 | 1.127 ± 0.007 |
|           |          | learned     | 1.0     | 0.722 ± 0.003 | 0.824 ± 0.001 | 0.047 ± 0.001 | 1.164 ± 0.005 |
| Mixup     | False    | initial     | 1.0     | 0.723 ± 0.002 | 0.824 ± 0.004 | 0.033 ± 0.001 | 1.120 ± 0.007 |
|           |          | learned     | 5.0     | 0.739 ± 0.003 | 0.838 ± 0.001 | 0.083 ± 0.002 | 1.082 ± 0.001 |
|           | True     | initial     | 5.0     | 0.739 ± 0.003 | 0.838 ± 0.001 | 0.075 ± 0.002 | 1.246 ± 0.006 |
|           |          | learned     | 1.0     | 0.703 ± 0.002 | 0.809 ± 0.002 | 0.105 ± 0.009 | 1.193 ± 0.030 |
| OC        | False    | initial     | 1.0     | 0.720 ± 0.001 | 0.823 ± 0.001 | 0.052 ± 0.002 | 1.147 ± 0.006 |
|           |          | learned     | 5.0     | 0.746 ± 0.002 | 0.824 ± 0.002 | 0.049 ± 0.002 | 1.051 ± 0.005 |
|           | True     | initial     | 5.0     | 0.721 ± 0.002 | 0.823 ± 0.003 | 0.048 ± 0.003 | 1.141 ± 0.003 |
|           |          | learned     | 1.0     | 0.746 ± 0.001 | 0.834 ± 0.001 | 0.045 ± 0.003 | 1.016 ± 0.002 |
| RND       | False    | initial     | 1.0     | 0.720 ± 0.001 | 0.823 ± 0.001 | 0.046 ± 0.002 | 1.141 ± 0.006 |
|           |          | learned     | 5.0     | 0.745 ± 0.001 | 0.842 ± 0.001 | 0.046 ± 0.002 | 1.019 ± 0.001 |
|           | True     | initial     | 5.0     | 0.720 ± 0.001 | 0.821 ± 0.002 | 0.050 ± 0.002 | 1.152 ± 0.007 |
|           |          | learned     | 1.0     | 0.745 ± 0.001 | 0.842 ± 0.001 | 0.048 ± 0.004 | 1.049 ± 0.006 |
| SoftLabeler | False  | initial     | 1.0     | 0.719 ± 0.003 | 0.824 ± 0.001 | 0.033 ± 0.002 | 1.180 ± 0.005 |
|           |          | learned     | 5.0     | 0.743 ± 0.002 | 0.842 ± 0.002 | 0.137 ± 0.001 | 1.179 ± 0.003 |
|           | True     | initial     | 5.0     | 0.744 ± 0.003 | 0.843 ± 0.001 | 0.098 ± 0.004 | 1.214 ± 0.006 |
|           |          | learned     | 1.0     | 0.720 ± 0.003 | 0.824 ± 0.001 | 0.102 ± 0.003 | 1.128 ± 0.008 |

Table 5: In-domain results for backbone ResNet18.
# Out-domain results

| algorithm | spectral | calibration | k  | AUC               | InAsIn     | OutAsOut  |
|-----------|----------|-------------|----|------------------|------------|-----------|
| ERM       | False    | initial     | 1.0 | 0.644 ± 0.017    | 0.949 ± 0.001 | 0.093 ± 0.016 |
|           | True     | learned     | 5.0 | 0.660 ± 0.011    | 0.956 ± 0.030 | 0.076 ± 0.066 |
|           |          | initial     | 1.0 | 0.649 ± 0.017    | 0.952 ± 0.001 | 0.090 ± 0.017 |
|           |          | learned     | 5.0 | 0.658 ± 0.009    | 0.949 ± 0.001 | 0.105 ± 0.011 |
|           |          | initial     | 1.0 | 0.648 ± 0.010    | 0.949 ± 0.000 | 0.097 ± 0.004 |
|           |          | learned     | 5.0 | 0.664 ± 0.005    | 0.948 ± 0.003 | 0.105 ± 0.006 |
|           |          | initial     | 1.0 | 0.649 ± 0.010    | 0.950 ± 0.001 | 0.097 ± 0.003 |
|           |          | learned     | 5.0 | 0.657 ± 0.006    | 0.951 ± 0.002 | 0.100 ± 0.004 |
| MCDropout | False    | initial     | 1.0 | 0.658 ± 0.008    | 0.965 ± 0.030 | 0.080 ± 0.070 |
|           | True     | learned     | 5.0 | 0.664 ± 0.010    | 0.966 ± 0.002 | 0.101 ± 0.008 |
|           |          | initial     | 1.0 | 0.655 ± 0.024    | 0.953 ± 0.003 | 0.095 ± 0.015 |
|           |          | learned     | 5.0 | 0.673 ± 0.006    | 0.949 ± 0.003 | 0.107 ± 0.007 |
|           |          | initial     | 1.0 | 0.656 ± 0.027    | 0.951 ± 0.003 | 0.093 ± 0.014 |
|           |          | learned     | 5.0 | 0.668 ± 0.007    | 0.951 ± 0.000 | 0.105 ± 0.003 |
| MIMO      | False    | initial     | 1.0 | 0.660 ± 0.034    | 0.950 ± 0.001 | 0.101 ± 0.026 |
|           | True     | learned     | 5.0 | 0.660 ± 0.012    | 0.949 ± 0.000 | 0.112 ± 0.010 |
|           |          | initial     | 1.0 | 0.655 ± 0.016    | 0.967 ± 0.029 | 0.164 ± 0.058 |
|           |          | learned     | 5.0 | 0.676 ± 0.004    | 0.948 ± 0.003 | 0.117 ± 0.002 |
|           |          | initial     | 1.0 | 0.652 ± 0.015    | 0.947 ± 0.001 | 0.101 ± 0.014 |
|           |          | learned     | 5.0 | 0.666 ± 0.003    | 0.948 ± 0.001 | 0.114 ± 0.005 |
| Mixup     | False    | initial     | 1.0 | 0.683 ± 0.007    | 0.966 ± 0.029 | 0.080 ± 0.069 |
|           | True     | learned     | 5.0 | 0.669 ± 0.020    | 0.951 ± 0.002 | 0.136 ± 0.019 |
|           |          | initial     | 1.0 | 0.676 ± 0.009    | 0.949 ± 0.001 | 0.122 ± 0.007 |
|           |          | learned     | 5.0 | 0.688 ± 0.011    | 0.951 ± 0.003 | 0.119 ± 0.009 |
|           |          | initial     | 1.0 | 0.689 ± 0.010    | 0.947 ± 0.002 | 0.126 ± 0.014 |
|           |          | learned     | 5.0 | 0.648 ± 0.021    | 0.950 ± 0.003 | 0.114 ± 0.022 |
|           |          | initial     | 1.0 | 0.679 ± 0.006    | 0.948 ± 0.003 | 0.121 ± 0.010 |
| OC        | False    | initial     | 1.0 | 0.641 ± 0.017    | 0.949 ± 0.003 | 0.092 ± 0.012 |
|           | True     | learned     | 5.0 | 0.667 ± 0.003    | 0.950 ± 0.001 | 0.104 ± 0.003 |
|           |          | initial     | 1.0 | 0.640 ± 0.017    | 0.950 ± 0.003 | 0.089 ± 0.013 |
|           |          | learned     | 5.0 | 0.662 ± 0.000    | 0.950 ± 0.003 | 0.106 ± 0.005 |
|           |          | initial     | 1.0 | 0.649 ± 0.030    | 0.947 ± 0.003 | 0.099 ± 0.011 |
|           |          | learned     | 5.0 | 0.670 ± 0.011    | 0.950 ± 0.001 | 0.106 ± 0.009 |
|           |          | initial     | 1.0 | 0.653 ± 0.029    | 0.950 ± 0.002 | 0.100 ± 0.012 |
|           |          | learned     | 5.0 | 0.659 ± 0.013    | 0.949 ± 0.002 | 0.106 ± 0.006 |
|           |          | initial     | 1.0 | 0.650 ± 0.030    | 0.951 ± 0.002 | 0.094 ± 0.023 |
|           |          | learned     | 5.0 | 0.671 ± 0.012    | 0.950 ± 0.003 | 0.107 ± 0.012 |
|           |          | initial     | 1.0 | 0.655 ± 0.031    | 0.967 ± 0.029 | 0.074 ± 0.066 |
|           |          | learned     | 5.0 | 0.661 ± 0.011    | 0.950 ± 0.000 | 0.108 ± 0.007 |
|           |          | initial     | 1.0 | 0.664 ± 0.007    | 0.948 ± 0.002 | 0.102 ± 0.006 |
|           |          | learned     | 5.0 | 0.645 ± 0.017    | 0.965 ± 0.031 | 0.062 ± 0.054 |
|           |          | initial     | 1.0 | 0.650 ± 0.007    | 0.948 ± 0.001 | 0.101 ± 0.004 |
|           |          | learned     | 5.0 | 0.630 ± 0.020    | 0.951 ± 0.003 | 0.089 ± 0.015 |
| SoftLabeler| False   | initial     | 1.0 | 0.644 ± 0.006    | 0.968 ± 0.028 | 0.062 ± 0.054 |
|           | True     | learned     | 5.0 | 0.646 ± 0.009    | 0.950 ± 0.001 | 0.097 ± 0.007 |
|           |          | initial     | 1.0 | 0.642 ± 0.013    | 0.950 ± 0.004 | 0.089 ± 0.005 |
|           |          | learned     | 5.0 | 0.656 ± 0.010    | 0.947 ± 0.002 | 0.103 ± 0.009 |

