IEEE Open Journal of the Computer Society

Received 14 November 2024; revised 9 December 2024; accepted 14 December 2024. Date of publication 18 December 2024; date of current version 9 January 2025. The review of this article was arranged by Associate Editor Jingcai Guo.

Digital Object Identifier 10.1109/OJCS.2024.3519984

Improving Accuracy and Calibration of Deep Image Classifiers With Agreement-Driven Dynamic Ensemble

PEDRO CONDE [1], RUI L. LOPES [2], AND CRISTIANO PREMEBIDA [1] (Senior Member, IEEE)

[1]Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, 3004-504 Coimbra, Portugal
[2]Critical Software, S.A. 3045-504 Coimbra, Portugal

CORRESPONDING AUTHOR: PEDRO CONDE (e-mail: pedro.conde@isr.uc.pt).

This work was supported in part by the Portuguese Foundation for Science and Technology (FCT), via the project *GreenBotics* under Grant PTDC/EEI-ROB/2459/2021, and in part by Critical Software, S.A.

ABSTRACT One of the biggest challenges when considering the applicability of Deep Learning systems to real-world problems is the possibility of failure in *critical* situations. Possible strategies to tackle this problem are two-fold: (i) models need to be highly accurate, consequently reducing this risk of failure; (ii) facing the impossibility of completely eliminating the risk of error, the models should be able to inform the level of uncertainty at the prediction level. As such, state-of-the-art DL models should be *accurate* and also *calibrated*, meaning that each prediction has to codify its confidence/uncertainty in a way that approximates the true likelihood of correctness. Nonetheless, relevant literature shows that improvements in *accuracy* and *calibration* are not usually related. This motivates the development of Agreement-Driven Dynamic Ensemble, a deep ensemble method that - by dynamically combining the advantages of two different ensemble strategies - is capable of achieving the highest possible accuracy values while obtaining also substantial improvements in calibration. The merits of the proposed algorithm are shown through a series of representative experiments, leveraging two different neural network architectures and three different datasets against multiple state-of-the-art baselines.

INDEX TERMS Deep ensembles, deep learning, image classification, reliability, probabilistic interpretation, uncertainty calibration.

I. INTRODUCTION

As Deep Neural Networks (DNNs) become ubquitous in critical fields such as autonomous driving, medicine, remote sensing, and robotics - where erroneous decisions can have significant repercussions - there rises a pressing need for Deep Learning (DL) models that not only exhibit high accuracy, but are also highly reliable when exposed to the variability of real-world data. It is therefore desirable that the DL models are capable to output predictions that codify trustworthy information. This underscores the importance of *uncertainty calibration* (also called *confidence calibration* or simply *calibration*), which ensures that the confidence scores generated by the DNNs accurately reflect the likelihood of correctness. This reliable quantification of predictive uncertainty will contribute to a probabilistically interpretable behaviour, when these models are applied to real-world data. Therefore assessing and improving the uncertainty calibration of modern DL systems, in diverse and representative scenarios, has become pivotal for the safe application of AI systems in real-world contexts.

As it will be formalized in Section III, the problem of uncertainty calibration focuses on the probabilistic correctness of a model's prediction. Intuitively, if a perfectly calibrated model makes 100 predictions that have a confidence value of 0.85, then 85 of those predictions will be in fact correct. A model that possesses this ability to accurately reflect how confident/uncertain it is in each prediction, is not only more reliable but also probabilistically interpretable, since - by leveraging calibrated confidence values - it is possible to quantify the level of uncertainty of the model for each prediction

and thus act accordingly within the context of the application. It is worth noting that modern DNNs have been found to be tendentiously uncalibrated, despite being increasingly accurate [7].

Understanding, evaluating and improving uncertainty calibration is essential for a trustworthy application of DNNs in real-world scenarios. This is particularly important in *critical* applications, where a malfunctioning AI system can lead to severe consequences. Remote sensing satellite systems are becoming increasingly vital in critical applications disaster prevention and response [15], [23], urban planning [24], resource management [21] and agriculture [5]. The variability in data induced by different sensors, resolutions, regions of acquisition and environmental factors, makes satellite image data a unique and challenging case study for evaluating the reliability of classification models. As such, given the critical aspects, the complexity and the real-world variability of this type of data, three different satellite image-based datasets were chosen as study case for the experiments presented in this article.

