Test Time Augmentation Meets Post-hoc Calibration: Uncertainty Quantification under Real-World Conditions

Achim Hekler[1,3], **Titus J. Brinker**[1], **and Florian Buettner**[1,2,3]

[1]German Cancer Research Center (DKFZ) Heidelberg, Germany
[2]German Cancer Consortium (DKTK), Germany
[3]Goethe University Frankfurt, Germany
{achim.hekler, titus.brinker, florian.buettner}@dkfz.de

Abstract

Communicating the predictive uncertainty of deep neural networks transparently and reliably is important in many safety-critical applications such as medicine. However, modern neural networks tend to be poorly calibrated, resulting in wrong predictions made with a high confidence. While existing post-hoc calibration methods like temperature scaling or isotonic regression yield strongly calibrated predictions in artificial experimental settings, their efficiency can significantly reduce in real-world applications, where scarcity of labeled data or distribution shifts are commonly present. In this paper, we first investigate the impact of these characteristics on post-hoc calibration and introduce an easy-to-implement extension of common post-hoc calibration methods based on test time augmentation. In extensive experiments, we demonstrate that our approach results in substantially better calibration on various architectures. We demonstrate the robustness of our proposed approach on a real-world application for skin cancer classification and show that it facilitates safe decision-making under real-world uncertainties.

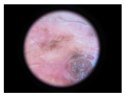
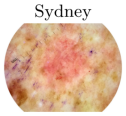
Hospital	Barcelona	Sydney
		
Ground truth	Melanoma	Melanoma
Model Prediction	Melanoma	Nevus
Methods	Confidence	Confidence
Uncalibrated model	0.99	0.99
Baseline calibration	0.85	0.91
Proposed calibration	0.92	0.78

Figure 1: Real-world distribution shift in skin cancer classification: recalibrated models make over-confident predictions. Left: in-distribution predictions by a model trained on images taken in Barcelona are made correctly with high confidence. Right: Under real-world distribution shift (image taken at a different hospital in Sidney) all models make wrong predictions, but only our proposed approach has a low confidence score reflecting model uncertainty.

Introduction

Deep neural networks are increasingly applied in safety-critical applications such as autonomous driving (Grigorescu et al. 2020) or medical diagnosis (Aggarwal et al. 2021). Such systems require not only high accuracy but also strongly calibrated uncertainty estimates. That is, the confidence score provided by the model should reflect its predictive uncertainty such that it matches the true likelihood of the prediction.

A lighthouse application for neural network-based classification in medicine is to support dermatologists in the diagnosis of skin cancer (Esteva et al. 2017). Studies have shown that image classifiers based on convolutional neural networks are on par or even superior to human experts in experimental settings (Brinker et al. 2019). However, several challenges arise when translating such systems from research to clinical practice. Due to different lighting conditions, acquisition systems or digital post-processing steps, the real-world deployment environment can generally be slightly different from the training domain (see Fig. 1). As a result, there may be a significant drop in the predictive power of the system.

For AI-based diagnostic systems to be applied routinely in the clinic, the underlying neural networks need to be able to transparently communicate the reliability of individual predictions throughout the entire life-cycle of the model. That is, predictive uncertainties need to be calibrated not only for in-distribution predictions, but also for distribution shift scenarios.

A variety of different post-hoc calibration methods have been proposed to ensure well-calibrated models (Guo et al. 2017; Platt et al. 1999; Zhang, Kailkhura, and Han 2020). These methods transform weakly calibrated model predictions such that in-distribution predictions are strongly calibrated. However, modern neural networks typically still make highly overconfident predictions under distribution shift, even after post-hoc calibration (Ovadia et al. 2019; Tomani and Buettner 2021).

Recent studies have examined calibration of deep neural networks under distribution shift typically on standard computer vision benchmarks such as IMAGENET-C or OBJECTNET for IMAGENET. However, calibration in the medical domain comes with requirements and challenges that are substantially different from those in these bench-

mark settings. In particular, binary classification tasks, such as whether a tumor is present or not, are common, and there is often a class imbalance between the positive and negative class. In addition, data labeling in the medical domain is typically a challenging and time-intensive task which requires domain-specific experts. Together with rigorous data protections regulations, this often results in data scarcity. However, these low-data medical regimes are precisely the safety-critical environments where models are required to communicate their uncertainty in a transparent and reliable manner.

Contributions. In this paper, we focus on the problem of post-hoc calibration under real-world uncertainties. We make the following main contributions:

- First, we introduce an easy-to-implement extension of common post-hoc calibration methods based on test time augmentation (TTA).
- With extensive experiments we show that the proposed TTA-based extensions consistently improve the performance of state-of-the-art post-hoc calibration methods in terms of calibration under distribution shift, without compromising on in-distribution calibration.
- At the example of skin cancer classification, we evaluate our approach on a real-world safety-critical application and showcase its data efficiency and practical relevance.