Table 6: Out-domain results for measure Augmentations and backbone ResNet18.
| algorithm | spectral | calibration | k | AUC   | InAsIn | OutAsOut |
|-----------|----------|-------------|---|-------|--------|----------|
| ERM       | False    | initial     | 1.0 | 0.663 ± 0.004 | 0.951 ± 0.001 | 0.106 ± 0.010 |
|           |          | learned     | 5.0 | 0.682 ± 0.004 | 0.966 ± 0.030 | 0.079 ± 0.068 |
|           | True     | initial     | 1.0 | 0.657 ± 0.010 | 0.951 ± 0.002 | 0.105 ± 0.017 |
|           |          | learned     | 5.0 | 0.662 ± 0.004 | 0.948 ± 0.002 | 0.106 ± 0.003 |
|           |          | initial     | 1.0 | 0.655 ± 0.002 | 0.951 ± 0.001 | 0.105 ± 0.006 |
|           |          | learned     | 5.0 | 0.662 ± 0.007 | 0.948 ± 0.002 | 0.101 ± 0.011 |
|           |          | initial     | 1.0 | 0.657 ± 0.001 | 0.947 ± 0.003 | 0.113 ± 0.009 |
|           |          | learned     | 5.0 | 0.652 ± 0.007 | 0.948 ± 0.002 | 0.099 ± 0.010 |
| MCDropout | False    | initial     | 1.0 | 0.635 ± 0.042 | 0.950 ± 0.002 | 0.089 ± 0.023 |
|           |          | learned     | 5.0 | 0.674 ± 0.010 | 0.967 ± 0.029 | 0.066 ± 0.057 |
|           | True     | initial     | 1.0 | 0.638 ± 0.040 | 0.951 ± 0.001 | 0.089 ± 0.017 |
|           |          | learned     | 5.0 | 0.654 ± 0.011 | 0.949 ± 0.001 | 0.096 ± 0.004 |
|           |          | initial     | 1.0 | 0.652 ± 0.020 | 0.949 ± 0.002 | 0.102 ± 0.009 |
|           |          | learned     | 5.0 | 0.666 ± 0.013 | 0.947 ± 0.001 | 0.106 ± 0.010 |
|           |          | initial     | 1.0 | 0.656 ± 0.021 | 0.949 ± 0.001 | 0.107 ± 0.012 |
|           |          | learned     | 5.0 | 0.659 ± 0.013 | 0.949 ± 0.001 | 0.103 ± 0.012 |
| MIMO      | False    | initial     | 1.0 | 0.664 ± 0.016 | 0.950 ± 0.001 | 0.110 ± 0.015 |
|           |          | learned     | 5.0 | 0.700 ± 0.006 | 0.949 ± 0.001 | 0.121 ± 0.005 |
|           | True     | initial     | 1.0 | 0.665 ± 0.016 | 0.950 ± 0.001 | 0.113 ± 0.013 |
|           |          | learned     | 5.0 | 0.679 ± 0.006 | 0.950 ± 0.001 | 0.114 ± 0.005 |
|           |          | initial     | 1.0 | 0.668 ± 0.020 | 0.950 ± 0.004 | 0.107 ± 0.017 |
|           |          | learned     | 5.0 | 0.686 ± 0.018 | 0.949 ± 0.002 | 0.111 ± 0.013 |
| Mixup     | False    | initial     | 1.0 | 0.643 ± 0.027 | 0.967 ± 0.029 | 0.061 ± 0.054 |
|           |          | learned     | 5.0 | 0.671 ± 0.015 | 0.946 ± 0.001 | 0.109 ± 0.010 |
| OC        | False    | initial     | 1.0 | 0.672 ± 0.021 | 0.945 ± 0.004 | 0.115 ± 0.012 |
|           |          | learned     | 5.0 | 0.700 ± 0.012 | 0.965 ± 0.031 | 0.084 ± 0.073 |
|           | True     | initial     | 1.0 | 0.700 ± 0.027 | 0.943 ± 0.001 | 0.151 ± 0.031 |
|           |          | learned     | 5.0 | 0.682 ± 0.010 | 0.948 ± 0.000 | 0.118 ± 0.007 |
| RND       | False    | initial     | 1.0 | 0.645 ± 0.025 | 0.951 ± 0.001 | 0.096 ± 0.014 |
|           |          | learned     | 5.0 | 0.664 ± 0.022 | 0.949 ± 0.002 | 0.102 ± 0.012 |
|           | True     | initial     | 1.0 | 0.651 ± 0.025 | 0.949 ± 0.002 | 0.104 ± 0.015 |
|           |          | learned     | 5.0 | 0.654 ± 0.017 | 0.948 ± 0.004 | 0.101 ± 0.012 |
| SoftLabeler | False  | initial    | 1.0 | 0.653 ± 0.021 | 0.966 ± 0.030 | 0.064 ± 0.057 |
|            |          | learned    | 5.0 | 0.675 ± 0.006 | 0.966 ± 0.030 | 0.074 ± 0.064 |
|            | True     | initial    | 1.0 | 0.661 ± 0.020 | 0.949 ± 0.001 | 0.117 ± 0.026 |
|            |          | learned    | 5.0 | 0.664 ± 0.005 | 0.949 ± 0.002 | 0.110 ± 0.010 |
|            |          | initial    | 1.0 | 0.655 ± 0.012 | 0.948 ± 0.001 | 0.104 ± 0.009 |
|            |          | learned    | 5.0 | 0.670 ± 0.013 | 0.948 ± 0.001 | 0.109 ± 0.006 |
|            |          | initial    | 1.0 | 0.655 ± 0.013 | 0.951 ± 0.002 | 0.102 ± 0.008 |
|            |          | learned    | 5.0 | 0.660 ± 0.013 | 0.949 ± 0.002 | 0.102 ± 0.004 |
|            |          | initial    | 1.0 | 0.630 ± 0.007 | 0.954 ± 0.003 | 0.083 ± 0.008 |
|            |          | learned    | 5.0 | 0.649 ± 0.007 | 0.963 ± 0.032 | 0.066 ± 0.057 |
|            | True     | initial    | 1.0 | 0.625 ± 0.002 | 0.950 ± 0.005 | 0.089 ± 0.009 |
|            |          | learned    | 5.0 | 0.647 ± 0.007 | 0.946 ± 0.001 | 0.096 ± 0.004 |
|            |          | initial    | 1.0 | 0.635 ± 0.044 | 0.950 ± 0.002 | 0.086 ± 0.021 |
|            |          | learned    | 5.0 | 0.644 ± 0.020 | 0.948 ± 0.002 | 0.087 ± 0.014 |
|            |          | initial    | 1.0 | 0.624 ± 0.043 | 0.951 ± 0.002 | 0.086 ± 0.021 |
|            |          | learned    | 5.0 | 0.632 ± 0.018 | 0.982 ± 0.031 | 0.034 ± 0.059 |

