Useful Confidence Measures: Beyond the Max Score

Gal Yona*
Weizmann Institute
gal.yona@gmail.com

Amir Feder
Columbia University
amirfeder@gmail.com

Itay Laish
Google
itaylaish@google.com

Abstract

An important component in deploying machine learning (ML) in safety-critic applications is having a reliable measure of confidence in the ML model's predictions. For a classifier f producing a probability vector $f(x)$ over the candidate classes, the confidence is typically taken to be $\max_i f(x)_i$. This approach is potentially limited, as it disregards the rest of the probability vector. In this work, we derive several confidence measures that depend on information beyond the maximum score, such as margin-based and entropy-based measures, and empirically evaluate their usefulness, focusing on NLP tasks with distribution shifts and Transformer-based models. We show that when models are evaluated on the out-of-distribution data "out of the box", using only the maximum score to inform the confidence measure is highly suboptimal. In the post-processing regime (where the scores of f can be improved using additional in-distribution held-out data), this remains true, albeit less significant. Overall, our results suggest that entropy-based confidence is a surprisingly useful measure.

1 Introduction

As machine learning (ML) is increasingly deployed in high-stakes decision-making applications, it becomes increasingly important that practitioners have access to reliable measure of how confident the ML model is in its various predictions. This becomes especially crucial in settings where the predictions are made in conditions significantly different than the ones present during development. In these cases, accuracy may unavoidably degrade, but a useful confidence measure can at least ensure practitioners "know" when the ML model "doesn't know".

In this work we assume classifiers f output a probability vector $f(x)$ over the candidate classes $\mathcal{Y}$ and treat *confidence* as a scalar quantity $c(f(x)) \in [0, 1]$ that represents how confident f is in its prediction, with scores near 1 representing highly confident predictions. Intuitively, a good confidence measure should give rise to scores that correlate well with the accuracy of f. In Section 3 we show that this objective can be decomposed into two familiar terms from the literature on forecasting (Murphy, 1973; Dawid, 1982): a *calibration error* term (encouraging that whenever we output a confidence value of e.g. 0.7, then on average, 70% of the time the model makes a correct prediction) and a *sharpness* term (encouraging the confidence values to also be varied).

Given a classifier f, what should we choose as our confidence measure c? One natural choice is to use the maximum class probability, $c(f(x)) = \max_i f(x)_i$. When f is itself calibrated, this will give rise to a calibrated confidence measure. However, we don't necessarily expect models to be

*Work done while an intern at Google.

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

well-calibrated "out of the box", especially not in the presence of distribution shifts. While it is common to post-process predictions to improve their calibration, this approach is not always feasible as it requires additional data, is observed to have limited success in settings of distributions shifts (Desai and Durrett, 2020; Dan and Roth, 2021), and may come at the cost of sharpness (Kumar et al., 2018).

The above discussion suggests that in the presence of model miscalibration, using the tail of the predictions to inform the confidence score could be beneficial.[2] Indeed, the literature on active learning (AL) has long since considered uncertainty scores that employ the rest of the probability vector (Settles, 2009). In the AL context, these are used to greedily select examples from a large pool of unlabeled examples for which labels will be requested. E.g., it is common to use the gap between the first and second largest entries of $f(x)$, and in some cases even the entropy of $f(x)$, to inform this selection.

Our contributions. In this work, we consider such uncertainty scores in the context of confidence measures, and perform a systematic evaluation of these measures in the presence of distribution shifts. We focus on large pre-trained Transformer-based language models like BERT (Devlin et al., 2018) for multi-class NLP tasks, which have observed to be well-calibrated on in-distribution data (Desai and Durrett, 2020). We use the Amazon reviews dataset from the WILDS benchmark (Koh et al., 2021), in which the out-of-distribution (OOD) test set consists of a set of reviewers that is disjoint from the training set and in-distribution (ID) validation set. We consider models trained on this task with different objectives (regular risk minimization, but also approaches that are designed to handle distribution shifts), and evaluate the different confidence measures on the OOD test data. Our key findings are:

1. When the confidence measures are evaluated "out of the box" (with no further tuning based on a validation set), using $\max_i f(x)_i$ is highly sub-optimal. Margin-based confidence measures perform better for most of the models considered (and by a significant gap), and the entropy-based confidence measure is consistently better.
2. We derive a variant of temperature scaling (TS), a popular post-processing technique for improving calibration, and show that it can be used to consistently improve the calibration for all the confidence measures we consider.
3. In the post-processing regime (namely, after applying TS), the entropy-based confidence measure Pareto dominates the max-based measure for most of the models (and is otherwise incomparable - has a marginally larger calibration error but is sharper).