Related Work

In general, existing approaches towards neural networks with calibrated predictive uncertainties can be broadly divided into two categories. On the one hand, several approaches proposed to modify the training process in order to obtain strongly calibrated predictions (Lakshminarayanan, Pritzel, and Blundell 2017; Thulasidasan et al. 2019; Mukhoti et al. 2020; Tomani and Buettner 2021).

In this paper, we focus on the second category of calibration methods, so-called post-hoc calibration methods, in which the uncalibrated output of a trained model is transformed such that the resulting confidence scores better match the true likelihood of a prediction.

Post-hoc Calibration Methods

A plethora of post-hoc calibration methods have been proposed in recent literature and include both parametric and non-parametric approaches (Guo et al. 2017; Zhang, Kailkhura, and Han 2020; Gupta et al. 2021; Ma and Blaschko 2021; Wang, Feng, and Zhang 2021).

The key idea of these methods is to use a validation set sampled from the same distribution as the training set in order to rescale the original outputs of a trained neural network such that in-distribution predictions are strongly calibrated. A simple non-parametric post-hoc method is histogram binning (Zadrozny and Elkan 2001), which partitions all confidence scores into S bins. Then, a calibrated score calculated by optimizing a bin-wise squared loss function on the validation set is assigned to each bin. For each test-time prediction, the original confidence score is then replaced by the optimized value associated with the bin, in which the prediction falls.

Isotonic regression (IR) (Zadrozny and Elkan 2002) is an extension of histogramm binning. Here, the outputs of the neural network are divided into M intervals and a piecewise constant function is fitted on the validation set to transform uncalibrated outputs to calibrated confidence scores.

Besides these non-parametric methods, also parametric approaches for post-hoc calibration exist. Their main difference lies in the parametric family of the respective calibration function. For binary classification, Platt scaling (Platt et al. 1999) is an approach for transforming uncalibrated logits to calibrated confidence scores using logistic regression. A simple extension of Platt scaling is temperature scaling (TS) (Guo et al. 2017). Here, a single parameter T is learned to rescale the logits of the network. A more expressive alternative to TS is extended temperature scaling (ETS) (Zhang, Kailkhura, and Han 2020), which is based on a weighted ensemble of 3 fixed temperatures. Dirichlet calibration (Kull et al. 2019) allows learning within the family of linear functions $f(x) = Wx+b$, where $W \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. A regularization of the off-diagonal elements of W is suggested in (Kull, Silva Filho, and Flach 2017) in order to avoid overfitting .

Zhang, Kailkhura, and Han (2020) have proposed a combination of parametric and non-parametric methods. The authors suggest to perform isotonic regression after a temperature scaling step. Additionally, the authors introduce an accuracy-preserving version of isotonic regression (IRM) (Zhang, Kailkhura, and Han 2020).

A first comprehensive evaluation of calibration under distribution shift is presented in (Ovadia et al. 2019). The authors show by means of artificially generated distribution shifts that the quality of predictive uncertainties decreases with increasing distribution shift, regardless of the calibration method.

Test Time Augmentation

While data augmentation is typically used as part of training neural networks, it can also be used at test time. The idea behind test time augmentation is to let a model predict multiple transformations of a given image and then aggregate the individual predictions via an adequate method, e.g. by calculating the arithmetic mean.
Various studies have employed TTA to improve accuracy (Krizhevsky, Sutskever, and Hinton 2012; Szegedy et al. 2015), enhance robustness (Maron et al. 2021; Prakash et al. 2018), or perform uncertainty estimation in the context of classification (Wang et al. 2019a; Combalia et al. 2020) and segmentation (Javadi et al. 2022; Wang et al. 2019b).
However, a combination of TTA with post-hoc calibration has not been investigated yet. This is particularly of interest since recent work has demonstrated that combining several calibration methods does not necessarily lead to better results. For example, Wang, Feng, and Zhang (2021) have shown that it is harder to further calibrate predictions with post-hoc calibration methods obtained from models with implicit or explicit regularization techniques.

Algorithm 1: TTA-based extension of common post-hoc re-calibration algorithms

Input - trained model $f(\cdot)$
- post-hoc calibrator $c(\cdot)$ tuned for $f(\cdot)$ on hold-out labeled validation set $\mathcal{D}^{\text{val}}$
- set of transformations $\mathcal{T}$
- test image x
- number of considered augmentations T