Table 7: Out-domain results for measure Augmentations and backbone ResNet50.
| Algorithm | Spectral | Calibration | k  | AUC       | InAsIn    | OutAsOut  |
|-----------|----------|-------------|----|-----------|-----------|-----------|
| ERM       | False    | initial     | 1.0| 0.811 ± 0.009 | 0.941 ± 0.001 | 0.258 ± 0.011 |
|           |          | learned     | 1.0| 0.810 ± 0.009 | 0.941 ± 0.001 | 0.257 ± 0.011 |
|           | True     | initial     | 1.0| 0.809 ± 0.006 | 0.941 ± 0.001 | 0.244 ± 0.008 |
|           |          | learned     | 1.0| 0.808 ± 0.005 | 0.941 ± 0.001 | 0.245 ± 0.009 |
|           |          | initial     | 1.0| 0.805 ± 0.008 | 0.941 ± 0.002 | 0.250 ± 0.014 |
|           |          | learned     | 1.0| 0.811 ± 0.006 | 0.939 ± 0.001 | 0.260 ± 0.011 |
|           |          | initial     | 1.0| 0.810 ± 0.004 | 0.940 ± 0.004 | 0.251 ± 0.008 |
|           |          | learned     | 1.0| 0.833 ± 0.004 | 0.940 ± 0.002 | 0.273 ± 0.007 |
|           |          | initial     | 1.0| 0.816 ± 0.013 | 0.939 ± 0.002 | 0.259 ± 0.023 |
|           |          | learned     | 1.0| 0.822 ± 0.004 | 0.939 ± 0.001 | 0.269 ± 0.015 |
|           |          | initial     | 1.0| 0.814 ± 0.003 | 0.939 ± 0.001 | 0.250 ± 0.012 |
|           |          | learned     | 1.0| 0.813 ± 0.009 | 0.940 ± 0.003 | 0.263 ± 0.017 |
|           |          | initial     | 1.0| 0.822 ± 0.006 | 0.940 ± 0.002 | 0.252 ± 0.007 |
|           |          | learned     | 1.0| 0.813 ± 0.005 | 0.940 ± 0.001 | 0.276 ± 0.010 |
|           |          | initial     | 1.0| 0.806 ± 0.011 | 0.941 ± 0.000 | 0.243 ± 0.013 |
|           |          | learned     | 1.0| 0.820 ± 0.008 | 0.938 ± 0.001 | 0.252 ± 0.011 |
|           |          | initial     | 1.0| 0.783 ± 0.010 | 0.938 ± 0.002 | 0.234 ± 0.004 |
|           |          | learned     | 1.0| 0.803 ± 0.004 | 0.938 ± 0.002 | 0.253 ± 0.009 |
|           |          | initial     | 1.0| 0.808 ± 0.005 | 0.941 ± 0.002 | 0.249 ± 0.009 |
|           |          | learned     | 1.0| 0.820 ± 0.004 | 0.939 ± 0.002 | 0.253 ± 0.006 |
|           |          | initial     | 1.0| 0.811 ± 0.003 | 0.940 ± 0.001 | 0.257 ± 0.009 |
|           |          | learned     | 1.0| 0.830 ± 0.005 | 0.939 ± 0.001 | 0.268 ± 0.009 |
|           |          | initial     | 1.0| 0.810 ± 0.003 | 0.941 ± 0.001 | 0.257 ± 0.009 |
|           |          | learned     | 1.0| 0.821 ± 0.004 | 0.938 ± 0.001 | 0.274 ± 0.010 |
|           |          | initial     | 1.0| 0.808 ± 0.003 | 0.941 ± 0.002 | 0.253 ± 0.005 |
|           |          | learned     | 1.0| 0.830 ± 0.001 | 0.939 ± 0.002 | 0.270 ± 0.010 |
|           |          | initial     | 1.0| 0.807 ± 0.003 | 0.941 ± 0.003 | 0.254 ± 0.005 |
|           |          | learned     | 1.0| 0.822 ± 0.001 | 0.938 ± 0.003 | 0.272 ± 0.011 |
|           |          | initial     | 1.0| 0.813 ± 0.002 | 0.940 ± 0.000 | 0.248 ± 0.011 |
|           |          | learned     | 1.0| 0.832 ± 0.001 | 0.941 ± 0.002 | 0.265 ± 0.012 |
|           |          | initial     | 1.0| 0.812 ± 0.002 | 0.940 ± 0.000 | 0.248 ± 0.011 |
|           |          | learned     | 1.0| 0.824 ± 0.001 | 0.941 ± 0.001 | 0.267 ± 0.008 |
|           |          | initial     | 1.0| 0.817 ± 0.006 | 0.939 ± 0.002 | 0.267 ± 0.017 |
|           |          | learned     | 1.0| 0.832 ± 0.003 | 0.938 ± 0.003 | 0.271 ± 0.012 |
|           |          | initial     | 1.0| 0.813 ± 0.004 | 0.939 ± 0.001 | 0.254 ± 0.007 |
|           |          | learned     | 1.0| 0.824 ± 0.001 | 0.938 ± 0.002 | 0.275 ± 0.007 |
|           |          | initial     | 1.0| 0.786 ± 0.021 | 0.941 ± 0.001 | 0.219 ± 0.022 |
|           |          | learned     | 1.0| 0.802 ± 0.004 | 0.939 ± 0.002 | 0.236 ± 0.007 |
|           |          | initial     | 1.0| 0.789 ± 0.019 | 0.940 ± 0.002 | 0.231 ± 0.018 |
|           |          | learned     | 1.0| 0.799 ± 0.002 | 0.938 ± 0.001 | 0.249 ± 0.003 |
|           |          | initial     | 1.0| 0.787 ± 0.007 | 0.941 ± 0.001 | 0.212 ± 0.015 |
|           |          | learned     | 1.0| 0.809 ± 0.008 | 0.940 ± 0.001 | 0.235 ± 0.011 |
|           |          | initial     | 1.0| 0.790 ± 0.003 | 0.940 ± 0.002 | 0.224 ± 0.011 |
|           |          | learned     | 1.0| 0.805 ± 0.004 | 0.939 ± 0.001 | 0.247 ± 0.002 |