Within the various methodologies proposed in recent years to improve uncertainty calibration of modern DNNs, the use of Deep Ensembles [9] has been shown to be an effective strategy to improve both accuracy and calibration in DL-based classification problems [8], [14], [18], [20]. From the preliminary work developed in [3], it can be concluded that leveraging calibration metrics within an ensemble strategy can lead to better results in terms of model calibration, when compared to a traditional average-based ensemble algorithm. Nonetheless, this comes at the cost of a lower accuracy than what can be obtained with the average-based strategy. This type of *trade-off* - where calibration is improved at the cost of accuracy - is often undesirable, particularly in critical applications.

Contribution: Therefore, this article extends the work developed in [3] by proposing the **Agreement-Driven Dynamic Ensemble** (ADDE) algorithm, that explores a dynamic ensemble system to improve the uncertainty calibration while retaining the same high accuracy obtained by an average-based Deep Ensemble. The merits of the proposed approach are shown through a series of representative experiments.

II. RELATED WORK

Through the work presented in [7], the problem of uncertainty calibration is introduced to the DL community. The authors highlight that while modern deep learning architectures have achieved superior performance across various tasks, they tend to be less well-calibrated compared to "older" models. This is shown through a series of experiments, using multiple state-of-the-art DNNs, applied to different datasets from both computer vision, pattern recognition, and natural language processing applications. This issue related to DNN calibration is illustrated by leveraging the use of *reliability diagrams* and explicitly evaluated with the Expected Calibration Error (ECE) metric.

To address the calibration issues identified in [7], the authors employ *post-hoc* calibration methods. They introduce temperature scaling, a generalization of the Platt scaling algorithm [19], [17], which shows better performance on most datasets compared to other methods like histogram binning [27] and isotonic regression [28]. Since then, other *post-hoc* calibration methods have been proposed, like Dirichlet calibration [12], that is applicable to larger range of classification models (beside DNNs) and can, in some cases, outperform temperature scaling.

Besides *post-hoc* approaches, addressing DNN's *epistemic uncertainty* through the use of approximate Bayesian inference has shown also good results in improving the calibration of DL-based models [4], [6]. Given the computational complexity of full Bayesian Neural Networks (BNNs), the authors in [4] propose Monte Carlo dropout (MC dropout) as a form of approximate Bayesian inference for DNNs, that achieves similar results with a significant decrease in complexity.

Alternatively, ensemble methods operate on the principle of generating predictions from multiple deterministic models. When multiple deep neural networks (DNNs) are combined, this is known as a Deep Ensemble. Initially developed to enhance prediction accuracy [9], Deep Ensembles were later proposed as an alternative to Bayesian neural networks (BNNs) for achieving better-calibrated predictions [14]. Unlike BNNs, Deep Ensembles are simpler to implement and, while requiring more computational effort than a single DNN, often demand less than Bayesian approaches. Intuitively, a Deep Ensemble will approximate the probabilistic nature of a BNN by making multiple point estimations. Earlier versions of Deep Ensembles were inspired by bootstrap aggregating (bagging) techniques [1]; however, the authors in [14] demonstrate that training individual DNNs with random initialization is sufficiently effective in practice, while bagging techniques usually negatively affect accuracy and calibration. In fact, simply averaging the predictions of independently trained and randomly initialized DNNs has shown better results than more complex strategies, even outperforming BNNs and MC dropout in most cases [8], [14], [18], [20].

The work done in [3] proposes an ECE-based weighting and filtering/pruning ensemble strategy that effectively outperforms traditional (average-based) Deep Ensembles in different calibration-related metrics. Nonetheless, this improvement in terms of calibration comes at the cost of reduced accuracy, relatively to the average-based counterpart. As such, balancing the improvement in terms of calibration and accuracy is the main motivation behind the work subsequently presented in this article.

III. BACKGROUND

In this section, we formally discuss the problem of uncertainty calibration and introduce two standard evaluation metrics used in this domain.

Notation: with $\Delta_k = \{(p_1, \ldots, p_k) \in [0, 1]^k : \sum_{j=1}^{k} p_j = 1\}$ denotes a probability simplex for a classification problem with k different classes; we will use bold notation to denote

vectors, like $\mathbf{p} = (p_1, \ldots, p_k)$; the j-th element of some vector $\mathbf{p}$ will be referred as $\mathbf{p}_{\{j\}} := p_j$; $\sigma : \mathbb{R}^k \rightarrow \Delta_k$ represents the *softmax* function; we define *bag* (also called a *multiset*) as an extension of the notion of *set*, that can have repeated elements (i.e., different instances of the same element). These remarks are valid for all sections of the article; other remarks on notation will be given along the text, when found relevant.