Output calibrated uncertainty prediction p for test image x

$t \leftarrow 0$ ▷ Initialize iteration counter
$\mathcal{L} \leftarrow \emptyset$
logits $\leftarrow f(x)$ ▷ Calculate logits of original image
$\mathcal{L} \leftarrow \mathcal{L} \cup$ logits
while $t < T$ **do**
  $\widetilde{x} \leftarrow \text{tta}(x, \mathcal{T})$ ▷ TTA of test image
  logits $\leftarrow f(\widetilde{x})$ ▷ Calculate logits of augmentations
  $\mathcal{L} \leftarrow \mathcal{L} \cup$ logits
  $t \leftarrow t + 1$
end while
$\hat{l} \leftarrow \underline{0}$
for logits in $\mathcal{L}$ **do** ▷ Calculation of mean logits
  $\hat{l} \leftarrow \hat{l}+$logits
end for
$\hat{l} \leftarrow \hat{l}/(T+1)$
$p = c(\hat{l})$ ▷ Calculate calibrated uncertainty score

Problem Formulation and Definitions

Top-label Recalibration under distribution Shift

In this paper, we address the problem of recalibrating an already trained neural network. Let $x \in \mathbb{R}^D$ represent the D-dimensonal input and $y \in \mathcal{Y} := \{1, \dots, K\}$ the corresponding labels in a classification task. Let $p^*(x, y) = p^*(y|x)p^*(x)$ be the true joint distribution of the data generating process. It is unknown and can only be observed through the training dataset $\mathcal{D}$, which consists of N i.i.d. samples $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$.

Furthermore, let $p_\theta(y|x)$ be a neural network, where the model parameters θ are optimized during the training using $\mathcal{D}$. For each input x, the model outputs a class prediction $\hat{y} = \text{argmax}_y p_\theta(y|x)$ and an associated confidence score $\hat{p} = \max_y p_\theta(y|x)$. Here, we do not focus primarily on the accuracy of the class predictions, but rather on an accurate estimation of the predictive uncertainties (i.e. of the confidence scores $\hat{p}$). We consider not only the in-distribution case, in which the test data originates from the same distribution $p^*(x, y)$ as $\mathcal{D}$ but also data shift scenarios, where the test data sample is from a different distribution $q(x|y) \neq p^*(x|y)$. More specifically, we study distribution shift scenarios, in which the distribution of the test data gradually shifts away from the true training distribution $p^*(x|y)$.

Calibration Metrics

Intuitively, a model is strongly calibrated if its outputs reflect reliably the predictive uncertainties. For example, if we consider all data samples having a confidence score $\hat{p} = 0.6$ for label y, we expect 60% of them to indeed take on the label y. More, formally we can define strong calibration such that accuracy and confidence match for all confidence levels (Gupta et al. 2021):

$$\mathbb{P}(\hat{y} = y|\hat{p} = p) = p, \quad \forall p \in [0, 1] \tag{1}$$

.

Based on this definition, a calibration error can be defined as the difference in expectation between accuracy and confidence. The most common estimator for this calibration error is the expected calibration error (ECE) (Naeini, Cooper, and Hauskrecht 2015). It calculates the average weighted gap between within-bin accuracy and within-bin predictive probabilities for S interval bins $B_s = \{n \in 1, \dots, N : p_\theta(y_n|x_n) \in (\rho_s, \rho_{s+1}]\}$, typically of equal size. Then, the ECE is calculated by

$$\text{ECE} = \sum_{s=1}^{S} \frac{|B_s|}{N} |\text{acc}(B_s) - \text{conf}(B_s)| \,, \tag{2}$$

where $\text{acc}(B_s) = \frac{1}{|B_s|} \sum_{n \in B_s} [y_n = \hat{y}_n]$ and $\text{conf}(B_s) = \frac{1}{|B_s|} \sum_{n \in B_s} p_\theta(\hat{y}_n|x_n)$.

Although the ECE is the most commonly used metric for measuring calibration, it has some drawbacks. It is not a proper score and for instance the chosen number of bins can result in biased estimates and/or volatility (Zhang, Kailkhura, and Han 2020). Therefore, we further calculate two alternative calibration metrics based on proper scores: Brier score (BS) and negative log likelihood (NLL).

Methods

Standard post-hoc calibration methods only utililze the original test image to determine well-calibrated predictions. The main idea of our approach is to obtain better calibrated confidence scores by applying data augmentation to the test data and recalibrating the resulting outputs with common post-hoc calibration methods.

First, a set of simple image transformations $\mathcal{T}$ (such as rotation, shift, changes in brightness) is defined. Each transformation has two parameters: 1) the probability that it will be performed and 2) the maximum magnitude of the transformation. The applied magnitude is randomly sampled out of the specified intervals during test time. The individual transformations are applied sequentially to an image x, such that the output of the composed function is a single transformed image $\tilde{x}$. Specific parameter settings used for all experiments are described in the next section.