Additional related work. (Desai and Durrett, 2020) evaluate the calibration of pre-trained Transformer models in both ID and OOD settings. Their results demonstrate that Transformer-based models tend to well-calibrated ID but that the calibration error can decrease significantly OOD. (Dan and Roth, 2021) empirically evaluate the relationship between scale and calibration, showing that OOD, smaller Transformer models tend to have worse calibration than larger models, even after applying TS. These works, together with earlier works (Guo et al., 2017), all conflate a model's confidence with the probability of the predicted label. One recent exception is (Taha et al., 2022), which consider confidence measures based on the margin and kurtosis of the logits. Their work is significantly different from ours as they do not directly evaluate calibration of these proposed measures and also do not consider distribution shifts.

2 Preliminaries

Setup. We consider a multi-class classification problem with feature space $\mathcal{X}$ and label space $\mathcal{Y}$, where $|\mathcal{Y}| = k \geq 2$ and $\Delta(\mathcal{Y})$ denotes the simplex over $\mathbb{R}^k$. Let $\mathbb{P}$ denote a joint distribution over $\mathcal{X} \times \mathcal{Y}$ and X, Y random variables w.r.t $\mathbb{P}$. A classifier is a mapping $f : \mathcal{X} \to \Delta(\mathcal{Y})$ and its predicted label is $\arg\max_i f(x)_i$. We use err to denote the 0-1 error of f: $\text{err}(x, y) = \mathbf{1}[\arg\max_i f(x)_i \neq y]$.

[2] As one illustrative example, consider a 10-class classification task and the predictions on two instances: $f(x_1) = [0.9, 0.1, 0.0, \ldots 0.0]$ and $f(x_2) = [0.9, 0.1/9, \ldots, 0.1/9]$. We might expect that the confidence on x_1 should be higher than on x_2: intuitively, for x_1 the model is "deliberating" between two concrete options (the first and second classes) whereas for x_2 there is no clear alternative to the first class. However, by definition, $\max_i f(x)_i$ can make no such distinctions.

Calibration. A binary classifier $f : \mathcal{X} \to [0,1]$ is said to be *calibrated* if $\forall v \in [0,1]$, $\Pr[Y = 1|f(X) = v] = v$. For classification problems with multiple classes ($k > 2$), there are different ways to define calibration (Widmann et al., 2019; Zhao et al., 2021). Arguably the simplest and most popular approach is to restrict the attention only to the most likely prediction (Guo et al., 2017). According to this definition, a classifier $f : \mathcal{X} \to \Delta(\mathcal{Y})$ is calibrated if $\forall v \in [0,1]$, $\Pr[Y = \arg\max_i f(X)_i | \max_i f(X)_i = v] = v$.

Post-hoc calibration. Since many ML models are not typically calibrated "out of the box", it is a common practice to post-process the model outputs in a way that improves their calibration. Methods include Histogram Binning (Zadrozny and Elkan, 2001), Isotonic Regression (Zadrozny and Elkan, 2002) and Platt scaling (Platt et al., 1999). In the context of modern ML models, the method of Temperature Scaling (TS) has demonstrated to be both simple and effective (Guo et al., 2017). For classifiers of the form $f(x) = \sigma(\mathbf{z})$ (where $\mathbf{z}$ is the logit vector and σ is the softmax operator), TS rescales the logit vector $\mathbf{z}$ by a factor of T before applying softmax σ. The hyperparameter T is called the temperature as it has the effect of "softening" the softmax (i.e. increasing the entropy of the output probability vector) when $T > 1$. In particular, $T = 1$ recovers the original model output and as $T \to \infty$ the model output approaches $1/K$. T is optimized by minimizing the Negative Log Liklihood on a labeled validation set.

3 Confidence measures

Notation. The entropy of a vector $\mathbf{v} \in \Delta(\mathcal{Y})$ is $H(\mathbf{v}) = -\sum_{i=1}^{k} v_i \cdot \log(v_i)$. We use $\tilde{H}(\mathbf{v}) = \frac{1}{\log k} \cdot H(\mathbf{v})$ to normalize the entropy to be in $[0,1]$. We also use $\tilde{\mathbf{v}}$ to denote the sorted version of $\mathbf{v}$ (in descending order), so that $\tilde{v}_i$ is the i-th largest value in $\mathbf{v}$.

