Beyond accuracy : score calibration in deep learning models for camera trap image sequences

Gaspard Dussert[1], Simon Chamaillé-Jammes*[2], Stéphane Dray[1] and Vincent Miele[1]

[1]Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

[2]CEFE, Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France

November 9, 2023

Abstract

1. In this paper, we investigate whether deep learning models for species classification in camera trap images are well calibrated, i.e. whether predicted confidence scores can be reliably interpreted as probabilities that the predictions are true. Additionally, as camera traps are often configured to take multiple photos of the same event, we also explore the calibration of predictions at the sequence level.

2. Here, we (i) train deep learning models on a large and diverse European camera trap dataset, using five established architectures; (ii) compare their calibration and classification performances on three independent test sets; (iii) measure the performances at sequence level using four approaches to aggregate individuals predictions; (iv) study the effect and the practicality of a post-hoc calibration method, for both image and sequence levels.

*simon.chamaille@cefe.cnrs.fr

3. Our results first suggest that calibration and accuracy are closely intertwined and vary greatly across model architectures. Secondly, we observe that averaging the logits over the sequence before applying softmax normalization emerges as the most effective method for achieving both good calibration and accuracy at the sequence level. Finally, temperature scaling can be a practical solution to further improve calibration, given the generalizability of the optimum temperature across datasets.

4. We conclude that, with adequate methodology, deep learning models for species classification can be very well calibrated. This considerably improves the interpretability of the confidence scores and their usability in ecological downstream tasks.

Keywords : calibration, camera trap, classification, confidence score, event, machine learning

1 Introduction

Camera traps have become a central tool in the monitoring and conservation of communities and populations. They generate a lot of data that can be used to infer, for instance, species richness, occupancy or activity patterns (Sollmann 2018). To exploit these data, it is first required to identify the species present in the photos or videos. This manual annotation task is generally long and tedious, but it has been shown in recent years that it can be replaced by an automatic classification made by deep learning models, often with an accuracy of over 90% (Norouzzadeh et al. 2018; Willi et al. 2019; Whytock, Świeżewski, et al. 2021).

Accuracy may not be the only model performance metrics to care about though. Accuracy is calculated from, for each image, the prediction that has the highest confidence score (i.e. the top-1 prediction). In many ecological studies, downstream tasks may however directly rely on the confidence score of the predictions. This can be the case for instance when considering that values above a certain threshold indicate true detections, or when propagating model uncertainty into subsequent statistical models.

Importantly, confidence scores are frequently interpreted as probabilities of the prediction being true, but this is not always the case as many models may provide biased confidence scores (Gawlikowski et al. 2023). In the context of classification models, a model returning confidence scores that can be reliably interpreted as probabilities of the prediction being true is said to be well calibrated. For instance, if a model predicts the label of 100 images with a confidence score of 0.8, we would expect to observe an actual accuracy of 80% on these images. However, deep learning models trained with the categorical crossentropy loss, a common practice, are often over-confident and poorly calibrated (Gawlikowski et al. 2023). Attention should therefore also be given to the properties of confidence scores, as seen in other disciplines. For instance, good calibration of deep learning models has been shown to be important for safety-critical applications such as autonomous driving (Bojarski et al. 2016) or medical diagnosis (Nair et al. 2018). In the field of ecology, a good calibration ease the interpretation of the scores, but could also be critical if the scores are used in downstream tasks such as occupancy estimation (Gimenez et al. 2022), inference of species interaction (Parsons et al. 2022), real-time alert to guide law-enforcement (Whytock, Suijten, et al. 2023), and confidence-score-based prediction checking on citizen science platforms (e.g. Zooniverse (Simpson et al. 2014)).

Here we explore the calibration of confidence scores in the context of species classification models for camera trap data. In that context, the recurring leading approach, as assessed in recent iWildcam competitions (Beery, Agarwal, et al. 2021), consists in two steps: (step 1) detecting animals, humans and vehicles and filtering out empty images using a robust detection model such as MegaDetector (Beery, Morris, et al. 2019; Mitterwallner et al. 2023) and (step 2) using a CNN classification model to identify the species in the bounding box returned by the detection model, when an animal has been detected. We therefore focus on these species classification models (step 2), which are developed for a large range of species all over the world. We explore the interplay between accuracy and calibration for different state-of-the-art model architectures, using camera trap data from different sources. Also, we consider the calibration of confidence scores at the level of sequences of images. Indeed, camera traps are often configured to take multiple photos at each trigger, and predictions aggregated at the level of the sequence of images (sometimes called 'the observation' or 'event'). The issue of the calibration of confidence scores at the level of sequences of images has not, to out knowledge, been addressed in the literature. Furthermore, we study the relevance of a popular post-hoc calibration method called temperature scaling (Platt 2000), for both image and sequence levels. Finally, we provide a set of good practices for researchers and practitioners in the field.

2 Material and Methods

2.1 The DeepFaune Dataset

We use the dataset of the DeepFaune initiative (Rigoudy et al. 2023), which is a collaborative effort involving over 50 partners who, together, have gathered over two millions images and twenty thousand videos that they had manually annotated. These partners are affiliated to a wide range of institutions, such as organizations managing protected areas, hunting federations, and academic research groups. Images and videos were mainly collected in France, but also in a few European countries. Most of the annotation were at the species level, but some were at a higher taxonomic level (e.g. mustelid). Videos were converted into images by extracting frames of the first four seconds, with a time step of one second. The dataset provides a great diversity of habitats, elevations and weather conditions, as well as a wide variety of camera trap models with different settings, resolutions, flash type and image processing.