For initializing the TTA-based extension for a given post-hoc calibration method, the base calibrator is tuned in a standard manner on the non-augmented validation set. During test time, not only the original test image is classified with the given model but also augmented versions of the test image modified with the image transformations $\mathcal{T}$. Based on the resulting logits of the individual predictions, the mean

Data	Meth.	Acc Base	Acc ours	ECE Base	ECE ours	BS Base	BS ours	NLL Base	NLL ours
C-10	TS	96.9±0.66	**97.1**±0.61	0.6±0.14	**0.4**±0.07	4.7±0.93	**4.3**±0.83	9.7±1.78	**8.8**±1.61
C-10	ETS	96.9±0.66	**97.1**±0.61	0.6±0.08	**0.5**±0.15	4.7±0.93	**4.3**±0.83	9.8±1.80	**9.0**±1.62
C-10	IR	96.9±0.66	**97.1**±0.61	0.8±0.09	**0.5**±0.09	4.8±0.93	**4.4**±0.83	13.8±2.16	**11.9**±2.04
C-10	IRM	96.9±0.66	**97.1**±**0.61**	0.6±0.12	**0.4**±0.10	4.8±0.93	**4.4**±0.83	9.9±1.81	**9.1**±1.58
C-100	TS	84.3±1.92	**85.2**±1.67	2.6±0.30	**2.3**±0.38	22.6±2.43	**21.5**±2.12	56.4±6.28	**52.9**±5.34
C-100	ETS	84.3±1.92	**85.2**±1.67	**1.9**±0.25	2.0±0.38	22.6±2.43	**21.4**±2.12	57.6±6.31	**54.2**±5.36
C-100	IR	84.3±1.92	**85.2**±1.67	2.6±0.30	**2.1**±0.26	23.4±2.50	**22.3**±2.19	87.6±7.63	**78.4**±6.37
C-100	IRM	84.3±1.92	**85.2**±1.67	**1.1**±0.31	1.3±0.34	22.7±2.49	**21.5**±2.14	58.1±6.56	**53.9**±5.45
Skin	TS	85.1±3.34	**86.9**±2.83	**2.7**±0.94	3.3±1.16	21.2±3.81	**19.3**±3.99	34.3±5.25	**31.7**±5.53
Skin	ETS	85.1±3.34	**86.9**±2.83	**2.8**±1.25	3.4±0.94	21.1±3.78	**19.3**±3.86	34.3±5.14	**31.7**±5.36
Skin	IR	85.1±3.34	**86.9**±2.83	**3.1**±0.99	3.8±1.68	21.2±3.40	**19.9**±4.12	41.6±7.66	**38.7**±9.69
Skin	IRM	85.1±3.34	**86.9**±2.83	**2.5**±0.88	3.3±1.64	21.2±3.68	**19.5**±3.97	35.8±5.40	**32.3**±5.58

Table 1: Performance comparison of the proposed TTA-based extensions to the corresponding standard post-hoc calibration methods on the in-distribution hold-out test set of CIFAR-10 (C-10), CIFAR-100 (C-100), and skin images (Skin). The mean and standard deviation of accuracy (Acc), expected calibration error (ECE), Brier score (BS), and negative log likelihood (NLL) over all considered architectures is shown. Better individual results (higher accuracy or lower calibration errors, respectively) are written in bold.

logit is calculated, which is subsequently used to compute the calibrated uncertainty score with the tuned post-hoc calibration method. The pseudocode of the proposed method is shown in Algorithm 1.

Recent theoretical work introducing a general bias variance decomposition (Gruber and Buettner 2022) allows us to motivate our approach from a theoretical perspective. In Gruber and Buettner (2022), the authors demonstrate that the classification log-likelihood can be decomposed such that the (tractable) Bregmann Information measures variance in logit space. They explicitly show that ensemble methods that average in the logit space - such as our proposed TTA-boosted calibration - provably reduce the variance of a classifier. The augmentations in our algorithm can thus be interpreted as a manifestation of one specific source of predictive noise (i.e. a representation of the source of domain drift due to changes in image acquisition systems, as in the skin lesion detection application); averaging the ensembled predictions in logit space will provably decrease the associated variance.

Experimental Setup

Model Architectures

We evaluate the calibration properties of deep neural networks using a variety of convolutional neural network architectures and sizes. We limit the focus on architectures which are commonly used in medical image classification.