A. DEFINING UNCERTAINTY CALIBRATION

In the literature, uncertainty calibration for DL systems in a multi-class context is typically defined in two primary ways. The most prevalent method, outlined in [7], extends the binary classification problem to a multi-class scenario using a *one-vs-all* approach, focusing solely on the calibration of the prediction with the highest confidence. Conversely, other studies, such as [27], adopt a more comprehensive definition that considers all the confidence values within the predicted probability distribution. As in [25], we differentiate between a *calibrated* model and a *strongly calibrated* model according to the definitions presented below. It is important to note that for binary classifiers, those definitions are equivalent.

Definition 1: **Calibrated model**

Let us consider a pair of random variables (X, Y), where X represents an input space (or feature space) and Y the corresponding set of true labels. Let us now take a model $f : X \rightarrow \Delta_k$ (this setting corresponds to a classification problem with k different classes). The model f is considered *calibrated* if

$$P\left(Y = \arg\max_{j\in\{1,\ldots,k\}} f(X) \,\middle|\, \max_{j\in\{1,\ldots,k\}} f(X)\right) = \max_{j\in\{1,\ldots,k\}} f(X). \tag{1}$$

Definition 2: **Strongly calibrated model**

Following the statements of Definition 1, the model f is considered *strongly calibrated* if, $\forall y \in \{1, \ldots, k\}$,

$$P\left(Y = y \,|\, f(X)_{\{y\}}\right) = f(X)_{\{y\}}. \tag{2}$$

B. EVALUATING UNCERTAINTY CALIBRATION

The probability values in the left hand side of both (1) and (2) cannot be computed using finitely many samples. Consequently, there is no *ground-truth* for the true likelihood values, which brings unique challenges when considering the evaluation of the calibration of the models. As such, there is a need for metrics tailored specifically for this type of evaluation. In this context, two variations of the well established ECE metric are described in this subsection, to address both "common" and strong calibration.

1) EXPECTED CALIBRATION ERROR

The ECE computes the bin-wise difference between the average confidence and the average accuracy of a given set of predictions. We start by creating the sets of bins $\{B_1, B_2 \ldots, B_M\}$, where each bin is a bag of confidence scores defined as

$$B_1 = \left\{p : \; p \in \max_{j\in\{1,\ldots,k\}} f(X)\right\} \cap [0, 1/M], \tag{3}$$

$$B_i = \left\{p : \; p \in \max_{j\in\{1,\ldots,k\}} f(X)\right\} \cap \,](i-1)/M, i/M], \tag{4}$$

for $i = 2, 3, \ldots, M$. We define the **confidence per bin** as

$$conf(B_i) = \frac{1}{|B_i|} \sum_{p\in B_i} p, \tag{5}$$

and the **accuracy per bin** as

$$acc(B_i) = \frac{1}{|B_i|} \sum_{p\in B_i} \delta(p), \tag{6}$$

where $\delta(p)$ equals 1 if the true class corresponds to the prediction and 0 otherwise. Finally, for a total of N predictions (i.e., N represents the number of samples used for evaluation) and a binning scheme $\{B_1, B_2, \ldots, B_M\}$, the ECE is defined as

$$\text{ECE} = \sum_{i=1}^{M} \frac{|B_i|}{N} |conf(B_i) - acc(B_i)|. \tag{7}$$

As is made clear by the binning scheme definition in (3), (4), the ECE addresses the calibration related to the maximum value of the prediction vector, being therefore associated with "common" calibration (Definition 1).