2.2 Training and validation datasets

For the species classification task, it is now known (Beery, Morris, et al. 2019; Norman et al. 2023) that two-step approaches (object detector followed by a classifier) are more efficient than classifiers that process the whole image. We use MegaDetector v5 (MdV5) (Beery, Morris, et al. 2019) to extract bounding boxes of animals, human and vehicles. Because MdV5 has already near-perfect accuracy on human and vehicles we only kept, for the training of our classifier, the bounding boxes that predicted the presence of an animal. For each bounding box, we created a cropped image of the original image, resulting in 429 347 cropped images of 22 different classes (the distribution of the classes is shown in Supporting Information Figure 1).

To avoid overfitting and shortcut learning between the background of the images (i.e. camera trap site) and the observed species, we designed the training and validation sets to have disjoint pairs of background and species while having the same balance of species and diversity of habitats. The validation set represented about 20% of the images available while being disjoint from the training set at the species level: for each species, the validation set is made of images originating from partners different than the ones used in the training set, while being as close as possible to a 80/20 split. This requires solving a problem of combinatorial optimization known as *subset sum problem*, which is a special case of the *knapsack problem* and which can be achieved using dedicated libraries (e.g. mknapsack). Ultimately, we had 368 786 images in the training set and 60 561 in the validation set.

2.3 Out-of-sample test sets

To demonstrate that the results of the classifier could generalize beyond the images collected in the DeepFaune initiative, 3 out-of-sample test datasets were used. These datasets originated from ecological programs conducted in three geographically distinct areas. We refer to these datasets by the name of the areas they originate from:

- **Pyrenees** : camera trap study in the national reserve of Orlu in the French Pyrenees, conducted by the French Biodiversity Agency (OFB), 100 266 images and 12 species after preprocessing.

- **Alps** : camera trap study in the Ecrins national park in the French Alps, conducted by S. Chamaillé-Jammes, 8 106 images and 12 species after preprocessing.

- **Portugal** : camera trap study in the Peneda-Gerês National Park in Portugal (Zuleger et al. 2023), publicly available. 99 750 cropped images and 16 species after processing.

2.4 Sequences of images

It is common to configure camera traps to take a series of images after each trigger. It is therefore relevant to have a single prediction for the whole series of images. We thereafter name such series 'sequences'. In our test sets, we considered that two consecutive images taken within 10s, at the same site (i.e. the same camera trap), belonged to the same sequence. We obtained sequences of 1 to 213 images.

2.5 Confidence score at sequence level

A sequence with S images has S individual predictions that can be aggregated in many different ways to produce a single prediction for the whole sequence. Formally, for a sequence of S images x_i, the model predicts the logits $z_i = (z_{i1}, ..., z_{iK})$ for each image, with K the number of classes. Confidence scores are derived using the softmax function : $p_i = (p_{i1}, ..., p_{iK}) = \text{softmax}(z_{i1}, ..., z_{iK})$. We aimed at predicting the confidence scores of the sequence $p_{seq} = (p_{seq1}, ..., p_{seqK})$ as a function of the predictions at the image level. We explored four different aggregation functions:

- **Average Score** : We averaged, over the sequence, the scores for individual pictures of the sequence:

$$p_{seq} = (\frac{1}{S}\sum_{i=1}^{S} p_{i1}, ..., \frac{1}{S}\sum_{i=1}^{S} p_{iK}) \tag{1}$$

- **Average Logit** : We averaged, over the sequence, the logits for individual pictures of the sequence, and then applied the softmax function:

$$p_{seq} = \text{softmax}(\frac{1}{S}\sum_{i=1}^{S} z_{i1}, ..., \frac{1}{S}\sum_{i=1}^{S} z_{iK}) \tag{2}$$

- **Max Score** : We kept the scores of the image that had the highest score amongst all scores of all images of the sequence:

$$p_{seq} = p_{i^*}, \text{ with } i^* = \underset{i\in[1,S]}{\arg\max}\{\max_{k\in[1,K]}\{p_{ik}\}\} \tag{3}$$

- **Max Logit** : We kept the scores of the image that had the highest logit amongst all logits of all images of the sequence:

$$p_{seq} = p_{i^*}, \text{ with } i^* = \underset{i\in[1,S]}{\arg\max}\{\max_{k\in[1,K]}\{z_{ik}\}\} \tag{4}$$

2.6 Calibration metrics

For a set of N images, we define the true class of the i-th image y_i and $p_i = (p_{i1}, ..., p_{iK})$ the confidences scores of the K classes. The predicted class $\hat{y}_i$ is the top-1 classification prediction, that is the class with the greatest confidence score, denoted s_i:

$$\hat{y}_i = \underset{k\in[1,K]}{\arg\max}\, p_i \quad \text{ and } \quad s_i = \max_{k\in[1,K]} p_i \tag{5}$$

For M evenly spaced bins, we can define b_m the set of indices i such as $s_i \in]\frac{m-1}{M}, \frac{m}{M}]$

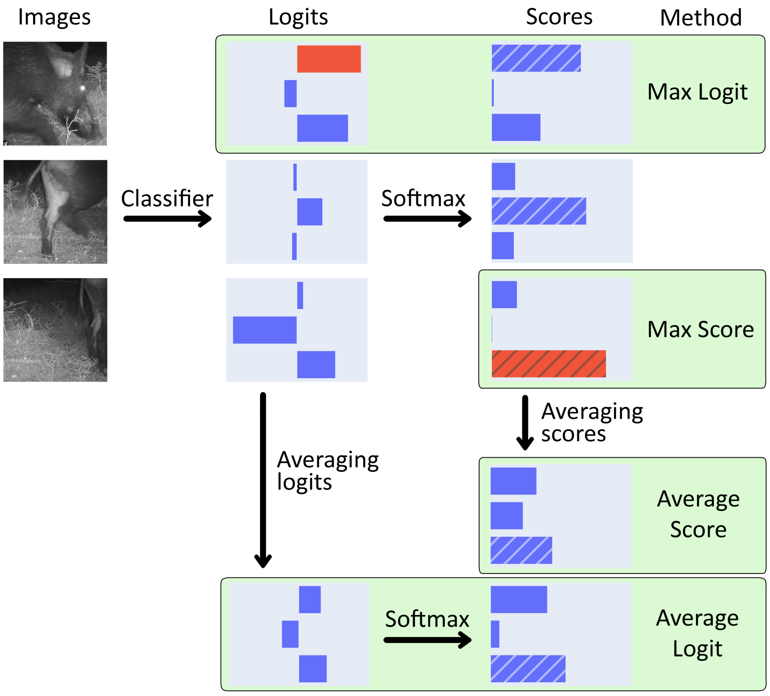

Figure 1: Illustration of the four aggregation methods. The greatest overall logit and score are in red. The top-1 score is hatched to emphasize that only this score is used to calculate the calibration.