The following 9 architectures are considered:

1. ResNets (He et al. 2016) with 18, 34, 50, 101, and 152 layers
2. DenseNets (Huang et al. 2017) with 101 and 169 layers
3. VGG architectures (Simonyan and Zisserman 2014) with 16 and 19 layers

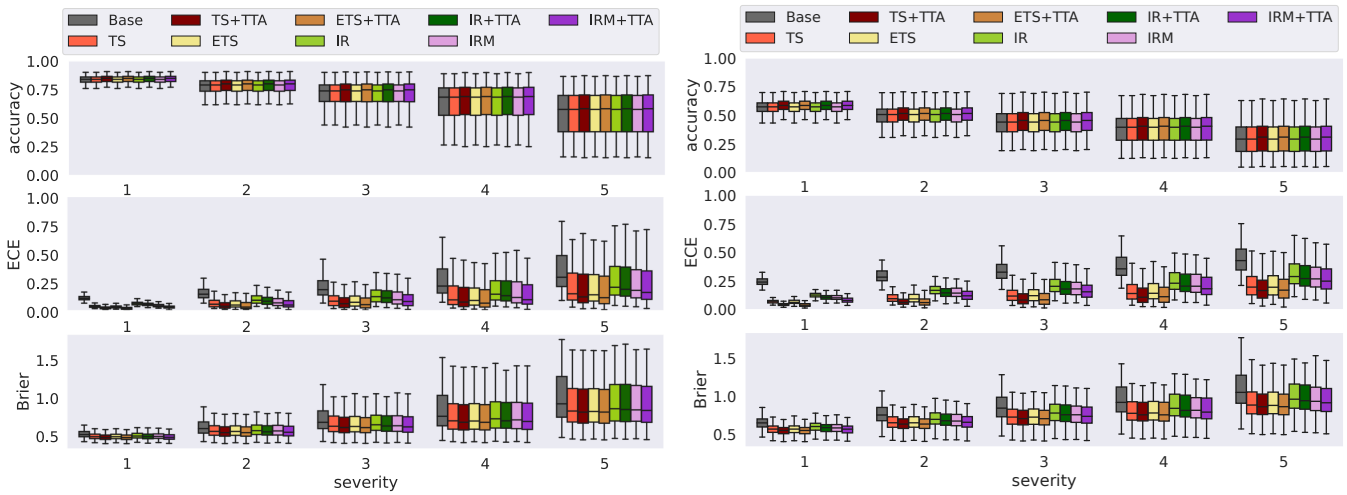
All models were initialized with standard IMAGENET pre-trained weights and then fine-tuned to the downstream task using transfer learning.

Datasets

We evaluate the proposed approach with both artificially generated distribution shift and real-world distribution shift scenarios.

Artificial distribution shift. First, we use two standard benchmark image data sets, CIFAR-10 and CIFAR-100. (Krizhevsky, Nair, and Hinton 2014). We randomly draw 20% of the samples from the original training set and use them as a validation set throughout the experiments. The reported results of the experiments are based on the standardized test data sets. For CIFAR-10 and CIFAR-100 the distribution shift is generated artificially by using 95 different corruptions (19 different types and 5 levels of severities each) proposed by Hendrycks and Dietterich (2019). Each corruption type mimics a distribution shift scenario in which the test data follows a distribution that gradually shifts away from the training distribution in a different manner.

Real-world distribution shift. For a more realistic distribution shift scenario, we further consider the classification of skin lesions. Here, we restrict ourselves to the binary classification task biopsy-verified melanoma (skin cancer) vs. biopsy-verified nevus (benign skin lesion such as a birthmark), which has a high practical relevance since it represents a challenging differential diagnosis, especially for inexperienced physicians. We use two open-source dermoscopic datasets for training the neural network: HAM10000 (Tschandl, Rosendahl, and Kittler 2018) and BCN20000 (Combalia et al. 2019). HAM10000 contains dermoscopic images (614 melanoma and 1155 nevi) which were acquired both at the University of Queensland and at the Department of Dermatology at the Medical University of Vienna. The



(a) Results on distribution shifted CIFAR-10.

(b) Results on distribution shifted CIFAR-100.

Figure 2: Accuracy, expected calibration error and Brier score of state-of-the-art post-hoc calibration methods and their corresponding proposed TTA-based extensions (+TTA) over 9 different model architectures.

dermoscopic images of BCN20000 (2832 melanoma and 2291 nevi) were collected at the Department of Dermatology at the Hospital Clinic of Barcelona. 1000 samples (500 per classes) each were randomly drawn for the validation and the hold-out test set. In addition to the in-distribution hold-out test set, two more datasets out of the 2020 ISIC Grand Challenge are used for testing the algorithm under real-world distribution shift (Rotemberg et al. 2021). We selected the sub-databases containing dermoscopic images from the Memorial Sloan Kettering Cancer Center in New York (216 melanoma and 614 nevi) as well as a collection of dermoscopic images originating from the Melanoma Institute Australia and the Sydney Melanoma Diagnosis Centre (134 melanoma and 161 nevi). Both external test data sets differ from the training set in terms of image acquisition systems and post processing steps (see Fig. 1).

Baseline Methods

The performance of the proposed approach is compared to the following baseline models: uncalibrated baseline model (**Base**), temperature scaling (**TS**) (Guo et al. 2017), ensemble temperature scaling (**ETS**) (Zhang, Kailkhura, and Han 2020), isotonic regression (**IR**) (Zadrozny and Elkan 2002), and the accuracy preserving version of isotonic regression (**IRM**) (Zhang, Kailkhura, and Han 2020).