2) MULTI-CLASS ECE

To assess *strong* calibration (Definition 2) the previous metric can be modified by altering the space of confidence values that constitute each bin. Therefore we simply define a new binning scheme $\{B_1^*, B_2^* \ldots, B_M^*\}$, where each bin is a bag of confidence scores defined as

$$B_1^* = \left\{p : \; p \in f(X)_{\{j\}}, \forall j \in \{1, \ldots, k\}\right\} \cap [0, 1/M], \tag{8}$$

$$B_i^* = \left\{p : \; p \in f(X)_{\{j\}}, \forall j \in \{1, \ldots, k\}\right\} \cap \,](i-1)/M, i/M], \tag{9}$$

for $i = 2, 3, \ldots, M$. Now, for a new space of samples of dimension $N^* = kN$, the *multi-class* ECE (mc-ECE) can be defined as

$$\text{mc-ECE} = \sum_{i=1}^{M} \frac{|B_i^*|}{N^*} |conf(B_i^*) - acc(B_i^*)|. \tag{10}$$

Following the guidelines present in [13], we will only consider predictions with a confidence value above a given threshold (in our case 0.05), when computing the mc-ECE; this avoids a bias to the behaviour of low-confidence prediction (close to 0), that is common in this type of multi-class bin-wise metrics.

IV. PROPOSED METHOD

Before going into a detailed description of the proposed ADDE, we will start by making some introductory considerations about ensembles of DNNs.

For n models $f_{\theta_i} : X \to \Delta_k$ $(i = 1, 2, \ldots, n)$, where θ_i represent the set of parameters/weights of the model f_{θ_i}, a classical average-based approach to a Deep Ensemble (who has shown to be an effective method for improving uncertainty calibration [8], [14], [18], [20]), approximates the probability vector

$$\mathbb{P}(Y|X) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\left(Y|f_{\theta_i}(X)\right). \tag{11}$$

A natural extension of this concept is to consider a weighted average to combine to outputs of the different models. In this context, we take a vector of weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ and make the estimation

$$\mathbb{P}(Y|X) \approx h(\mathbf{w}) \cdot \left[\mathbb{P}\left(Y|f_i(X)\right]_{i\in\Theta}^{T}, \tag{12}$$

where $\Theta = \{\theta_i : i = 1, 2, \ldots, n\}$ and $h : \mathbb{R}^n \to \Delta_n$ is a normalization function. The vector $\mathbf{w}$ defines different biases with respect to different members of the ensemble. Additionally, a filtering/pruning mechanism can also be added by defining a set $\Gamma \subset \{1, 2, \ldots, n\}$ and $\tilde{\Theta} = \{\theta_i : i \in \Gamma\}$, therefore considering only part a subset of the members of the ensemble, approximating the probability vector in the form

$$\mathbb{P}(Y|X) \approx \tilde{h}\left([w_i]_{i\in\Gamma}\right) \cdot \left[\mathbb{P}\left(Y|f_i(X)\right)\right]_{i\in\tilde{\Theta}}^{T}, \tag{13}$$

where $\tilde{h} : \mathbb{R}^{|\Gamma|} \to \Delta_{|\Gamma|}$. This process of reducing the size of an ensemble through a given criteria is usually referred as *ensemble pruning* [16]. The preliminary work developed in [3] is an example of an DNN ensemble strategy of the form (13).

As previously referred, the main motivation behind this work is to leverage the key insights from [3] while improving the biggest disadvantage of that preliminary method, that is a decrease in accuracy when compared to an average-based Deep Ensemble. From that work it is possible to conclude that, while an average-based Deep Ensemble is a better *decider* (i.e., it is better at defining the predicted class, which will naturally result in higher accuracy values), the ECE-based Deep Ensemble is better at encoding the predictive uncertainty of most predictions. Therefore, ADDE is based in a simple yet effective strategy to combine both of these properties, by proposing a *dynamic convex combination* of an ensemble of type (11) and another of type (13). This type of *convex combination* will be based on two principles:

1) ADDE works *dynamically* i.e., the behaviour of the algorithm can differ from prediction to prediction.
2) The final prediction of ADDE is dependent on the nature of the *class-related agreement* between two parts: a) an ECE-based Deep Ensemble, b) an average-based Deep Ensemble.

The description of the ADDE algorithm follows in the next subsection. We start by formalizing the idea behind the method and then describe its practical application.