and compute the average bin accuracy and the average bin confidence score :

$$\mathrm{acc}\left(b_m\right) = \frac{1}{|b_m|} \sum_{i \in b_m} \mathbb{1}\left(\hat{y}_i = y_i\right) \tag{6}$$

$$\mathrm{conf}\left(b_m\right) = \frac{1}{|b_m|} \sum_{i \in b_m} s_i \tag{7}$$

The bin-wise accuracy can be plotted to construct the reliability histogram (Guo et al. 2017) (e.g. Supporting Information Figure 3). It allows to visualize the calibration of a model: the closest the tops of the histogram bars are from the identity line, the better calibrated the model is. In addition, if the tops of the histogram bars are mostly above (resp. below) the line, the model is said to be under-confident (resp. over-confident).

The most common metric to measure the model's calibration quantitatively is the

Expected Calibration Error (ECE) (Guo et al. 2017). ECE is defined as the bin-wise calibration error weighted by the size of the bin :

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|b_m|}{N} \left|\mathrm{acc}\left(b_m\right) - \mathrm{conf}\left(b_m\right)\right| \tag{8}$$

Due to the large amount of images in our test sets, we decided to use a greater number of bins, specifically 20 instead of the standard 15, to obtain a more precise measurement of calibration with the ECE. In addition to this metric, we evaluated the classification performance of our classifier with the accuracy metric. These two metrics can also be used to evaluate the classification and the calibration at the sequence level, using the score p_{seq} and the associated predicted label $\hat{y}_{seq} = \underset{k \in [1,K]}{\arg\max}\, p_{seq}$.

2.7 Temperature Scaling

Temperature scaling (Platt 2000) is a post-processing method to improve the calibration of the model after the training. The scores predicted by the model are rescaled by a temperature parameter $T > 0$ using a generalization of the softmax function :

$$p_{ij} = \frac{\exp^{z_{ij}/T}}{\sum_{k=1}^{K} \exp^{z_{ik}/T}} \tag{9}$$

For $T = 1$ the scores obtained are the same as with the standard softmax function. $T > 1$ leads to lower scores and helps when the model is over-confident. Conversely, $T < 1$ increases the scores and helps under-confident models. For a given dataset, it is possible to determine the optimal temperature T^*, that minimize the ECE. However, this optimum temperature may differ from one dataset to another, and determining the optimum requires access to the labels. It is therefore unrealistic to use this individual temperature T^* to compare methods, as it cannot be calculated for a new dataset without manually annotating a fraction of the data. Instead, we propose to look at performance using a single temperature $\bar{T}$ shared across the three datasets. We define

$\bar{T}$ as the temperature that minimizes the average ECE across the 3 test datasets. Temperature scaling can be combined with the four aggregation method (Section 2.5) to calibrate sequence level predictions by simply replacing the standard softmax function with Equation 9.

2.8 Deep learning models

We used 5 established machine learning architectures: EfficientNetV2, ConvNext, ViT, Swin Transformer V2, and MobileNetV3. (Tan and Le 2021; Zhuang Liu et al. 2022; Dosovitskiy et al. 2021; Ze Liu et al. 2022; Howard et al. 2019). We have selected these architectures to represent CNNs (EfficientNetV2, ConvNext), Transformers (Swin, ViT), as well as lightweight architectures that could be deployed in camera traps that do the classification at the edge (MobileNetV3). Models were trained using the TIMM library (Wightman 2019) with transfer-learning from ImageNet-22k (Ridnik et al. 2021), the largest publicly available database. Data augmentation was applied using the imgaug library (A. B. Jung et al. 2020) using only standard transformations such as flips, crops, conversion to grayscale and affine transformation. The optimization was done using SGD, with a batch size of 32 and a different learning rate adapted for each architecture. To avoid overfitting, early stopping was used while monitoring the validation accuracy and with a patience of 10 epochs.

3 Results

3.1 Calibration at the image level

Generally, we observed that calibration (as measured by ECE) was negatively correlated with accuracy across models, for the 3 test datasets (Figure 2). ConvNext was the model providing the best overall performance. In particular, this model was slightly better in accuracy but much more efficient in terms of calibration (ECE of 2.37%, more than

2 times less than the second-best model, Swin Transformer V2, which has an ECE of 5.04%) on the Portugal dataset. In the meantime, the lightweight model, MobileNet, had bad to very bad (ECE of 34.27% on the Portugal dataset) accuracy and calibration performances.

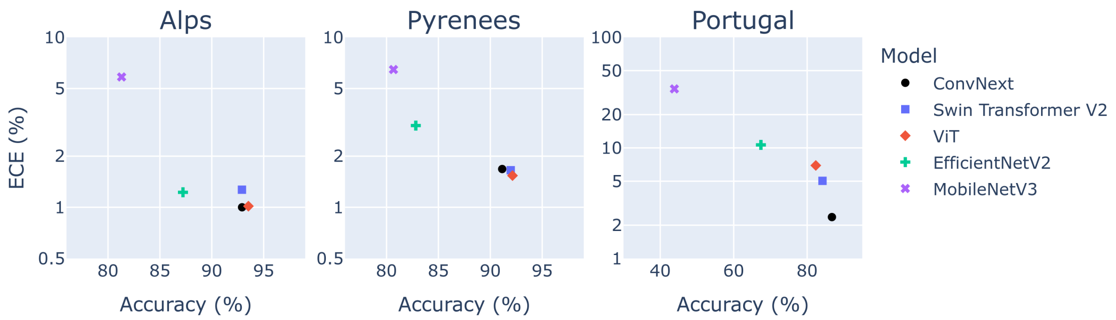

Figure 2: Scatterplot of ECE vs. accuracy values for five models (colored points) and three test data sets (panels), computed at the image level. Here, the ECE is not post-calibrated with temperature scaling (i.e. the temperature is 1 for all models).