Test Time Augmentation and Parameter Settings

Throughout the experiments, the configuration of the test-time augmentation methods are fixed. In particular, the selection of the individual image transformations used as well as their parameters are neither optimized for the data sets, model architectures, or severities of distribution shift. Instead, they are initialized with standard parameters that have proven suitable for many image classification tasks. This consisted of a random horizontal flip, rotation ($\pm 10°$), zoom (1.0-1.1), a change in brightness (± 0.1) and a symmetric warp (± 0.2). In all experiments, $T = 4$ test-time augmented images were generated and the reported expected calibration errors are based on $S = 20$ interval bins. The implementation of the experiments can be found on GitHub[1]

Results

Test Time Augmentation Improves In-distribution Post-hoc Calibration

First, we compared the proposed approach with common state-of-the-art post-hoc calibration methods on in-distribution hold-out test sets of CIFAR-10, CIFAR-100, and skin images. The in-distribution results are shown in Table 1. Regarding the calibration error, we analyzed the in-distribution results using three different metrics, namely expected calibration error, Brier score, and negative log likelihood. The experiments show that our proposed TTA-based extension consistently results in a lower Brier score and negative log likelihood on each of the three data sets and each every post-hoc calibrator. Moreover, in these cases, the standard deviation is also notably reduced with the proposed method. In terms of ECE, our TTA-based extension achieved competitive performance across all data-sets, with consistent improvement across all methods on CIFAR-10.

Test Time Augmentation Improves Calibration on Artificially Shifted Datasets

For a systematic in-depth analysis of the performance of TTA-boosted post-hoc calibration methods in distribution-shift scenarios, we first performed experiments with artificially distribution-shifted test sets of CIFAR-10 and CIFAR-100 (see Fig. 2). For this, 95 additional test sets each modified with a specific corruption of a given severity were gener-

[1]https://github.com/achimhekler/TTABoostedCalibration

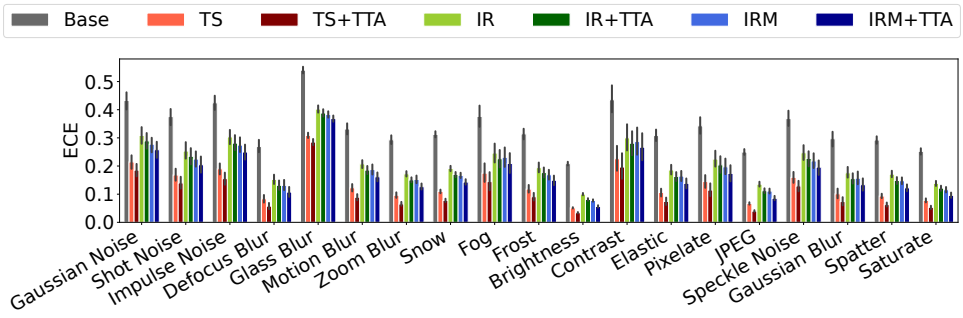


Figure 3: Averaged error across all severities of perturbations and models considered for CIFAR-100. Using TTA in combination with standard post-hoc calibration methods consistently improves ECE for any given perturbation and calibration method

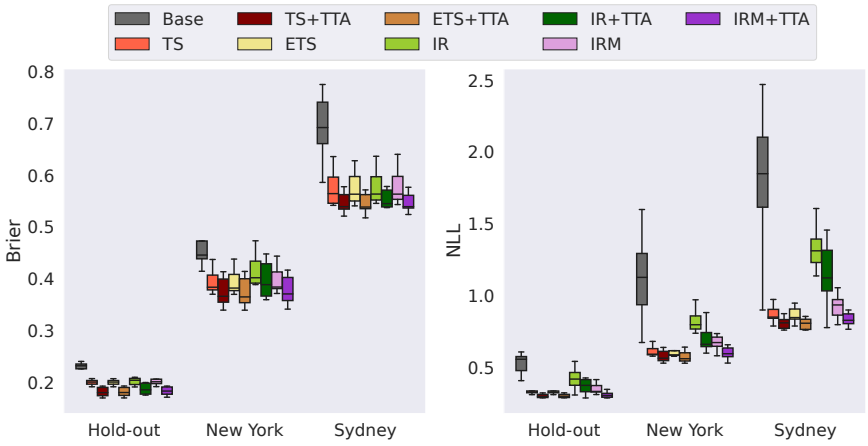


Figure 4: Brier score and negative log likelihood on the in-distribution hold-out test set and on the external datasets from New York (NY) and Sydney (SYD) over all 9 achitectures considered

ated. As expected, the accuracy drops strongly with increasing severity of distribution shift.