Table 8: Out-domain results for measure Entropy and backbone ResNet18.
| algorithm | spectral | calibration | k | AUC       | InAsIn    | OutAsOut    |
|-----------|----------|-------------|---|-----------|-----------|-------------|
| ERM       | False    | initial     | 1.0 | 0.775 ± 0.007 | 0.944 ± 0.002 | 0.177 ± 0.010 |
|           |         | learned     | 5.0 | 0.795 ± 0.004 | 0.942 ± 0.001 | 0.192 ± 0.008 |
|           | True     | initial     | 1.0 | 0.774 ± 0.007 | 0.944 ± 0.002 | 0.176 ± 0.010 |
|           |         | learned     | 5.0 | 0.785 ± 0.004 | 0.942 ± 0.002 | 0.178 ± 0.008 |
| MCDropout | False    | initial     | 1.0 | 0.774 ± 0.007 | 0.944 ± 0.001 | 0.172 ± 0.007 |
|           |         | learned     | 5.0 | 0.796 ± 0.003 | 0.944 ± 0.002 | 0.183 ± 0.010 |
|           | True     | initial     | 1.0 | 0.777 ± 0.004 | 0.945 ± 0.002 | 0.171 ± 0.008 |
|           |         | learned     | 5.0 | 0.786 ± 0.003 | 0.944 ± 0.001 | 0.171 ± 0.009 |
| MIMO      | False    | initial     | 1.0 | 0.783 ± 0.007 | 0.944 ± 0.001 | 0.179 ± 0.010 |
|           |         | learned     | 5.0 | 0.788 ± 0.002 | 0.943 ± 0.001 | 0.179 ± 0.004 |
|           | True     | initial     | 1.0 | 0.782 ± 0.007 | 0.944 ± 0.001 | 0.178 ± 0.010 |
|           |         | learned     | 5.0 | 0.775 ± 0.002 | 0.944 ± 0.001 | 0.166 ± 0.004 |
| Mixup     | False    | initial     | 1.0 | 0.783 ± 0.007 | 0.947 ± 0.001 | 0.172 ± 0.009 |
|           |         | learned     | 5.0 | 0.790 ± 0.004 | 0.942 ± 0.002 | 0.185 ± 0.009 |
|           | True     | initial     | 1.0 | 0.750 ± 0.009 | 0.944 ± 0.003 | 0.151 ± 0.007 |
|           |         | learned     | 5.0 | 0.772 ± 0.001 | 0.943 ± 0.002 | 0.159 ± 0.004 |
| OC        | False    | initial     | 1.0 | 0.778 ± 0.003 | 0.946 ± 0.002 | 0.174 ± 0.003 |
|           |         | learned     | 5.0 | 0.794 ± 0.003 | 0.941 ± 0.001 | 0.184 ± 0.001 |
|           | True     | initial     | 1.0 | 0.777 ± 0.003 | 0.946 ± 0.002 | 0.174 ± 0.004 |
|           |         | learned     | 5.0 | 0.784 ± 0.002 | 0.943 ± 0.001 | 0.172 ± 0.002 |
| RND       | False    | initial     | 1.0 | 0.780 ± 0.005 | 0.944 ± 0.002 | 0.174 ± 0.004 |
|           |         | learned     | 5.0 | 0.797 ± 0.002 | 0.942 ± 0.001 | 0.188 ± 0.003 |
|           | True     | initial     | 1.0 | 0.780 ± 0.005 | 0.944 ± 0.002 | 0.173 ± 0.004 |
|           |         | learned     | 5.0 | 0.788 ± 0.002 | 0.943 ± 0.002 | 0.174 ± 0.005 |
| SoftLabeler | False   | initial    | 1.0 | 0.771 ± 0.018 | 0.944 ± 0.002 | 0.175 ± 0.012 |
|           |         | learned     | 5.0 | 0.786 ± 0.000 | 0.942 ± 0.003 | 0.183 ± 0.004 |
|           | True     | initial    | 1.0 | 0.764 ± 0.017 | 0.944 ± 0.001 | 0.165 ± 0.012 |
|           |         | learned     | 5.0 | 0.771 ± 0.003 | 0.944 ± 0.004 | 0.162 ± 0.003 |

Table 9: Out-domain results for measure Gap and backbone ResNet18.
| algorithm | spectral | calibration | k | AUC         | InAsIn     | OutAsOut    |
|-----------|----------|-------------|---|-------------|------------|-------------|
| ERM       | False    | initial     | 1.0 | 0.822 ± 0.001 | 0.945 ± 0.002 | 0.223 ± 0.000 |
|           |          | learned     | 5.0 | 0.829 ± 0.002 | 0.944 ± 0.001 | 0.225 ± 0.005 |
|           | True     | initial     | 1.0 | 0.819 ± 0.003 | 0.945 ± 0.001 | 0.216 ± 0.006 |
|           |          | learned     | 5.0 | 0.820 ± 0.003 | 0.944 ± 0.002 | 0.209 ± 0.007 |
| MCDropout | False    | initial     | 1.0 | 0.815 ± 0.004 | 0.944 ± 0.001 | 0.214 ± 0.005 |
|           |          | learned     | 5.0 | 0.827 ± 0.002 | 0.942 ± 0.003 | 0.223 ± 0.005 |
|           | True     | initial     | 1.0 | 0.817 ± 0.005 | 0.944 ± 0.001 | 0.218 ± 0.006 |
|           |          | learned     | 5.0 | 0.818 ± 0.003 | 0.943 ± 0.003 | 0.207 ± 0.004 |
| MIMO      | False    | initial     | 1.0 | 0.817 ± 0.004 | 0.945 ± 0.003 | 0.216 ± 0.003 |
|           |          | learned     | 5.0 | 0.827 ± 0.003 | 0.943 ± 0.002 | 0.219 ± 0.008 |
|           | True     | initial     | 1.0 | 0.818 ± 0.005 | 0.945 ± 0.003 | 0.221 ± 0.003 |
|           |          | learned     | 5.0 | 0.818 ± 0.004 | 0.943 ± 0.002 | 0.208 ± 0.009 |
| Mixup     | False    | initial     | 1.0 | 0.822 ± 0.004 | 0.944 ± 0.001 | 0.229 ± 0.010 |
|           |          | learned     | 5.0 | 0.822 ± 0.006 | 0.944 ± 0.001 | 0.226 ± 0.006 |
|           | True     | initial     | 1.0 | 0.820 ± 0.002 | 0.944 ± 0.002 | 0.223 ± 0.006 |
|           |          | learned     | 5.0 | 0.819 ± 0.003 | 0.942 ± 0.004 | 0.217 ± 0.007 |
| OC        | False    | initial     | 1.0 | 0.819 ± 0.004 | 0.944 ± 0.003 | 0.221 ± 0.010 |
|           |          | learned     | 5.0 | 0.828 ± 0.002 | 0.943 ± 0.002 | 0.230 ± 0.005 |
|           | True     | initial     | 1.0 | 0.802 ± 0.007 | 0.946 ± 0.003 | 0.194 ± 0.011 |
|           |          | learned     | 5.0 | 0.811 ± 0.003 | 0.944 ± 0.002 | 0.201 ± 0.007 |
| RND       | False    | initial     | 1.0 | 0.818 ± 0.003 | 0.947 ± 0.001 | 0.219 ± 0.003 |
|           |          | learned     | 5.0 | 0.830 ± 0.001 | 0.943 ± 0.003 | 0.231 ± 0.008 |
|           | True     | initial     | 1.0 | 0.808 ± 0.006 | 0.945 ± 0.001 | 0.202 ± 0.007 |
|           |          | learned     | 5.0 | 0.813 ± 0.001 | 0.946 ± 0.003 | 0.201 ± 0.009 |
| SoftLabeler | False  | initial    | 1.0 | 0.814 ± 0.002 | 0.944 ± 0.002 | 0.216 ± 0.007 |
|           |          | learned     | 5.0 | 0.826 ± 0.001 | 0.942 ± 0.001 | 0.225 ± 0.003 |
|           | True     | initial     | 1.0 | 0.816 ± 0.002 | 0.944 ± 0.002 | 0.220 ± 0.008 |
|           |          | learned     | 5.0 | 0.817 ± 0.002 | 0.943 ± 0.002 | 0.208 ± 0.002 |
|           | False    | initial    | 1.0 | 0.813 ± 0.005 | 0.943 ± 0.003 | 0.214 ± 0.008 |
|           |          | learned     | 5.0 | 0.827 ± 0.003 | 0.943 ± 0.004 | 0.231 ± 0.009 |
|           | True     | initial    | 1.0 | 0.815 ± 0.005 | 0.943 ± 0.003 | 0.218 ± 0.008 |
|           |          | learned     | 5.0 | 0.818 ± 0.003 | 0.943 ± 0.003 | 0.214 ± 0.009 |