A. AGREEMENT-DRIVEN DYNAMIC ENSEMBLE

Let us consider a pair of random variables (X, Y), where X represents an input space and Y the corresponding set of true labels. Let us now take a set of n models $f_{\theta_i} : X \to \Delta_k$ $(i = 1, 2, \ldots, n)$. Let us also consider $\Gamma \subseteq \{1, 2, \ldots, n\}$ and $\Theta = \{\theta_i : i \in \Gamma\}$; for now, Γ can be considered a generic subset, as its nature will be defined when discussing the practical computation of ADDE. To simplify the notation let us take $\hat{p}_{\theta_i} = \max_{j\in\{1,\ldots,k\}} f_{\theta_i}(X)$, $\hat{y}_{\theta_i} = \arg\max_{j\in\{1,\ldots,k\}} f_{\theta_i}(X)$ and

$$\mathcal{D}_i = P(Y = \hat{y}_i|\hat{p}_i) - \hat{p}_i. \tag{14}$$

In this context, the ADDE algorithm will approximate the probability vector

$$P(Y|X) \approx \frac{\gamma}{n} \sum_{i=1}^{n} P\left(Y|f_{\theta_i}(X)\right) + (1-\gamma)\sigma\left(\mathbb{E}[\mathcal{D}_i]_{i\in\Gamma}\right) \cdot \left[P\left(Y|f_{\tilde{\theta}}(X)\right)\right]_{\tilde{\theta}\in\Theta}^{T}, \tag{15}$$

where γ equals 0 if

$$\arg\max_{j\in\{1,\ldots,k\}} \sum_{i=1}^{n} P\left(Y|f_{\theta_i}(X)\right) \tag{16}$$

equals

$$\arg\max_{j\in\{1,\ldots,k\}} \sigma\left(\mathbb{E}[\mathcal{D}_i]_{i\in\Gamma}\right) \cdot \left[P\left(Y|f_{\tilde{\theta}}(X)\right)\right]_{\tilde{\theta}\in\Theta}^{T}, \tag{17}$$

and 1 otherwise.

In the computation of ADDE, the values for $P(Y|f_{\theta_i}(X))$ are approximated by the members of the Deep Ensemble (independently trained DNNs with random initialization), whereas $\mathbb{E}[P(Y = \hat{y}_\Theta|\hat{p}_\Theta) - \hat{p}_\Theta]$ is approximated through the computation of the ECE on a validation set [7]. Additionally, the subset Θ is defined by selecting the $\tau \leq n$ elements of the Deep Ensemble that perform best regarding the referred ECE computation, where τ is a previously defined threshold parameter. In Algorithm 1 we describe the practical application of ADDE. Additionally, in Fig. 1 is illustrated an example of the ADDE inference process using an ensemble with 10 DNNs and threshold value of $\tau = 2$.

V. EXPERIMENTS AND RESULTS

The following experimental section is made by leveraging three different datasets - EuroSAT [11], Aerial Image Dataset (AID) [26] and NWPU-RESISC45 (RESISC45) [2] (details on each dataset can be found in the Appendix), as well as two different DNN architectures, EfficientNet-B0 [22] and ResNet-50 [10]. The referred datasets are randomly divided into 70%, 10%, 20% for training, validation and test, respectively. All the DNNs used - either as members of an ensemble or for single-network baselines - are trained in an analogous manner for 40 epochs, using an Adam optimizer with learning rate of 0.001 and random initialization of weights.

Besides the proposed ADDE, we used in the experiments two other types of Deep Ensembles: an average-based Deep Ensemble (that we will from here onwards refer solely as Deep Ensemble) and an ECE-based Deep Ensemble (ECE Ensemble) [3]. The experiments are made separately for each