As expected, temperature scaling allowed improving ECE values, for all models and datasets. We almost always observed a V-shape relationship between ECE and temperature, with ECE increasing quickly and by several percents around the optimum temperature value (Figure 3). This optimum temperature was generally greater than 1, suggesting that all models were initially overconfident to a greater or lesser extent. Interestingly, the V-shape curves of the different datasets overlapped well for the most accurate models (ConvNext and transformed-based models, ViT and Swin), and optimum temperature were similar across datasets. This suggested that a single optimum temperature would be sufficient to achieve efficient post-processing calibration. Indeed, using temperature scaling with temperature $\bar{T}$, the models exhibited on average a relative reduction in ECE of 38% compared to without temperature scaling ($T = 1$) (dashed line in Figure 3).

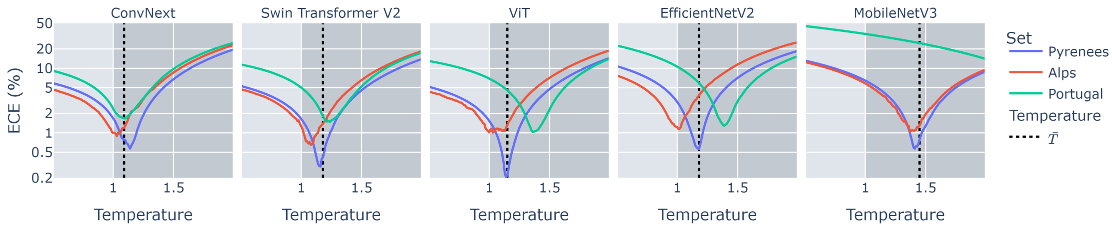

Figure 3: Curves of ECE values along the gradient of temperature values, for five models (panels) and three test data sets (colored curves). An optimum temperature below 1 indicates an underconfident model (light gray area), and above 1 indicates an overconfident model (dark gray area). The vertical dashed line shows $\bar{T}$, the temperature that minimized the average ECE across the 3 test datasets.

3.2 Calibration at the sequence level

Classification accuracy was much greater at the sequence level than at the image level (Figure 4 top). This was true for all models and all datasets, with up to +10% of accuracy for MobileNetV3 on the Portugal dataset. The Average Score and Average Logit were the two best methods for maximizing accuracy, with a slight advantage for the former. The gain in accuracy was however lower for models that were already efficient at the image level (ConvNext, ViT and SwinTransformer), but those remained the best models at the sequence level. Importantly, of the two aggregation methods that improved accuracy most, Average Score and Average Logit, only Average Logit provided well calibrated scores (Figure 4 bottom). The Average score was actually the worst aggregation method for calibration. Therefore, considering both accuracy and calibration metrics, the Average Logit was the best aggregation method.

Figure 4: ΔAccuracy (top, the greater the better) and ECE (bottom, the lower the better) for the four aggregation methods (colored curves) and five models (x-axis) on three test data sets (3 panels). ΔAccuracy is the difference between the accuracy at the sequence level and the accuracy at the image level.