Confidence measures. A confidence measure is a mapping $c : \Delta(\mathcal{Y}) \to [0,1]$. The confidence of a classifier f on input $x \in \mathcal{X}$ according to c is then $c(f(x))$, where values near 1 represent high-confidence predictions and values near 0 represent low-confidence predictions. In this work, we consider the following measures: $\texttt{max} : \mathbf{v} \mapsto \tilde{v}_1$; $\texttt{margin12} : \mathbf{v} \mapsto \tilde{v}_1 - \tilde{v}_2$; $\texttt{margin123} : \mathbf{v} \mapsto \tilde{v}_1 - (0.5\tilde{v}_2 + 0.5\tilde{v}_3)$; and $\texttt{entropy} : \mathbf{v} \mapsto \tilde{H}(\mathbf{v})$. See Figure 2 in Appendix A for a comparison between these measures in a classification problem with $k = 3$ classes.

3.1 Evaluating confidence measures: calibration and sharpness

A good confidence signal c should give rise to calibrated confidence predictions; Namely, the binary classifier $c \circ f \in [0,1]$ should be calibrated. However, calibration by itself does not guarantee that the confidence measure is useful. For example, the confidence measure that always output the marginal accuracy $c(x) = \Pr[1 - \text{err}(X,Y)]$ will be perfectly calibrated but useless for practical purposes. Thus, a good confidence measure should also give rise to a variety of confidence values; this requirement is referred to as sharpness (Gneiting et al., 2007)[3]. Letting $T(x) = \mathbf{E}[\text{err}(X,Y)|c(X) = c(x)]$, we can decompose the ℓ_2 "loss" of a confidence measure c as:

$$\mathbf{E}[(\text{err}(X,Y) - c(X))^2] = \mathbf{Var}[\text{err}(X,Y)] - \underbrace{\mathbf{Var}[T(x)]}_{\text{sharpness}} + \underbrace{\mathbf{E}[(T(x) - c(X))^2]}_{\text{calibration error}} \quad (1)$$

The decomposition is a direct application of a similar decomposition for binary classifiers[4]. Consider the three terms in the decomposition of (1). The first term, from the perspective of the choice of the confidence measure, is irreducible. The sharpness term measures the variation in the error across confidence predictions. The calibration term measures how closely the confidence predictions track the error. Overall, this suggests that a good confidence measure should strike a balance between minimizing miscalibration and maximizing sharpness.

Estimation from finite samples. In principle, estimating $T(x)$ accurately for every value $c(x) \in [0,1]$ requires an infinite amount of data. To estimate both calibration and sharpness from finite data,

[3] Another approach is to consider calibration on overlapping structured subgroups of the data (Hébert-Johnson et al., 2018; Barda et al., 2021), but we focus instead on global calibration and sharpness.

[4] (Kuleshov and Liang, 2015) prove that for any $y : \mathcal{X} \to [0,1]$ and $F : \mathcal{X} \to [0,1]$, $\mathbf{E}[(y(x) - F(x))^2] = \mathbf{Var}[y(x)] - \mathbf{Var}[T(x)] + \mathbf{E}[T(x) - F(x))^2]$, where $T(x) = \mathbf{E}[y(x)|F(x)]$. The decomposition we employ for confidence measures can be obtained by taking $y(x)$ to be $\text{err}(x) = \mathbf{1}[f(x) \neq y(x)]$ and $F \equiv c$.

we use discretized versions of both notions. Specifically, let $\mathcal{B}$ be a partitioning ("binning") of the interval $[0, 1]$, where $B : [0, 1] \to \mathcal{B}$ maps any $v \in [0, 1]$ to the bin $B(v)$ that contains it. Then, we redefine $T(x)$ as follows: $T_{\mathcal{B}}(x) = \mathbf{E}[\text{err}(X, Y)|B(c(X)) = B(c(x))]$.

Binning strategy. Fix a granularity parameter n representing the number of target bins. We distinguish between a *fixed* binning strategy – in which $\mathcal{B}$ is formed by partitioning $[0, 1]$ interval into n equally-spaced bins; and an *adaptive* binning strategy – in which $\mathcal{B}$ is formed in a way that depend on the classifier f. Specifically, such that each bin has equal mass under $c \circ f$. In terms of measuring the calibration error, this is the different between the Expected Calibration Error (ECE) (Naeini et al., 2015) and the Adaptive Calibration Error (ACE) (Nixon et al., 2019). Since our focus is on comparing different models in terms of their calibration error[5], we prefer to use adaptive binning.