Table 10: Out-domain results for measure Gap and backbone ResNet50.
| algorithm | spectral | calibration | k  | AUC       | InAsIn      | OutAsOut    |
|-----------|----------|-------------|----|-----------|-------------|-------------|
| ERM       | False    | initial     | 1.0| 0.552 ± 0.007 | 0.956 ± 0.001 | 0.045 ± 0.002 |
|           |          | learned     | 5.0| 0.218 ± 0.012 | 0.960 ± 0.000 | 0.001 ± 0.000 |
|           | True     | initial     | 1.0| 0.552 ± 0.007 | 0.956 ± 0.001 | 0.045 ± 0.002 |
|           |          | learned     | 5.0| 0.218 ± 0.012 | 0.960 ± 0.000 | 0.001 ± 0.000 |
| MIMO      | False    | initial     | 1.0| 0.269 ± 0.027 | 0.961 ± 0.001 | 0.003 ± 0.002 |
|           |          | learned     | 5.0| 0.244 ± 0.004 | 0.961 ± 0.001 | 0.001 ± 0.000 |
|           | True     | initial     | 1.0| 0.269 ± 0.027 | 0.961 ± 0.001 | 0.003 ± 0.002 |
|           |          | learned     | 5.0| 0.244 ± 0.004 | 0.961 ± 0.001 | 0.001 ± 0.000 |
| Mixup     | False    | initial     | 1.0| 0.565 ± 0.005 | 0.954 ± 0.002 | 0.056 ± 0.002 |
|           |          | learned     | 5.0| 0.259 ± 0.013 | 0.960 ± 0.001 | 0.002 ± 0.000 |
|           | True     | initial     | 1.0| 0.539 ± 0.006 | 0.955 ± 0.002 | 0.041 ± 0.006 |
|           |          | learned     | 5.0| 0.259 ± 0.013 | 0.960 ± 0.001 | 0.002 ± 0.000 |
| OC        | False    | initial     | 1.0| 0.569 ± 0.016 | 0.956 ± 0.001 | 0.051 ± 0.006 |
|           |          | learned     | 5.0| 0.216 ± 0.008 | 0.959 ± 0.000 | 0.000 ± 0.000 |
|           | True     | initial     | 1.0| 0.539 ± 0.013 | 0.956 ± 0.001 | 0.044 ± 0.006 |
|           |          | learned     | 5.0| 0.218 ± 0.005 | 0.959 ± 0.002 | 0.001 ± 0.000 |
| RND       | False    | initial     | 1.0| 0.557 ± 0.017 | 0.955 ± 0.003 | 0.047 ± 0.007 |
|           |          | learned     | 5.0| 0.217 ± 0.003 | 0.959 ± 0.002 | 0.001 ± 0.000 |
|           | True     | initial     | 1.0| 0.537 ± 0.013 | 0.956 ± 0.001 | 0.041 ± 0.005 |
|           |          | learned     | 5.0| 0.220 ± 0.006 | 0.959 ± 0.002 | 0.001 ± 0.000 |
| SoftLabeler | False   | initial    | 1.0| 0.587 ± 0.005 | 0.953 ± 0.001 | 0.049 ± 0.009 |
|            |          | learned    | 5.0| 0.406 ± 0.005 | 0.958 ± 0.002 | 0.020 ± 0.001 |
|            | True     | initial    | 1.0| 0.586 ± 0.006 | 0.953 ± 0.003 | 0.047 ± 0.004 |
|            |          | learned    | 5.0| 0.394 ± 0.009 | 0.957 ± 0.002 | 0.016 ± 0.002 |

Table 11: Out-domain results for measure Jacobian and backbone ResNet18.
Table 12: Out-domain results for measure Jacobian and backbone ResNet50.

| algorithm | spectral | calibration | k   | AUC         | InAsIn     | OutAsOut   |
|-----------|----------|-------------|-----|-------------|------------|------------|
| ERM       | False    | initial     | 1.0 | 0.545 ± 0.036 | 0.951 ± 0.003 | 0.058 ± 0.013 |
|           | True     | initial     | 1.0 | 0.522 ± 0.020 | 0.950 ± 0.003 | 0.053 ± 0.004 |
|           |          | learned     | 1.0 | 0.522 ± 0.020 | 0.950 ± 0.003 | 0.053 ± 0.004 |
| MIMO      | False    | initial     | 1.0 | 0.192 ± 0.007 | 0.955 ± 0.002 | 0.003 ± 0.001 |
|           | True     | initial     | 1.0 | 0.194 ± 0.009 | 0.955 ± 0.001 | 0.003 ± 0.000 |
|           |          | learned     | 1.0 | 0.230 ± 0.009 | 0.955 ± 0.002 | 0.005 ± 0.001 |
| Mixup     | False    | initial     | 1.0 | 0.561 ± 0.054 | 0.951 ± 0.003 | 0.061 ± 0.021 |
|           | True     | initial     | 1.0 | 0.595 ± 0.029 | 0.948 ± 0.002 | 0.072 ± 0.018 |
| OC        | False    | initial     | 5.0 | 0.529 ± 0.018 | 0.951 ± 0.001 | 0.050 ± 0.005 |
|           | True     | initial     | 5.0 | 0.626 ± 0.000 | 0.958 ± 0.001 | 0.001 ± 0.000 |
| RND       | False    | initial     | 5.0 | 0.166 ± 0.001 | 0.958 ± 0.001 | 0.001 ± 0.000 |
|           |          | learned     | 5.0 | 0.543 ± 0.013 | 0.956 ± 0.003 | 0.054 ± 0.008 |
| SoftLabeler | False  | initial     | 5.0 | 0.161 ± 0.009 | 0.958 ± 0.003 | 0.001 ± 0.000 |
|           | True     | initial     | 5.0 | 0.536 ± 0.005 | 0.950 ± 0.001 | 0.045 ± 0.003 |
|           |          | learned     | 5.0 | 0.379 ± 0.011 | 0.953 ± 0.002 | 0.015 ± 0.001 |
What classifiers know what they don’t?

| algorithm | spectral calibration | k   | AUC    | InAsIn  | OutAsOut |
|-----------|----------------------|-----|--------|---------|-----------|
| ERM       | initial              | 1.0 | 0.795 ± 0.009 | 0.940 ± 0.002 | 0.245 ± 0.006 |
|           | learned              | 5.0 | 0.819 ± 0.005 | 0.937 ± 0.002 | 0.271 ± 0.009 |
|           | initial              | 1.0 | 0.794 ± 0.009 | 0.940 ± 0.002 | 0.244 ± 0.006 |
|           | learned              | 5.0 | 0.804 ± 0.005 | 0.937 ± 0.000 | 0.253 ± 0.005 |
| MCDropout | initial              | 1.0 | 0.796 ± 0.004 | 0.941 ± 0.000 | 0.235 ± 0.007 |
|           | learned              | 5.0 | 0.818 ± 0.005 | 0.937 ± 0.001 | 0.263 ± 0.011 |
|           | initial              | 1.0 | 0.795 ± 0.004 | 0.941 ± 0.000 | 0.234 ± 0.007 |
|           | learned              | 5.0 | 0.804 ± 0.005 | 0.937 ± 0.002 | 0.247 ± 0.012 |
| MIMO      | initial              | 1.0 | 0.792 ± 0.008 | 0.940 ± 0.001 | 0.239 ± 0.017 |
|           | learned              | 5.0 | 0.819 ± 0.004 | 0.939 ± 0.003 | 0.260 ± 0.017 |
|           | initial              | 1.0 | 0.796 ± 0.005 | 0.941 ± 0.001 | 0.243 ± 0.012 |
|           | learned              | 5.0 | 0.805 ± 0.004 | 0.939 ± 0.001 | 0.243 ± 0.016 |
|           | initial              | 1.0 | 0.798 ± 0.004 | 0.940 ± 0.001 | 0.244 ± 0.005 |
|           | learned              | 5.0 | 0.821 ± 0.005 | 0.939 ± 0.001 | 0.268 ± 0.006 |
|           | initial              | 1.0 | 0.797 ± 0.004 | 0.940 ± 0.001 | 0.243 ± 0.006 |
|           | learned              | 5.0 | 0.806 ± 0.005 | 0.939 ± 0.001 | 0.248 ± 0.011 |
| Mixup     | initial              | 1.0 | 0.802 ± 0.010 | 0.939 ± 0.001 | 0.251 ± 0.018 |
|           | learned              | 5.0 | 0.814 ± 0.003 | 0.939 ± 0.001 | 0.247 ± 0.012 |
|           | initial              | 1.0 | 0.801 ± 0.010 | 0.939 ± 0.001 | 0.251 ± 0.018 |
|           | learned              | 5.0 | 0.796 ± 0.002 | 0.940 ± 0.000 | 0.231 ± 0.012 |
|           | initial              | 1.0 | 0.801 ± 0.007 | 0.942 ± 0.002 | 0.244 ± 0.014 |
|           | learned              | 5.0 | 0.814 ± 0.005 | 0.939 ± 0.001 | 0.242 ± 0.011 |
|           | initial              | 1.0 | 0.800 ± 0.007 | 0.942 ± 0.002 | 0.243 ± 0.014 |
|           | learned              | 5.0 | 0.794 ± 0.005 | 0.939 ± 0.000 | 0.231 ± 0.010 |
| OC        | initial              | 1.0 | 0.799 ± 0.009 | 0.941 ± 0.002 | 0.237 ± 0.015 |
|           | learned              | 5.0 | 0.814 ± 0.006 | 0.938 ± 0.003 | 0.254 ± 0.013 |
|           | initial              | 1.0 | 0.768 ± 0.009 | 0.939 ± 0.001 | 0.216 ± 0.006 |
|           | learned              | 5.0 | 0.786 ± 0.002 | 0.940 ± 0.001 | 0.220 ± 0.007 |
|           | initial              | 1.0 | 0.798 ± 0.002 | 0.942 ± 0.002 | 0.242 ± 0.009 |
|           | learned              | 5.0 | 0.815 ± 0.002 | 0.938 ± 0.001 | 0.257 ± 0.009 |
|           | initial              | 1.0 | 0.773 ± 0.007 | 0.942 ± 0.001 | 0.213 ± 0.017 |
|           | learned              | 5.0 | 0.794 ± 0.002 | 0.940 ± 0.001 | 0.233 ± 0.005 |
| RND       | initial              | 1.0 | 0.797 ± 0.003 | 0.940 ± 0.003 | 0.246 ± 0.011 |
|           | learned              | 5.0 | 0.818 ± 0.003 | 0.939 ± 0.001 | 0.265 ± 0.009 |
|           | initial              | 1.0 | 0.796 ± 0.003 | 0.940 ± 0.003 | 0.244 ± 0.011 |
|           | learned              | 5.0 | 0.803 ± 0.003 | 0.939 ± 0.002 | 0.247 ± 0.005 |
|           | initial              | 1.0 | 0.796 ± 0.002 | 0.941 ± 0.002 | 0.243 ± 0.001 |
|           | learned              | 5.0 | 0.819 ± 0.001 | 0.940 ± 0.001 | 0.263 ± 0.008 |
|           | initial              | 1.0 | 0.795 ± 0.002 | 0.941 ± 0.001 | 0.241 ± 0.002 |
|           | learned              | 5.0 | 0.805 ± 0.002 | 0.940 ± 0.001 | 0.244 ± 0.004 |
| SoftLabeler | initial             | 1.0 | 0.800 ± 0.003 | 0.940 ± 0.002 | 0.239 ± 0.004 |
|           | learned              | 5.0 | 0.821 ± 0.001 | 0.939 ± 0.001 | 0.265 ± 0.003 |
|           | initial              | 1.0 | 0.798 ± 0.003 | 0.940 ± 0.002 | 0.237 ± 0.004 |
|           | learned              | 5.0 | 0.806 ± 0.002 | 0.939 ± 0.001 | 0.248 ± 0.003 |
|           | initial              | 1.0 | 0.804 ± 0.001 | 0.939 ± 0.001 | 0.242 ± 0.004 |
|           | learned              | 5.0 | 0.778 ± 0.018 | 0.941 ± 0.001 | 0.224 ± 0.019 |
|           | initial              | 1.0 | 0.784 ± 0.003 | 0.940 ± 0.001 | 0.222 ± 0.001 |
|           | learned              | 5.0 | 0.811 ± 0.006 | 0.940 ± 0.002 | 0.241 ± 0.001 |
|           | initial              | 1.0 | 0.778 ± 0.003 | 0.939 ± 0.001 | 0.214 ± 0.009 |
|           | learned              | 5.0 | 0.789 ± 0.004 | 0.940 ± 0.002 | 0.224 ± 0.003 |