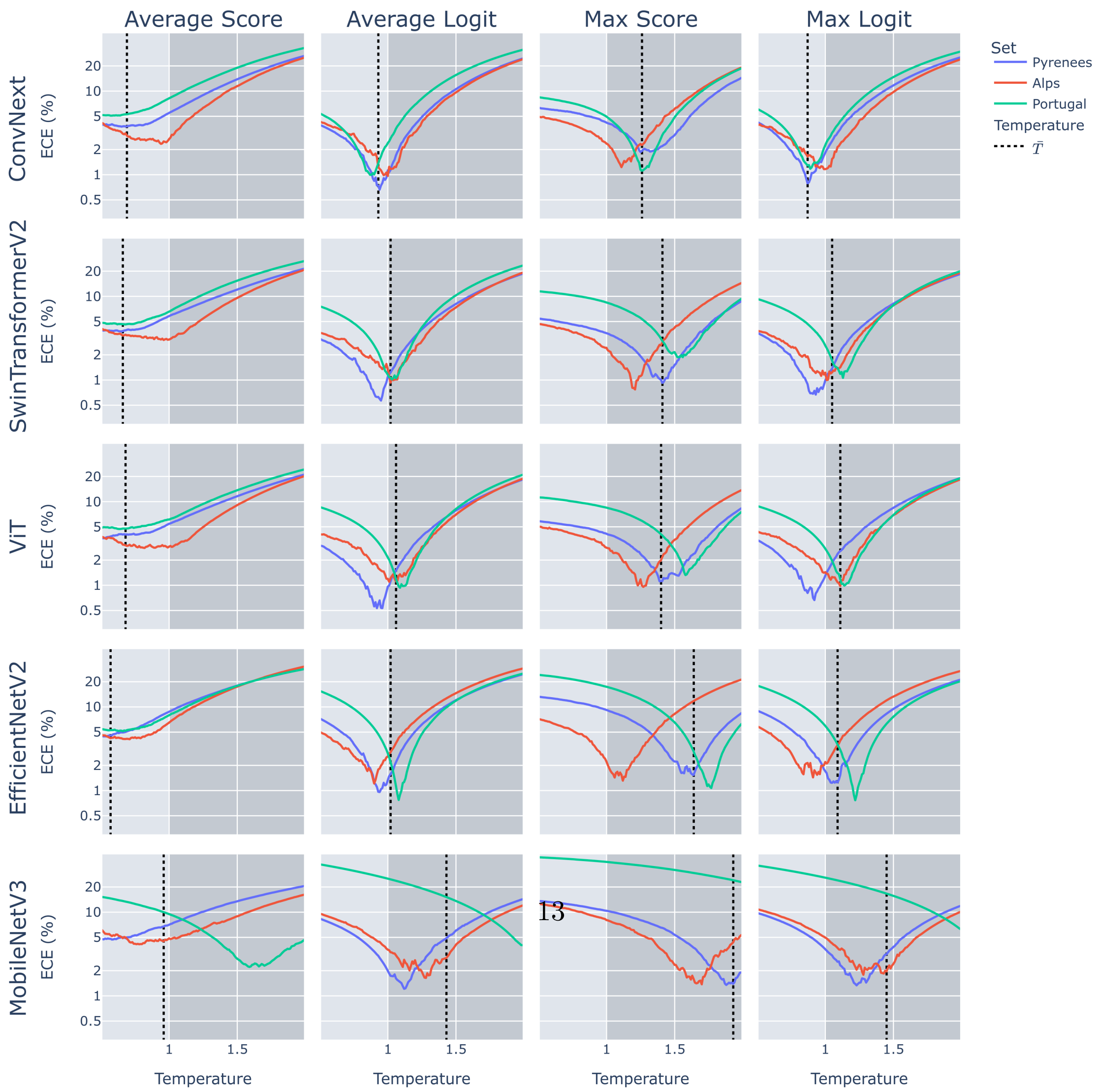

Figure 5: Curves of ECE values along the gradient of temperature values,for the four aggregation methods (columns), the five models (rows) and the three test datasets (colored curves). Light/gray area and dashed line defined as in Figure 3.

We finally studied the interplay between temperature scaling and aggregation methods. We observed that the aforementioned V-shape was more flat for the Average Score method than for the other methods (first column in Figure 5 versus the others). This confirmed that this method was the worst method, even with temperature scaling. We also noted that the Average Logit method provided the lowest ECE values overall (1.17% on average), and thus remained the best method, with temperature scaling further improving calibration at sequence level.

Looking at $\bar{T}$ and the optimum temperature of each set (minima and dashed lines on Figure 5), it can be noted that using the Average Score methods led models towards underconfidence, whereas using the Max Score methods led them towards overconfidence. This observation is also visible in the reliability histogram, as shown in Supporting Information (Supplementary Figure 3). Meanwhile, models using the Average Logit method displayed optimum temperatures close to 1 and, as a consequence, a temperature $\bar{T}$ close to 1 as well. We therefore concluded that the Average Logit method did not led to under- or over-confidence of models in our experiments. Also, and as observed at the image level, a single temperature (possibly close to 1) would be sufficient to achieve good post-processing calibration with the Average Logit method.

4 Discussion

This study assessed the calibration of confidence scores, at image and sequence level, for different deep learning models in the context of species classification in camera trap data. Using five state-of-the-art models and three out-of-sample test datasets, we showed that score calibration can vary greatly across model architectures, in a way that is consistent across test sets. Further, we showed that the different aggregation methods to obtain scores at the sequence level gave very different calibration values, and that the Average Logit method must be preferred over the others for optimizing both accuracy and calibration. Finally, we showed that temperature scaling can be used both at image level and sequence level, with a single temperature $\bar{T}$ that do not have to vary across datasets, to further improve the calibration. These observations pave the way for practitioners that are invited to 1/ monitor calibration as well as accuracy, 2/ use the Average Logit method and 3/ estimate the optimal temperature on their own test dataset and use it for the model deployment.

Differences in models' performance can be partly explained by model size. Indeed we found that models with the highest number of parameters (ConvNext, ViT, SwinTransformer) gave the best accuracy and ECE values. On the other hand, the only lightweight model, MobileNet, was consistently the worst model. Despite some literature showing that neural networks can be poorly calibrated, our result shows that this is not always the case (see also Minderer et al. (2021)), and that certain families of model architectures, such as ConvNext here, are intrinsically better calibrated than others, independently of the size of the model. The calibration of each model can be further improved on each dataset using temperature scaling as post-processing function. However, determining the optimal T requires annotating at least a fraction of the target set of images, which is something that practitioners would like to avoid if possible. Fortunately, we showed empirically with three very different datasets that the optimal temperatures are very close from one dataset to another, which suggests the generalizability of a single temperature that can be determined and fixed for future test sets. That said, we do not claim that the optimal temperatures defined in this paper can be used directly when using one of the studied architectures. Indeed, these temperatures are valid for a given training procedure (datasets, hyperparameters). In practice, it is mandatory to estimate the temperature using available test dataset(s) and subsequently maintain this temperature for deployment (since we showed it will be generalizable). This way, when the model will be used to classify new unseen data, the previously estimated temperature will ensure a better calibration of the predicted scores.

Proper model calibration at the image level is not always sufficient, as many software programs and scientific studies operate at the scale of the sequences that define the relevant 'observations' or 'events' from an ecological viewpoint. It is therefore extremely important to be able to calibrate the predictions at sequence level. For the first time, we showed that the most intuitive approach, in which scores are averaged, did not provide the best accuracy and had the worst calibration, with largely under-confident predictions. Interestingly, our findings can be confirmed by the analogy with ensemble models. These approaches use N models to make a prediction on *one* image, whereas we use N images to make a prediction with *one* model at the sequence level. Wu and Gales (2021) showed that for ensemble models, individual model calibration is not sufficient to yield a calibrated ensemble prediction, and that their own method, which is equivalent to Average Score approach also leads to under-confidence. Moreover, Rahaman and Thiery (2021) show that, thanks to this natural shift in the optimal temperature when models are ensembled, if the individual models were slightly overconfident ($T > 1$, as is often the case in deep learning) then the ensemble model was naturally calibrated ($T \sim 1$). Our results greatly support the use of the Average Logit method for aggregating individual scores at the sequence level. It shifts slightly the optimal temperature towards underconfidence, which counterbalanced the overconfident nature of deep learning networks, and resulted in sequence level prediction that are almost calibrated without post-processing. With Average Logit, it is still interesting to use temperature scaling to improve calibration as much as possible, especially given that the ECE minima are again very close to each other and allow a single temperature to be set.