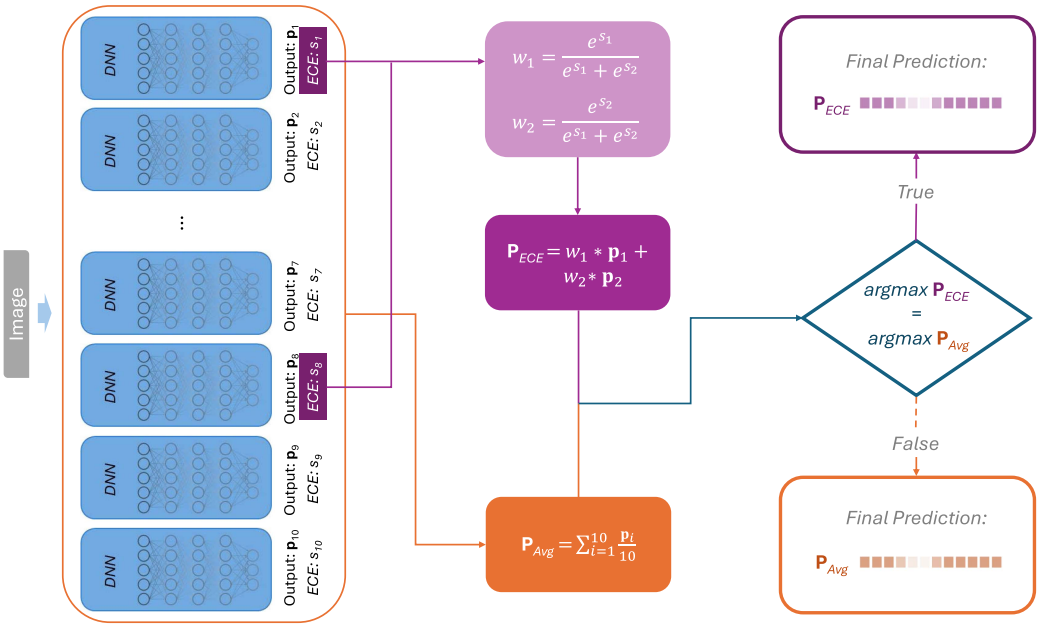


FIGURE 1. Graphical representation exemplifying the ADDE inference process, when using an emsemble of 10 DNNs and a threshold value of $\tau = 2$.

Algorithm 1: Agreement-Driven Dynamic Ensemble.

PRELIMINARIES:

Initialize n *DNNs* $f_{\theta_i^*} : X \to \Delta_k$ $(i = 1, 2, \ldots, n)$, *with random weights* $\theta_1^*, \theta_2^*, \ldots, \theta_n^*$.

Train each $f_{\theta_i^*}$ *independently in the training set.*
Output: *Trained DNNs,* $f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_n}$.

Using a validation set, evaluate each f_{θ_i} *regarding the ECE.*
Output: *ECE score for each DNN,* $s_1, s_2, \ldots, s_n$.

For $\tau \leq n$, *define* $\Gamma \subseteq \{1, 2, \ldots, n\}$, *as the subset of indexes of the* τ *best scores (lower values) in the set* $\{s_1, s_2, \ldots, s_n\}$.

Define $\Theta = \{\theta_i : i \in \Gamma\}$.

INFERENCE: (For each input $x \in X$)

$\mathbf{P}_{ECE}(x) = \sigma\left([s_i]_{i\in\Gamma}\right) \cdot \left[f_{\tilde{\theta}}(x)\right]_{\tilde{\theta}\in\Theta}^{T}$

$\mathbf{P}_{Avg}(x) = n^{-1} \sum_{i=1}^{n} f_{\theta_i}(x)$

if $\arg\max \mathbf{P}_{ECE}(x) = \arg\max \mathbf{P}_{Avg}(x)$ **then**
  $\mathbf{P}_{Final}(x) = \mathbf{P}_{ECE}(x)$

else
  $\mathbf{P}_{Final}(x) = \mathbf{P}_{Avg}(x)$

DNN architecture, meaning that each ensemble is composed by 10 randomly initialized DNNs of the same architecture. Additionally, some single-network baselines are also included in the experimental section, including a *vanilla* baseline (i.e., the DNN without any additional methodology applied) and different calibration methods: temperature scaling [7], L2-norm Dirichlet calibration (Dirichlet calibration) [12], ODIR Dirichlet calibration (ODIR calibration) [12] and MC-dropout [4]. MC-dropout is used by leveraging the existing dropout layers of the trained DNN; as such, this method is not included when using the ResNet-50 architecture, since the original model does not have dropout layers. The methods chosen in this article (besides ECE Ensemble which is the preliminary method that inspired ADDE) are widely used in the literature and have shown strong results in recent studies on uncertainty calibration of DNNs [3], [4], [7], [8], [12], [14], [18], [20].

In the following experiments, the threshold parameter τ of the ADDE is set as $\tau = 2$, following the recommendations made in [3] regarding the use of an ECE-based ensemble. Nonetheless, the nature of the ADDE algorithm opens the possibility for experimentation with different threshold parameter values. Naturally, the ECE Ensemble baseline follows the same guidelines.

In the following subsections, each experiment is described and its results discussed. All the results presented are obtained using the respective test set from each of the three datasets. In summary, each subsection can be categorized as follows:

Reliability diagrams: Improvements on calibration resulting from the application of ADDE are illustrated through reliability diagrams.

General quantitative results: The performance of ADDE in terms of accuracy, ECE and mc-ECE, is evaluated using the three datasets, with both DNN architectures, against multiple baselines.