3.2 Post-processing to improve the calibration of arbitrary confidence measures

In principle, a post-processing approach such as temperature scaling (TS) is designed to improve the calibration error of `max`. To generalize this idea to general confidence measures, instead of tuning the temperature value T to minimize the NLL, we simply perform a line search over the relevant region (e.g. $[0, 2]$) to choose the value T that minimizes the calibration error of the desired confidence signal over a validation set. From this point, when we refer to TS, we mean this approach.

4 Evaluation

To empirically study optimal confidence signals under realistic and naturally occurring distribution shifts, we use the Wilds benchmark (Koh et al., 2021). We continue to our evaluation procedure.

Data. We use the Amazon Wilds dataset, which is a multi-class sentiment classification task derived from the Amazon Reviews dataset (Ni et al., 2019). The input x is the test of a review, the label y is a corresponding star rating (from 1 to 5), and the domain d is the identifier of the reviewer who wrote the review. The validation and test splits are comprised of both an in-distribution (ID) set and an out-of-distribution (OOD) set. The ID sets consist of reviews that are not in the training set, but are written by reviewers that *are* in the training set. The OOD sets consists of reviews written by a collection of reviewers that are disjoint from the training set and ID set reviewers.

Models. We evaluate the calibration of DistilBERT-base-uncased models (Sanh et al., 2019) that were finetuned on the Amazon Wilds training set[6]. All models were fine-tuned using a similar setup (grid search over learning rate and objective-specific parameters, with the rest of the hyperparameters set to standard/default values; see Appendix E in (Koh et al., 2021) for a detailed description). The models are standard ERM, IRM (Arjovsky et al., 2019) and GroupDRO (Sagawa et al., 2019).

Evaluation and results. For each model, we measure the performance of each confidence measure (`max`, `margin12`, `margin123` and `entropy`) via its sharpness and calibration errors. Results are means over three independent training runs. We distinguish between "out of the box" performance and performance with temperature scaling, with the temperature value T (per confidence measure) optimized on the ID validation set. We visualize the results in Figure 1; see Appendix A for numerical results. We see that in the OOB regime, `max` is never optimal (`entropy` is always preferrable, and `margin123` is better for ERM and IRM type models). In the TS regime, `entropy` Pareto-dominates `max` for both ERM and IRM model (for groupDRO, they are incomparable).

5 Discussion

In this work we define and empirically evaluate several confidence measures, and show that measures beyond the max score achieve favourable results. These observations join other recent work (Rothblum and Yona, 2022) in demonstrating that that in the presence of model miscalibration, additional attention should be given to otherwise trivial choices, in this case the confidence measure.

Our empirical findings raise several natural directions for future exploration. First, the striking efficacy of using the entropy as a confidence measure seems surprising, and it is interesting to explore

[5](Nixon et al., 2019) note several important limitations of the fixed binning strategy. Namely, that when predictions are skewed, many regions of the $[0, 1]$ interval sparsely populated, so only a few bins are "active".

[6]We use the weights provided in (Koh et al., 2021).

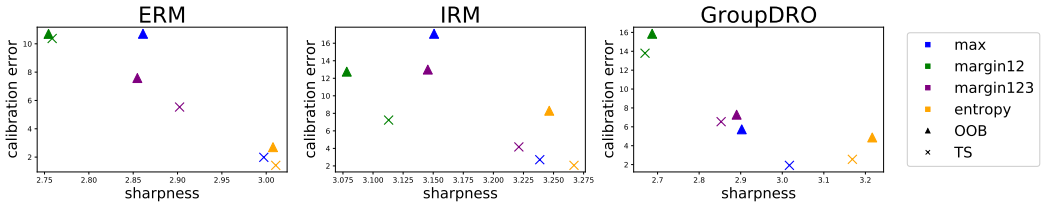

Figure 1: The calibration errors (y-axis, *lower is better*) vs sharpness (x-axis, *higher is better*) for each confidence measure, before and after temperature scaling, for different models.

whether this finding generalizes to additional settings and whether there are cases in which it can be justified theoretically. It is also natural to consider arbitrary confidence measures, beyond those we chose to evaluate here. In principle, the problem of finding the *optimal* confidence measure could itself be cast as an ML problem, and solved using ML tools. Our preliminary findings suggest that this type of "confidence meta-learning" is a promising direction for future work.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Barda, N., Yona, G., Rothblum, G. N., Greenland, P., Leibowitz, M., Balicer, R., Bachmat, E., and Dagan, N. (2021). Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, 28(3):549–558.