In this work, we focused on temperature scaling and did not consider other methods that have been shown to sometimes improve calibration, such as label smoothing and mixup (Szegedy et al. 2015; Zhang et al. 2018). We did so because these two methods are actually debated, as several studies have showed that they can actually worsen calibration when combined with temperature scaling (Wang et al. 2023; Minderer et al.

2021). As Minderer et al. (2021) state, "label smoothing creates artificially underconfident models and may therefore improve calibration for a specific amount of distribution shift". Label smoothing also assumes that all incorrect classes are equally likely (Maher and Kull 2021), which is obviously problematic in ecology (e.g., a wrongly predicted roe deer is much more likely to be a red deer than a wolf). Mixup also deteriorates calibration properties of networks by creating non-realistic images in the training set and leading to substantial distributional shift (Rahaman and Thiery 2021; Gawlikowski et al. 2023).

We believe that our results could be of use to researchers and practitioners in the field of computer vision of camera trap images. Firstly, we encourage everyone to select the architecture of their model using not only accuracy but also by calculating the ECE. Secondly, we recommend using the Average Logit method to aggregate information at sequence level, as it performs very well in terms of accuracy and calibration. Finally, to use temperature scaling and make calibration even better, the optimum temperature can be calculated on a test dataset and kept for future datasets.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD010113729 made by GENCI.

We acknowledge the organizations and people contributing to the DeepFaune initiative.

Conflict of interest

None of the authors has a conflict of interest.

Author contributions

G.D., S.C.J., S.D. and V.M. conceived the ideas and designed the methodology. G.D., S.C.J. and V.M. gathered the training data. S.C.J collected the data of the Alps test set. G.D. and V.M. coded and performed the analysis. G.D. wrote the first version of the manuscript, S.C.J., S.D. and V.M. contributed critically to the drafts and gave final approval for publication.

Data availability statement

The five trained models, all derived data used in the analysis, and the code for the inference and metric calculation are available at https://doi.org/10.5281/zenodo.10014376. The Portugal and Alps datasets are available at https://doi.org/10.15468/rah33j and https://doi.org/10.5281/zenodo.10014376. The Pyrenees dataset is available upon request only, because of the presence of a sensitive species (brown bear).

References

Beery, Sara, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar (2021). *The iwildcam 2021 competition dataset.* arXiv:2105.03494 [cs.

Beery, Sara, Dan Morris, and Siyu Yang (July 2019). *Efficient Pipeline for Camera Trap Image Review.* arXiv:1907.06772 [cs]. DOI: `10.48550/arXiv.1907.06772`.

Bojarski, Mariusz, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba (Apr. 2016). *End to End Learning for Self-Driving Cars.* arXiv:1604.07316 [cs]. DOI: `10.48550/arXiv.1604.07316`.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (June 2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* arXiv:2010.11929 [cs]. DOI: 10.48550/arXiv.2010.11929.

Gawlikowski, Jakob, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. (2023). "A survey of uncertainty in deep neural networks". In: *Artificial Intelligence Review*, pp. 1–77.

Gimenez, Olivier, Maëlis Kervellec, Jean-Baptiste Fanjul, Anna Chaine, Lucile Marescot, Yoann Bollet, and Christophe Duchamp (Apr. 2022). "Trade-off between deep learning for species identification and inference about predator-prey co-occurrence". en. In: *Computo*. ISSN: 2824-7795. DOI: 10.57750/yfm2-5f45.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger (Aug. 2017). *On Calibration of Modern Neural Networks.* arXiv:1706.04599 [cs].

Howard, Andrew, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam (Nov. 2019). *Searching for MobileNetV3.* arXiv:1905.02244 [cs].

Jung, Alexander B., Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. (2020). *imgaug.* https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

Liu, Ze, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo (Apr. 2022). *Swin Transformer V2: Scaling Up Capacity and Resolution.* arXiv:2111.09883 [cs].

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie (Mar. 2022). *A ConvNet for the 2020s*. arXiv:2201.03545 [cs]. DOI: 10.48550/arXiv.2201.03545.

Maher, Mohamed and Meelis Kull (Oct. 2021). *Instance-based Label Smoothing For Better Calibrated Classification Networks*. arXiv:2110.05355 [cs].

Minderer, Matthias, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic (Oct. 2021). *Revisiting the Calibration of Modern Neural Networks*. arXiv:2106.07998 [cs].

Mitterwallner, Veronika, Anne Peters, Hendrik Edelhoff, Gregor Mathes, Hien Nguyen, Wibke Peters, Marco Heurich, and Manuel J. Steinbauer (2023). "Automated visitor and wildlife monitoring with camera traps and machine learning". In: *Remote Sensing in Ecology and Conservation* n/a.n/a. DOI: https://doi.org/10.1002/rse2.367. eprint: https://zslpublications.onlinelibrary.wiley.com/doi/pdf/10.1002/rse2.367.

Nair, Tanya, Doina Precup, Douglas L. Arnold, and Tal Arbel (Oct. 2018). *Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation*. arXiv:1808.01200 [cs]. DOI: 10.48550/arXiv.1808.01200.