Accuracy-calibration divergence from optimum: We evaluate how ADDE balances the improvements on both accuracy and calibration, compared to multiple baselines.

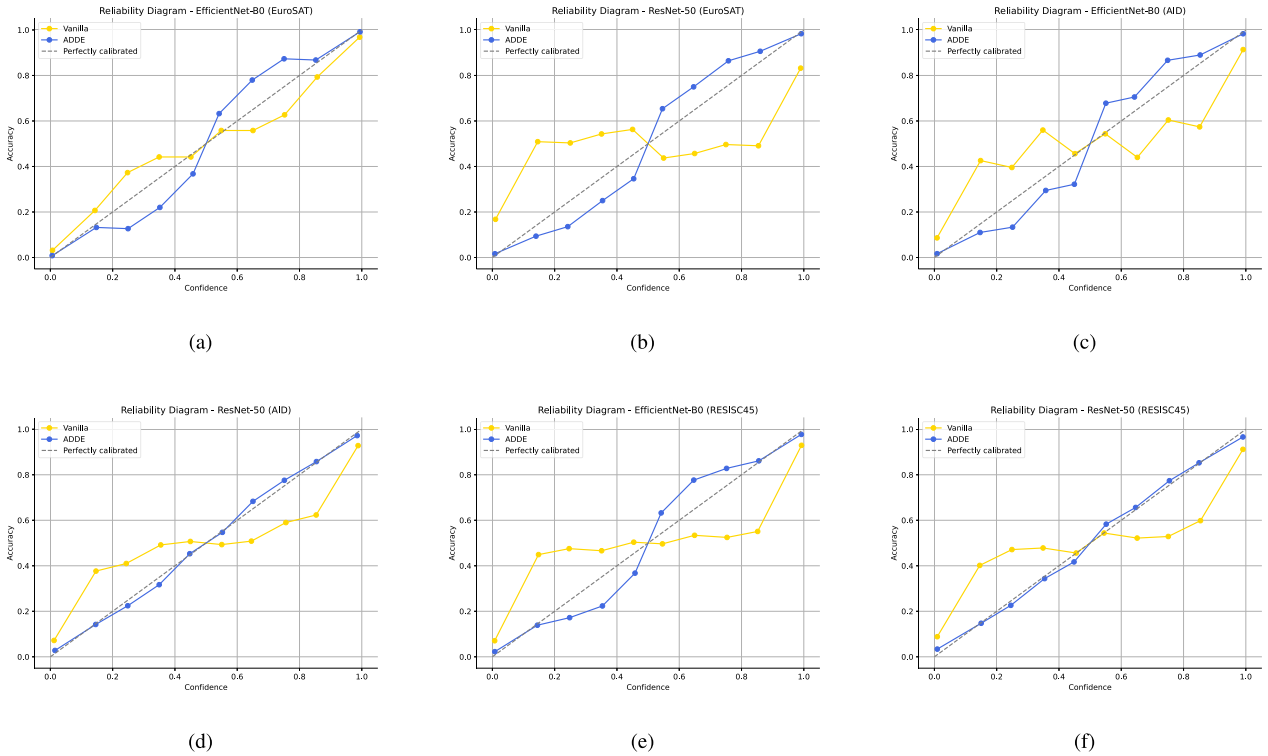


FIGURE 2. Reliability diagrams illustrating the graphical differences in terms of calibration between ADDE and a vanilla approach, using both EfficientNet-B0 and ResNet-50 DNN architectures, on the test sets of the EuroSAT, AID, and RESISC45 datasets.

Additional results: Additional experiments and respective results for deeper insights on the previous findings and on the behavior of ADDE.

A. RELIABILITY DIAGRAMS

In the present subsection, we leverage a graphical representation of calibration in the form of reliability diagrams - presented in Fig. 2 - to intuitively illustrate the effects of ADDE against a vanilla approach. Unlike a traditional reliability diagram, like those presented in [7] (that are presented in the form of bar charts) we take advantage of connected scatter plots for the graphical representation of calibration, since it allows for clearer comparisons.