Considering ECE and Brier score, both calibration errors increase significantly with distribution shift severity (as also reported in (Ovadia et al. 2019)). However, the proposed TTA-based extensions consistently improve their respective baseline calibration methods with respect to both metrics. To provide further insights, we assessed the effect of our TTA-boosted post-hoc calibrators on ECE for each individual type of synthetic corruption (see Fig. 3). This revealed that TTA-boosted post-hoc calibration methods result in consistently better calibration even at the level of each individual perturbation.

Safer Decision-Making under Real-World Distribution Shift

Next, we investigated the performance of the proposed TTA-based extension on real-world distribution shifts in skin cancer detection. Fig. 4 shows the Brier score and the negative log likelihood on the in-distribution hold-out test set and on the external datasets from New York (NY) and Sydney (SYD) over all 9 architectures considered. As for the experiments with artificial distribution shift, Brier score and negative log likelihood increases for the two external datasets from New York and Sydney. But for all test data sets, our TTA-based extensions result in a substantial reduction of the two calibration metrics. The best results over all data sets were obtained with our proposed TTA-based extension of ETS.

TTA-boosted Calibration Improves Downstream Tasks

We consider the use of a confidence threshold as an exemplary downstream task, which is especially relevant for sensitive domains. Here, only individual predictions with a confidence greater than a given threshold are retained, all others are e.g. referred to an expert. In our example application of machine learning-based melanoma detection, an expert opinion from a dermatologist is consulted if the confidence score of prediction is lower the given threshold. This task is especially relevant under distribution-shift, since it is not a-priori clear how well a model will perform when data shifts occur during deployment.

Table 2 shows the result for confidence thresholding under distribution shift, for the New York data. Here, only the individual predictions of the ResNet34 trained on the skin images from Barcelona, Vienna, and Queensland with a confidence greater than a given threshold are only used for di-

		Threshold								
		0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
Accuracy	Base	0.712	0.713	0.716	0.718	0.719	0.726	0.728	0.733	0.743
	TS	0.715	0.726	0.730	0.741	0.748	0.752	0.772	0.806	0.838
	TS+TTA	0.737	0.750	0.752	0.752	0.771	0.787	0.812	0.832	0.857
no. of predictions	Base	816	806	795	786	773	760	742	711	666
	TS	794	760	725	676	612	540	464	330	130
	TS+TTA	787	743	714	674	612	540	451	310	119

Table 2: Results of ResNet34 for certainty thresholding on external test set from New York. Only individual predictions with a confidence greater than a given threshold are retained.

agnosis. All other predictions are discarded. While TS alone improves accuracy for any given threshold due to the improved calibration of the model, TTA-boosted calibration further results in a consistently higher accuracy compared to standard TS, for all confidence thresholds.

Our proposed TTA-based extension of TS shows an additional practical advantage in this task. On the one hand, it improves the accuracy by approximately 3% for each threshold. However, it is particularly interesting to note that the number of discarded predictions, does not increase compared to the TS-calibrated model. For example, for a threshold of 0.75 and 0.8, the number of rejected predictions is the same , whereas the corresponding accuracy is 2.3% and 3.5% higher for our proposed approach, respectively.

Data-Efficient Calibration

A key feature of TS-based methods for post-hoc calibration over other approaches is their high data-efficiency. Even for small validation sets, the optimal calibration parameter can be determined in a reliable and robust manner (Zhang, Kailkhura, and Han 2020). In order to examine the data-efficiency of the proposed TTA-based extensions, we conducted experiments using validation sets of different sizes to determine the optimal calibration parameters. For this purpose, we varied the size of the subsets from 10% to 100% of the respective standard validation set size. Analyzing the Brier score on the in-distribution test set, we observed that the calibration quality of the proposed TTA-based extensions of TS and ETS do not strongly depend on the size of the validation set (see Fig. 5): TTA-boosted TS and ETS are similarly data-efficient as their corresponding base methods. That is, our proposed TTA-boost maintains one of the key advantages of TS-based methods.

These findings are in contrast to non-parametric models IR and IRM. We found that calibration errors depend strongly on the size of the validation set and increased substantially with decreasing validation set size. This behavior can also be observed for the proposed extensions IR+TTA and IRM+TTA; however, the absolute calibration error after TTA-boosted post-hoc calibration is consistently lower for each validation set size compared to calibration with the respective base calibration methods IR and IRM.