Norman, Danielle L., Philipp H. Bischoff, Oliver R. Wearn, Robert M. Ewers, J. Marcus Rowcliffe, Benjamin Evans, Sarab Sethi, Philip M. Chapman, and Robin Freeman (2023). "Can CNN-based species classification generalise across variation in habitat within a camera trap survey?" en. In: *Methods in Ecology and Evolution* 14.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14031, pp. 242–251. ISSN: 2041-210X. DOI: 10.1111/2041-210X.14031.

Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune (June 2018). "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning". en. In: *Proceedings of the National Academy of Sciences* 115.25. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1719367115.

Parsons, Arielle W., Kenneth F. Kellner, Christopher T. Rota, Stephanie G. Schuttler, Joshua J. Millspaugh, and Roland W. Kays (2022). "The effect of urbanization on spatiotemporal interactions between gray foxes and coyotes". en. In: *Ecosphere* 13.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecs2.3993, e3993. ISSN: 2150-8925. DOI: 10.1002/ecs2.3993.

Platt, John (June 2000). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Adv. Large Margin Classif.* 10.

Rahaman, Rahul and Alexandre H. Thiery (Nov. 2021). *Uncertainty Quantification and Deep Ensembles*. arXiv:2007.08792 [cs, stat].

Ridnik, Tal, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor (Aug. 2021). *ImageNet-21K Pretraining for the Masses*. arXiv:2104.10972 [cs]. DOI: 10.48550/arXiv.2104.10972.

Rigoudy, Noa, Gaspard Dussert, Abdelbaki Benyoub, Aurélien Besnard, Carole Birck, Jérome Boyer, Yoann Bollet, Yoann Bunz, Gérard Caussimont, Elias Chetouane, Jules Chiffard Carriburu, Pierre Cornette, Anne Delestrade, Nina De Backer, Lucie Dispan, Maden Le Barh, Jeanne Duhayer, Jean-François Elder, Jean-Baptiste Fanjul, Jocelyn Fonderflick, Nicolas Froustey, Mathieu Garel, William Gaudry, Agathe Gérard, Olivier Gimenez, Arzhela Hemery, Audrey Hemon, Jean-Michel Jullien, Daniel Knitter, Isabelle Malafosse, Mircea Marginean, Louise Ménard, Alice Ouvrier, Gwennaelle Pariset, Vincent Prunet, Julien Rabault, Malory Randon, Yann Raulet, Antoine Régnier, Romain Ribière, Jean-Claude Ricci, Sandrine Ruette, Yann Schneylin, Jérôme Sentilles, Nathalie Siefert, Bethany Smith, Guillaume Terpereau, Pierrick Touchet, Wilfried Thuiller, Antonio Uzal, Valentin Vautrain, Ruppert Vimal, Julian Weber, Bruno Spataro, Vincent Miele, and Simon Chamaillé-Jammes (Oct. 2023). "The DeepFaune initiative: a collaborative effort towards the automatic identification of European fauna in camera trap images". In: *European Journal of Wildlife Research* 69.6, p. 113. DOI: 10.1007/s10344-023-01742-7.

Simpson, Robert, Kevin R. Page, and David De Roure (Apr. 2014). "Zooniverse: observing the world's largest citizen science platform". en. In: *Proceedings of the 23rd International Conference on World Wide Web*. Seoul Korea: ACM, pp. 1049–1054. ISBN: 978-1-4503-2745-9. DOI: 10.1145/2567948.2579215.

Sollmann, Rahel (2018). "A gentle introduction to camera-trap data analysis". en. In: *African Journal of Ecology* 56.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/aje.12557, pp. 740–749. ISSN: 1365-2028. DOI: 10.1111/aje.12557.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna (Dec. 2015). *Rethinking the Inception Architecture for Computer Vision*. arXiv:1512.00567 [cs].

Tan, Mingxing and Quoc V. Le (June 2021). *EfficientNetV2: Smaller Models and Faster Training*. arXiv:2104.00298 [cs].

Wang, Deng-Bao, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Min-Ling Zhang (2023). "On the Pitfall of Mixup for Uncertainty Calibration". en. In: pp. 7609–7618.

Whytock, Robin C., Thijs Suijten, Tim van Deursen, Jedrzej Świeżewski, Hervé Mermiaghe, Nazaire Madamba, Narcisse Mouckoumou, Joeri A. Zwerts, Aurélie Flore Koumba Pambo, Laila Bahaa-el-din, Stephanie Brittain, Anabelle W. Cardoso, Philipp Henschel, David Lehmann, Brice Roxan Momboua, Loïc Makaga, Christopher Orbell, Lee J. T. White, Donald Midoko Iponga, and Katharine A. Abernethy (2023). "Real-time alerts from AI-enabled camera traps using the Iridium satellite network: A case-study in Gabon, Central Africa". en. In: *Methods in Ecology and Evolution* 14.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14036, pp. 867–874. ISSN: 2041-210X. DOI: 10.1111/2041-210X.14036.

Whytock, Robin C., Jedrzej Świeżewski, Joeri A. Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa-el-din, Kelly Boekee, Stephanie Brittain, Anabelle W. Cardoso, Philipp Henschel, David Lehmann, Brice Momboua, Cisquet Kiebou Opepa, Christopher Orbell, Ross T. Pitman, Hugh S. Robinson, and Katharine A. Abernethy (2021). "Robust ecological analysis of camera trap data labelled by a machine learning model". en. In: *Methods in Ecology and Evolution* 12.6. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13576, pp. 1080–1092. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13576.

Wightman, Ross (2019). *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models. DOI: 10.5281/zenodo.4414861.

Willi, Marco, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson (2019). "Identifying animal species in camera trap images using deep learning and citizen science". en. In: *Methods in Ecology and Evolution* 10.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13099, pp. 80–91. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13099.

Wu, Xixin and Mark Gales (Jan. 2021). *Should Ensemble Members Be Calibrated?* arXiv:2101.05397 [cs, stat].

Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz (Apr. 2018). *mixup: Beyond Empirical Risk Minimization*. arXiv:1710.09412 [cs, stat]. DOI: 10.48550/arXiv.1710.09412.