Table 13: Out-domain results for measure Largest and backbone ResNet18.
| Algorithm | Spectral | Calibration | k  | AUC   | InAsIn   | OutAsOut   |
|-----------|----------|-------------|----|-------|----------|-------------|
| ERM       | False    | initial     | 1.0| 0.846 | 0.942    | 0.336       |
|           |          | learned     | 5.0| 0.856 | 0.940    | 0.342       |
|           |          | initial     | 1.0| 0.843 | 0.943    | 0.323       |
|           |          | learned     | 5.0| 0.841 | 0.941    | 0.311       |
|           | True     | initial     | 1.0| 0.837 | 0.942    | 0.313       |
|           |          | learned     | 5.0| 0.854 | 0.941    | 0.334       |
| MCDropout | False    | initial     | 1.0| 0.838 | 0.943    | 0.315       |
|           |          | learned     | 5.0| 0.854 | 0.940    | 0.337       |
|           |          | initial     | 1.0| 0.839 | 0.943    | 0.318       |
|           |          | learned     | 5.0| 0.839 | 0.941    | 0.309       |
| MIMO      | False    | initial     | 1.0| 0.847 | 0.941    | 0.348       |
|           |          | learned     | 5.0| 0.846 | 0.941    | 0.345       |
|           |          | initial     | 1.0| 0.826 | 0.938    | 0.288       |
|           |          | learned     | 5.0| 0.845 | 0.942    | 0.330       |
| Mixup     | False    | initial     | 1.0| 0.841 | 0.940    | 0.327       |
|           |          | learned     | 5.0| 0.856 | 0.940    | 0.343       |
|           |          | initial     | 1.0| 0.820 | 0.942    | 0.290       |
|           |          | learned     | 5.0| 0.830 | 0.941    | 0.304       |
| OC        | False    | initial     | 1.0| 0.840 | 0.943    | 0.319       |
|           |          | learned     | 5.0| 0.853 | 0.942    | 0.331       |
|           |          | initial     | 1.0| 0.842 | 0.942    | 0.323       |
|           |          | learned     | 5.0| 0.859 | 0.941    | 0.346       |
| RND       | False    | initial     | 1.0| 0.837 | 0.941    | 0.317       |
|           |          | learned     | 5.0| 0.853 | 0.940    | 0.338       |
|           |          | initial     | 1.0| 0.840 | 0.941    | 0.322       |
|           |          | learned     | 5.0| 0.839 | 0.939    | 0.309       |
| SoftLabeler | False | initial   | 1.0| 0.835 | 0.942    | 0.310       |
|            |          | learned     | 5.0| 0.851 | 0.941    | 0.325       |
|            |          | initial     | 1.0| 0.821 | 0.941    | 0.284       |
|            |          | learned     | 5.0| 0.826 | 0.941    | 0.282       |