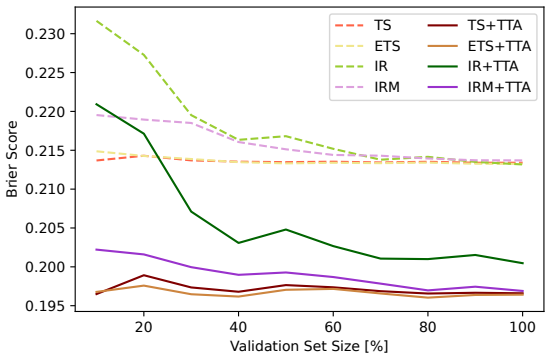


Figure 5: Brier score for different validation set sizes on the in-distribution test set of skin images averaged over the nine models

Conclusion

We present a simple and versatile approach for post-hoc calibration of neural networks that boost existing post-hoc calibration methods with test time augmentation. Our proposed TTA-based extension is easy to implement and can be applied to any given algorithm for post-hoc calibration, thereby making it accessible to practitioners working in safety-critical applications such as medical diagnostics.
Through extensive experiments across various data sets, model architectures and post-hoc calibration methods, a consistent improvement of the state-of-the-art is shown. In particular, we demonstrate on a real-world safety-critical application that TTA-boosted post-hoc calibration yields consistently better calibration also in real-world distribution shifts. This directly translates into a reliably better predictive power in a downstream application where only high-confidence predictions are retained.

Acknowledgements

This study was partially funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project 2). We thank Sebastian G. Gruber for insightful discussions on theoretical foundation of the work.

References

Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S.; Karthikesalingam, A.; King, D.; Ashrafian, H.; and Darzi, A. 2021. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ digital medicine*, 4(1): 1–23.

Brinker, T. J.; Hekler, A.; Enk, A. H.; Klode, J.; Hauschild, A.; Berking, C.; Schilling, B.; Haferkamp, S.; Schadendorf, D.; Holland-Letz, T.; et al. 2019. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113: 47–54.

Combalia, M.; Codella, N. C.; Rotemberg, V.; Helba, B.; Vilaplana, V.; Reiter, O.; Carrera, C.; Barreiro, A.; Halpern, A. C.; Puig, S.; et al. 2019. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

Combalia, M.; Hueto, F.; Puig, S.; Malvehy, J.; and Vilaplana, V. 2020. Uncertainty Estimation in Deep Neural Networks for Dermoscopic Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639): 115–118.

Grigorescu, S.; Trasnea, B.; Cocias, T.; and Macesanu, G. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386.

Gruber, S.; and Buettner, F. 2022. Uncertainty Estimates of Predictions via a General Bias-Variance Decomposition. *arXiv preprint arXiv:2210.12256*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330. PMLR.

Gupta, K.; Rahimi, A.; Ajanthan, T.; Mensink, T.; Sminchisescu, C.; and Hartley, R. 2021. Calibration of Neural Networks using Splines. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Javadi, G.; Bayat, S.; Kazemi Esfeh, M. M.; Samadi, S.; Sedghi, A.; Sojoudi, S.; Hurtado, A.; Chang, S.; Black, P.; Mousavi, P.; and Abolmaesumi, P. 2022. Towards targeted ultrasound-guided prostate biopsy by incorporating model and label uncertainty in cancer detection. *International Journal of Computer Assisted Radiology and Surgery*, 17(1): 121–128.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2014. The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5).

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Kull, M.; Silva Filho, T.; and Flach, P. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, 623–631. PMLR.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Ma, X.; and Blaschko, M. B. 2021. Meta-Cal: Well-controlled Post-hoc Calibration by Ranking. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 7235–7245. PMLR.

Maron, R. C.; Haggenmüller, S.; von Kalle, C.; Utikal, J. S.; Meier, F.; Gellrich, F. F.; Hauschild, A.; French, L. E.; Schlaak, M.; Ghoreschi, K.; et al. 2021. Robustness of convolutional neural networks in recognition of pigmented skin lesions. *European journal of cancer*, 145: 81–91.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33: 15288–15299.

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

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

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.

Prakash, A.; Moran, N.; Garber, S.; DiLillo, A.; and Storer, J. 2018. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8571–8580.

Rotemberg, V.; Kurtansky, N.; Betz-Stablein, B.; Caffery, L.; Chousakos, E.; Codella, N.; Combalia, M.; Dusza, S.; Guitera, P.; Gutman, D.; et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1): 1–8.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Thulasidasan, S.; Chennupati, G.; Bilmes, J. A.; Bhattacharya, T.; and Michalak, S. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Tomani, C.; and Buettner, F. 2021. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 3.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Wang, D.-B.; Feng, L.; and Zhang, M.-L. 2021. Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 11809–11820. Curran Associates, Inc.

Wang, G.; Li, W.; Aertsen, M.; Deprest, J.; Ourselin, S.; and Vercauteren, T. 2019a. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338: 34–45.

Wang, G.; Li, W.; Aertsen, M.; Deprest, J.; Ourselin, S.; and Vercauteren, T. 2019b. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338: 34–45.

Zadrozny, B.; and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, 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*, 694–699.

Zhang, J.; Kailkhura, B.; and Han, T. Y.-J. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, 11117–11128. PMLR.