Zuleger, Annika, Andrea Perino, Florian Wolf, Helen C. Wheeler, and Henrique M. Pereira (2023). "Long-term monitoring of mammal communities in the Peneda-Gerês National Park using camera trap data". In: *ARPHA Preprints* 4, ARPHA Preprints. DOI: 10.3897/arphapreprints.e99983. eprint: https://doi.org/10.3897/arphapreprints.e99983.

492 # Supplementary materials

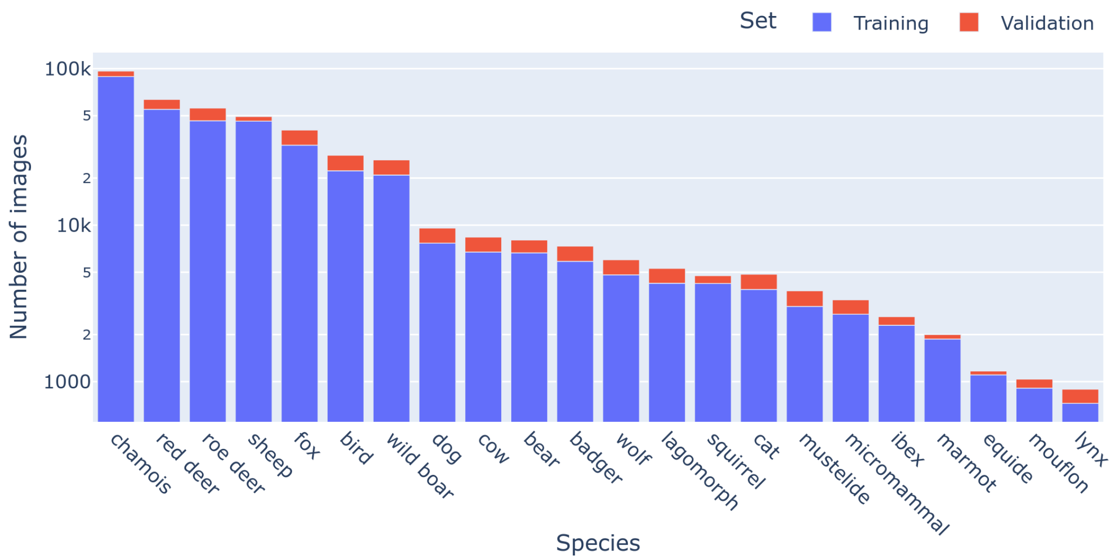

Figure 6: Number of images in the training and validation sets, for each species. Log scale is used to improve the readability of the rarer classes.

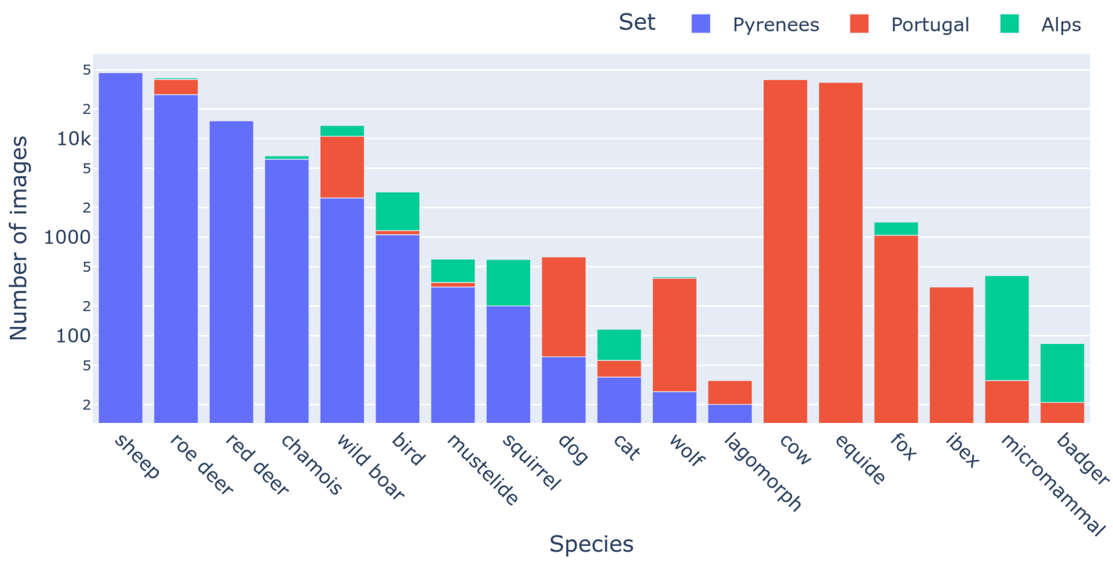

Figure 7: Number of images in the three out-of-sample datasets, for each species. Log scale is used to improve the readability of the rarer classes.

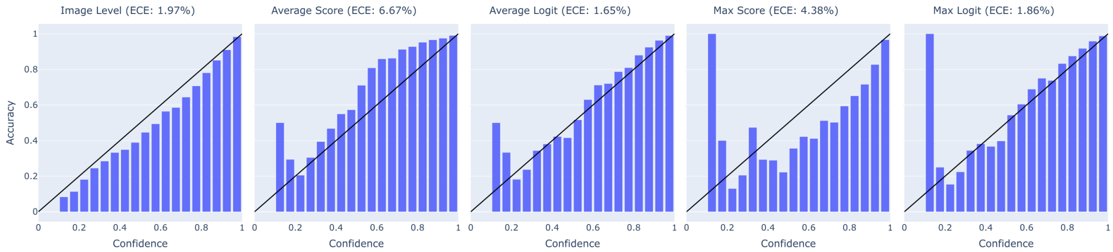

Figure 8: Reliability histogram of the ConvNext model, using the 3 test sets pooled together, and without temperature scaling.