Table 14: Out-domain results for measure Largest and backbone ResNet50.
| algorithm | spectral calibration | k  | AUC   | InAsIn | OutAsOut |
|-----------|----------------------|----|-------|--------|-----------|
| ERM       | False                | 5.0| 0.768 ± 0.004 | 0.942 ± 0.002 | 0.126 ± 0.002 |
|           | learned              | 5.0| 0.768 ± 0.004 | 0.942 ± 0.002 | 0.126 ± 0.002 |
|           | initial              | 5.0| 0.771 ± 0.007 | 0.943 ± 0.003 | 0.128 ± 0.009 |
|           | learned              | 5.0| 0.771 ± 0.007 | 0.943 ± 0.003 | 0.128 ± 0.009 |
|           | initial              | 5.0| 0.519 ± 0.000 | 0.960 ± 0.001 | 0.094 ± 0.003 |
|           | learned              | 5.0| 0.772 ± 0.003 | 0.945 ± 0.002 | 0.126 ± 0.002 |
| MCDropout | False                | 5.0| 0.517 ± 0.004 | 0.958 ± 0.003 | 0.099 ± 0.004 |
|           | learned              | 5.0| 0.772 ± 0.003 | 0.945 ± 0.002 | 0.126 ± 0.002 |
|           | initial              | 5.0| 0.521 ± 0.003 | 0.959 ± 0.001 | 0.099 ± 0.000 |
|           | learned              | 5.0| 0.770 ± 0.002 | 0.942 ± 0.002 | 0.126 ± 0.003 |
| MIMO      | False                | 5.0| 0.730 ± 0.008 | 0.946 ± 0.004 | 0.088 ± 0.014 |
|           | learned              | 5.0| 0.730 ± 0.008 | 0.946 ± 0.004 | 0.088 ± 0.014 |
|           | initial              | 5.0| 0.733 ± 0.003 | 0.947 ± 0.002 | 0.086 ± 0.005 |
|           | learned              | 5.0| 0.733 ± 0.003 | 0.947 ± 0.002 | 0.086 ± 0.005 |
| Mixup     | False                | 1.0| 0.438 ± 0.009 | 0.942 ± 0.007 | 0.041 ± 0.002 |
|           | learned              | 5.0| 0.754 ± 0.002 | 0.945 ± 0.001 | 0.109 ± 0.005 |
|           | initial              | 1.0| 0.498 ± 0.008 | 0.948 ± 0.003 | 0.055 ± 0.002 |
|           | learned              | 5.0| 0.754 ± 0.002 | 0.945 ± 0.001 | 0.109 ± 0.005 |
|           | initial              | 5.0| 0.745 ± 0.002 | 0.946 ± 0.001 | 0.098 ± 0.002 |
|           | learned              | 5.0| 0.745 ± 0.002 | 0.946 ± 0.001 | 0.098 ± 0.002 |
| OC        | False                | 1.0| 0.500 ± 0.000 | 1.000 ± 0.000 | 0.000 ± 0.000 |
|           | learned              | 5.0| 0.769 ± 0.004 | 0.942 ± 0.003 | 0.125 ± 0.007 |
|           | initial              | 1.0| 0.500 ± 0.000 | 1.000 ± 0.000 | 0.000 ± 0.000 |
|           | learned              | 5.0| 0.769 ± 0.004 | 0.944 ± 0.003 | 0.120 ± 0.004 |
| RND       | False                | 1.0| 0.390 ± 0.011 | 0.957 ± 0.001 | 0.011 ± 0.005 |
|           | learned              | 5.0| 0.771 ± 0.012 | 0.943 ± 0.002 | 0.125 ± 0.009 |
|           | initial              | 1.0| 0.390 ± 0.011 | 0.957 ± 0.001 | 0.011 ± 0.005 |
|           | learned              | 5.0| 0.771 ± 0.012 | 0.943 ± 0.002 | 0.125 ± 0.009 |
|           | initial              | 1.0| 0.468 ± 0.049 | 0.956 ± 0.001 | 0.031 ± 0.014 |
|           | learned              | 5.0| 0.769 ± 0.002 | 0.944 ± 0.003 | 0.123 ± 0.005 |
| SoftLabeler| False              | 1.0| 0.370 ± 0.042 | 0.956 ± 0.000 | 0.009 ± 0.005 |
|           | learned              | 5.0| 0.769 ± 0.002 | 0.944 ± 0.003 | 0.123 ± 0.005 |
|           | initial              | 1.0| 0.769 ± 0.024 | 0.941 ± 0.002 | 0.228 ± 0.019 |
|           | learned              | 5.0| 0.735 ± 0.007 | 0.947 ± 0.002 | 0.103 ± 0.008 |
|           | initial              | 1.0| 0.683 ± 0.027 | 0.941 ± 0.001 | 0.224 ± 0.019 |
|           | learned              | 5.0| 0.735 ± 0.007 | 0.947 ± 0.002 | 0.103 ± 0.008 |
|           | initial              | 1.0| 0.771 ± 0.004 | 0.940 ± 0.000 | 0.221 ± 0.010 |
|           | learned              | 5.0| 0.734 ± 0.004 | 0.948 ± 0.001 | 0.096 ± 0.004 |
|           | initial              | 1.0| 0.683 ± 0.002 | 0.939 ± 0.001 | 0.214 ± 0.009 |
|           | learned              | 5.0| 0.734 ± 0.004 | 0.948 ± 0.001 | 0.096 ± 0.004 |

Table 15: Out-domain results for measure Native and backbone ResNet18.
| algorithm | spectral | calibration | k   | AUC   | InAsIn | OutAsOut |
|-----------|----------|-------------|-----|-------|--------|-----------|
| ERM       | False    | initial     | 5.0 | 0.792 ± 0.005 | 0.941 ± 0.001 | 0.144 ± 0.003 |
|           |          | learned     | 5.0 | 0.792 ± 0.005 | 0.941 ± 0.001 | 0.144 ± 0.003 |
|           | True     | initial     | 5.0 | 0.793 ± 0.002 | 0.942 ± 0.000 | 0.143 ± 0.006 |
|           |          | learned     | 5.0 | 0.793 ± 0.002 | 0.942 ± 0.000 | 0.143 ± 0.006 |
|           | False    | initial     | 1.0 | 0.523 ± 0.003 | 0.968 ± 0.001 | 0.094 ± 0.004 |
|           |          | learned     | 5.0 | 0.792 ± 0.005 | 0.942 ± 0.003 | 0.148 ± 0.009 |
|           | True     | initial     | 1.0 | 0.523 ± 0.003 | 0.968 ± 0.001 | 0.094 ± 0.004 |
|           |          | learned     | 5.0 | 0.792 ± 0.005 | 0.942 ± 0.003 | 0.148 ± 0.009 |
| MCDropout | False    | initial     | 5.0 | 0.793 ± 0.004 | 0.941 ± 0.004 | 0.147 ± 0.002 |
|           |          | learned     | 5.0 | 0.793 ± 0.004 | 0.941 ± 0.004 | 0.147 ± 0.002 |
|           | True     | initial     | 5.0 | 0.766 ± 0.005 | 0.943 ± 0.002 | 0.097 ± 0.007 |
|           |          | learned     | 5.0 | 0.766 ± 0.005 | 0.943 ± 0.002 | 0.097 ± 0.007 |
| MIMO      | False    | initial     | 1.0 | 0.457 ± 0.003 | 0.928 ± 0.003 | 0.053 ± 0.004 |
|           |          | learned     | 5.0 | 0.774 ± 0.008 | 0.945 ± 0.000 | 0.115 ± 0.003 |
|           | True     | initial     | 1.0 | 0.471 ± 0.019 | 0.948 ± 0.004 | 0.041 ± 0.003 |
|           |          | learned     | 5.0 | 0.774 ± 0.008 | 0.945 ± 0.000 | 0.115 ± 0.003 |
| Mixup     | False    | initial     | 1.0 | 0.458 ± 0.015 | 0.945 ± 0.000 | 0.036 ± 0.002 |
|           |          | learned     | 5.0 | 0.776 ± 0.002 | 0.943 ± 0.002 | 0.119 ± 0.007 |
|           | True     | initial     | 1.0 | 0.497 ± 0.008 | 0.945 ± 0.002 | 0.049 ± 0.005 |
|           |          | learned     | 5.0 | 0.776 ± 0.002 | 0.943 ± 0.002 | 0.119 ± 0.007 |
| OC        | False    | initial     | 1.0 | 0.500 ± 0.000 | 1.000 ± 0.000 | 0.000 ± 0.000 |
|           |          | learned     | 5.0 | 0.794 ± 0.007 | 0.941 ± 0.000 | 0.147 ± 0.010 |
|           | True     | initial     | 1.0 | 0.500 ± 0.000 | 1.000 ± 0.000 | 0.000 ± 0.000 |
|           |          | learned     | 5.0 | 0.791 ± 0.006 | 0.940 ± 0.003 | 0.144 ± 0.006 |
| RND       | False    | initial     | 1.0 | 0.488 ± 0.023 | 0.957 ± 0.001 | 0.019 ± 0.004 |
|           |          | learned     | 5.0 | 0.793 ± 0.003 | 0.940 ± 0.002 | 0.146 ± 0.005 |
|           | True     | initial     | 1.0 | 0.488 ± 0.023 | 0.957 ± 0.001 | 0.019 ± 0.004 |
|           |          | learned     | 5.0 | 0.793 ± 0.003 | 0.940 ± 0.002 | 0.146 ± 0.005 |
| SoftLabeler | False  | initial    | 1.0 | 0.826 ± 0.004 | 0.942 ± 0.001 | 0.310 ± 0.013 |
|           |          | learned     | 5.0 | 0.766 ± 0.003 | 0.947 ± 0.000 | 0.114 ± 0.003 |
|           | True     | initial     | 1.0 | 0.716 ± 0.003 | 0.941 ± 0.003 | 0.284 ± 0.011 |
|           |          | learned     | 5.0 | 0.766 ± 0.003 | 0.947 ± 0.000 | 0.114 ± 0.003 |

Table 16: Out-domain results for measure Native and backbone ResNet50.
C ImageNot dataset

Our ImageNot dataset is a class partition of the ImageNet ILSVRC2012 dataset (Russakovsky et al., 2015).