Dan, S. and Roth, D. (2021). On the effects of transformer size on in-and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Desai, S. and Durrett, G. (2020). Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Kuleshov, V. and Liang, P. S. (2015). Calibrated structured prediction. *Advances in Neural Information Processing Systems*, 28.

Kumar, A., Sarawagi, S., and Jain, U. (2018). Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *CVPR Workshops*, volume 2.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Rothblum, G. N. and Yona, G. (2022). Decision-making under miscalibration. *arXiv preprint arXiv:2203.09852*.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Settles, B. (2009). Active learning literature survey.

Taha, A. A., Hennig, L., and Knoth, P. (2022). Confidence estimation of classification based on the distribution of the neural network output layer. *arXiv preprint arXiv:2210.07745*.

Widmann, D., Lindsten, F., and Zachariah, D. (2019). Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems*, 32.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Zhao, S., Kim, M., Sahoo, R., Ma, T., and Ermon, S. (2021). Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324.

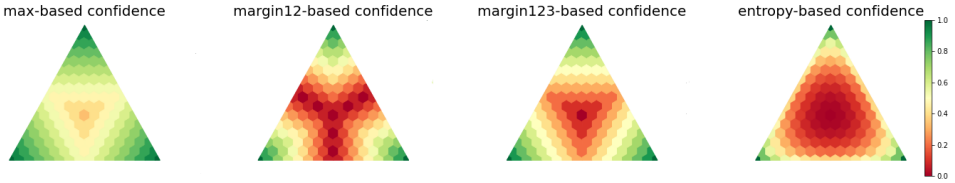

Figure 2: Illustrating the different confidence measures in a 3-way classification problem.

Table 1: **Out-of-the-box evaluation** (ACE)

	`max`	`margin12`	`margin123`	`entropy`
ERM	10.71 (1.16)	10.70 (0.10)	7.58 (0.76)	**2.70** (0.48)
IRM	17.08 (0.51)	12.76 (0.52)	12.99 (0.63)	**8.30** (0.62)
GroupDRO	5.73 (1.24)	15.86 (1.91)	7.29 (1.18)	**4.87** (1.42)

Table 2: **With temperature scaling** (ACE).

	`max`	`margin12`	`margin123`	`entropy`
ERM	1.99 (0.15)	10.38 (0.22)	5.53 (0.15)	**1.42** (0.10)
IRM	2.71 (0.24)	7.23 (0.17)	4.17 (0.12)	**2.06** (0.05)
GroupDRO	**1.92** (0.16)	13.79 (1.28)	6.54 (0.22)	2.56 (1.20)

Table 3: **Out-of-the-box evaluation** (ECE)

	`max`	`margin12`	`margin123`	`entropy`
ERM	1.34 (0.15)	1.02 (0.02)	0.84 (0.09)	**0.30** (0.07)
IRM	2.13 (0.06)	1.34 (0.07)	1.45 (0.07)	**0.94** (0.07)
GroupDRO	0.72 (0.16)	1.49 (0.22)	0.81 (0.13)	**0.52** (0.17)

Table 4: **With temperature scaling** (ECE)

	`max`	`margin12`	`margin123`	`entropy`
ERM	0.24 (0.02)	0.96 (0.02)	0.60 (0.01)	**0.17** (0.01)
IRM	0.33 (0.03)	0.65 (0.02)	0.45 (0.02)	**0.23** (0.01)
GroupDRO	**0.24** (0.02)	1.27 (0.14)	0.72 (0.03)	0.29 (0.13)

A Additional Figures

In Figure 2 we show heatmaps illustrating the behaviour of each confidence measure in a classification problem with three classes. The high confident predictions are marked in green (e.g. the simplex corners) and low confident predictions are in red (e.g., for all signals, confidence is minimized at the simplex center where the distribution over classes is uniform).

In Tables 1 - 4 we report results of the calibration errors for the various confidence measures, with and without applying temperature scaling on the in-distribution validation data, and when calibration is measured both with adaptive binning (ACE) and with fixed binning (ECE).