C.1 In-domain classes

| n02666196 | n03032252 | n03482405 | n03793489 | n04118776 | n04447861 |
| n02669723 | n03042490 | n03495258 | n03794056 | n04131690 | n04456115 |
| n02672831 | n03045698 | n03496892 | n03803284 | n04143272 | n04456333 |
| n02690373 | n03063599 | n03498985 | n03814639 | n04146614 | n04462240 |
| n02699494 | n03075370 | n03527444 | n03841143 | n04153751 | n04465501 |
| n02776631 | n03085013 | n03534580 | n03843555 | n04154565 | n04476259 |
| n02783161 | n03095699 | n03553780 | n03857828 | n04162706 | n04482393 |
| n02786058 | n03124170 | n03584254 | n03866082 | n04179913 | n04485082 |
| n02791124 | n03127747 | n03590841 | n03868863 | n04192698 | n04487394 |
| n02793495 | n03127925 | n03594734 | n03874293 | n04201297 | n04501370 |
| n02794156 | n03131574 | n03594945 | n03874599 | n04208210 | n04517823 |
| n02797295 | n03160309 | n03627232 | n03877845 | n04209133 | n04522168 |
| n02799071 | n03180011 | n03633091 | n03884397 | n04228054 | n04525038 |
| n02804610 | n03187595 | n03657121 | n03887697 | n04235860 | n04525305 |
| n02808304 | n03201208 | n03661043 | n03891332 | n04238763 | n04540053 |
| n02808440 | n03207743 | n03670208 | n03895866 | n04243546 | n04548280 |
| n02814860 | n03207941 | n03680355 | n03903868 | n04254120 | n04548362 |
| n02817516 | n03216828 | n03690938 | n03920288 | n04254777 | n04550184 |
| n02834397 | n03223299 | n03706229 | n03924679 | n04258138 | n04552348 |
| n02835271 | n03240683 | n03709823 | n03930313 | n04259630 | n04553703 |
| n02835779 | n03249569 | n03710193 | n03933933 | n04264628 | n04560804 |
| n02840245 | n03272010 | n03710721 | n03938244 | n04266014 | n04579145 |
| n02859443 | n03272562 | n03717622 | n03947888 | n04270147 | n04584207 |
| n02860847 | n03290653 | n03720891 | n03950228 | n04273569 | n04590129 |
| n02869837 | n03291819 | n03721384 | n03956157 | n04286575 | n04591157 |
| n02871525 | n03325584 | n03729826 | n03958227 | n04311004 | n04591713 |
| n02877765 | n03337140 | n03733281 | n03967562 | n04311174 | n04592741 |
| n02883205 | n03344393 | n03733805 | n03976657 | n04317175 | n04604644 |
| n02892767 | n03347037 | n03742115 | n03982430 | n04325704 | n04612504 |
| n02894605 | n03372029 | n03743016 | n03983396 | n04326547 | n04613696 |
| n02895154 | n03384352 | n03759954 | n03995372 | n04330267 | n06359193 |
| n02909870 | n03388549 | n03761084 | n04004767 | n04332243 | n07802026 |
| n02927161 | n03399312 | n03763968 | n04008634 | n04346328 | n07930864 |
| n02951358 | n03394916 | n03769881 | n04019541 | n04347754 | n09193705 |
| n02951585 | n03417042 | n03773504 | n04023962 | n04355933 | n09246464 |
| n0296687 | n03425413 | n03775071 | n04033995 | n04356056 | n09286635 |
| n02974003 | n03445777 | n03775546 | n04037443 | n04357314 | n09332890 |
| n02977058 | n03445924 | n03777568 | n04039381 | n04371430 | n09421951 |
| n02979186 | n03447721 | n03781244 | n04041544 | n04371774 | n09472597 |
| n02980441 | n03452741 | n03782006 | n04065272 | n04372370 | n10148035 |
| n02981792 | n03459775 | n03785016 | n04067472 | n04378676 | n15075141 |
| n02992211 | n03461385 | n03786901 | n04070727 | n04399382 | |
| n03000684 | n03467068 | n03787032 | n04081281 | n04404412 | |
| n03026506 | n03478589 | n03788365 | n04086273 | n04409515 | |
| n03028079 | n03481172 | n03791053 | n04090263 | n04418357 | |
C.2 Out-domain classes

| n01440764 | n01692333 | n01820546 | n02013706 | n02093428 | n02105162 |
| n01443537 | n01693334 | n01824575 | n02017213 | n02093647 | n02105251 |
| n01468450 | n01694178 | n01828970 | n02018207 | n02093754 | n02105412 |
| n01491361 | n01695060 | n01829413 | n02018795 | n02093859 | n02105505 |
| n01494475 | n01697457 | n01833805 | n02025239 | n02093991 | n02105641 |
| n01496331 | n01698640 | n01843065 | n02027492 | n02094114 | n02105855 |
| n01498041 | n01704323 | n01843383 | n02028035 | n02094258 | n02106030 |
| n01514668 | n01728672 | n01847000 | n02033041 | n02094433 | n02106166 |
| n01514859 | n01728920 | n01855032 | n02037110 | n02095314 | n02106382 |
| n01518878 | n01729322 | n01855672 | n02051845 | n02095570 | n02106550 |
| n01530575 | n01729977 | n01860187 | n02056570 | n02095889 | n02106662 |
| n01531178 | n01734418 | n01871265 | n02058221 | n02096051 | n02107142 |
| n01532829 | n01735189 | n01872401 | n02066245 | n02096177 | n02107312 |
| n01534433 | n01737021 | n01873310 | n02071294 | n02096294 | n02107874 |
| n01537544 | n01739381 | n01877812 | n02074367 | n02096437 | n02107683 |
| n01558993 | n01740131 | n01882714 | n02077923 | n02096585 | n02107908 |
| n01560419 | n01742172 | n01883070 | n02085620 | n02097047 | n02108000 |
| n01580077 | n01744401 | n01910747 | n02085782 | n02097130 | n02108089 |
| n01582220 | n01748264 | n01914609 | n02085936 | n02097209 | n02108422 |
| n01592084 | n01749939 | n01917289 | n02086079 | n02097298 | n02108551 |
| n01601694 | n01751748 | n01924916 | n02086240 | n02097474 | n02108915 |
| n01608432 | n01753488 | n01930112 | n02086646 | n02097658 | n02109047 |
| n01614925 | n01755681 | n01943899 | n02086910 | n02098105 | n02109525 |
| n01616318 | n01756291 | n01944390 | n02087046 | n02098286 | n02109961 |
| n01622779 | n01768244 | n01945685 | n02087394 | n02098413 | n02110063 |
| n01629819 | n01770081 | n01950731 | n02088094 | n02099267 | n02110185 |
| n01630670 | n01770393 | n01955084 | n02088238 | n02099429 | n02110341 |
| n01631663 | n01773157 | n01968897 | n02088364 | n02099601 | n02110627 |
| n01632458 | n01773549 | n01978287 | n02088466 | n02099712 | n02110806 |
| n01632777 | n01773797 | n01978455 | n02088632 | n02099849 | n02110958 |
| n01641577 | n01774384 | n01980166 | n02089078 | n02100236 | n02111129 |
| n01644373 | n01774750 | n01981276 | n02089867 | n02100583 | n02111277 |
| n01644900 | n01775062 | n01983481 | n02089973 | n02100735 | n02111500 |
| n01664065 | n01776313 | n01984695 | n02090379 | n02100877 | n02111889 |
| n01665541 | n01784875 | n01985128 | n02090622 | n02101006 | n02112018 |
| n01667114 | n01795545 | n01986214 | n02090721 | n02101388 | n02112137 |
| n01667778 | n01796340 | n01990800 | n02091032 | n02101556 | n02112350 |
| n01669191 | n01797986 | n02002556 | n02091134 | n02102040 | n02112706 |
| n01675722 | n01798484 | n02002724 | n02091244 | n02102177 | n02113023 |
| n01677366 | n01806143 | n02006656 | n02091467 | n02102318 | n02113186 |
| n01682214 | n01806567 | n02007558 | n02091635 | n02102480 | n02113624 |
| n01685808 | n01807496 | n02009229 | n02091831 | n02102973 |
| n01687978 | n01817953 | n02009912 | n02092002 | n02104029 |
| n01688243 | n01818515 | n02011460 | n02092339 | n02104365 |
| n01689811 | n01819313 | n02012849 | n02093256 | n02105056 |