To plot these diagrams, the interval [0, 1] is divided into 10 equally spaced bins, where the confidence values are distributed (similar to the computation of the ECE). Then, in each dot we plot the average confidence per bin against the average accuracy per bin. The confidence values used in this graphical representation are the maximum value and its counterpart ($1 -$ maximum) resulting in a *one-vs-all* representation for each prediction. In Fig. 2 we show, for each dataset and DNN architecture, the reliability diagrams of both a vanilla approach and the proposed ADDE; additionally, the expected behavior of a perfectly calibrated model is also represented.

Although an intuitive look into calibration, reliability diagrams can be misleading. For example, in Fig. 2(a) the behaviour of the vanilla DNN and the ADDE seem closely similar; however, since most predictions are situated at the initial and last bins, the difference in behaviour at the extremities will result in clear differences in terms of calibration (as will be clear in the following subsection). Nonetheless, it is already possible to illustrate - especially when considering the ResNet-50 architecture, i.e., Fig. 2(b), (d), and (f) - the improvements in calibration caused by ADDE.

Finally, it is possible to observe that while the *vanilla* outputs are are consistently over-confident, the ADDE confidence values are sightly under-confident, which is illustrated by the behavior of the respective curves in relation to the perfectly calibrated curve.

B. GENERAL QUANTITATIVE RESULTS

In Tables 1 and 2 are presented the values obtained for accuracy, ECE and mc-ECE in each experiment, allowing a comparison of the performance of ADDE against all the different baselines. We highlight in bold the best obtained value(s), while the values that are less than 2 *p.p.* from the best value are underlined (for accuracy this means less then 2 *p.p.* bellow the best value, and for ECE and mc-ECE it translates as less then 2 *p.p.* above that value). ECE and mc-ECE are computed using 15 bins.

We begin by comparing the two DNN architectures and their responses to different ensemble and single-network strategies. Without applying any additional techniques (vanilla), EfficientNet-B0 generally outperforms ResNet-50 in terms of both accuracy and calibration, except on the

TABLE 1. Results for ADDE and All the Different Baselines, Regarding Accuracy, ECE and mc-ECE, Using the EfficientNet-B0 DNN Architecture on the Test Sets of the EuroSAT, AID, and RESISC45 Datasets

EfficientNet-B0									
	EuroSAT			AID			RESISC45		
	Acc.	ECE	mc-ECE	Acc.	ECE	mc-ECE	Acc.	ECE	mc-ECE
Vanilla	93.31	3.18	3.33	83.65	9.78	9.04	85.44	8.82	8.57
Deep Ensemble	**96.65**	3.25	4.25	**93.90**	5.46	5.65	**91.65**	4.46	4.50
ECE Ensemble	95.39	1.41	1.84	89.90	2.66	3.00	88.95	2.68	3.60
T. Scaling	93.31	**1.05**	**1.25**	83.65	**2.31**	**1.68**	85.44	**1.66**	**1.34**
Dirichlet Calib.	93.44	8.67	7.15	86.85	10.96	6.49	85.14	12.13	6.42
ODIR Calib.	93.96	1.58	1.60	86.85	3.58	2.69	86.24	1.71	1.60
MC Dropout	93.26	3.15	3.49	83.55	9.98	9.37	85.41	8.87	8.53
ADDE	**96.65**	2.31	2.72	**93.90**	3.88	4.42	**91.65**	4.27	4.31

In bold are the best achieved performance scores while uncerlined ones are the values within 2 *p.p.* form the best value.

TABLE 2. Results for ADDE and All the Different Baselines (Except MC-Dropout), Regarding Accuracy, ECE and mc-ECE, Using the ResNet-50 DNN Architecture on the Test Sets of the EuroSAT, AID, and RESISC45 Datasets

ResNet-50									
	EuroSAT			AID			RESISC45		
	Acc.	ECE	mc-ECE	Acc.	ECE	mc-ECE	Acc.	ECE	mc-ECE
Vanilla	87.59	7.77	8.53	83.65	7.73	6.51	80.30	12.17	10.45
Deep Ensemble	**94.57**	7.76	7.77	**89.40**	9.49	7.66	**89.29**	5.73	5.42
ECE Ensemble	89.39	5.03	6.47	82.60	5.92	5.75	84.21	6.45	5.83
T. Scaling	87.59	**0.97**	**1.21**	83.65	**1.39**	**1.50**	80.30	3.59	2.14
Dirichlet Calib.	91.11	7.33	5.94	84.10	11.76	6.20	81.00	12.06	5.78
ODIR Calib.	91.33	1.08	1.13	85.10	3.55	2.71	82.81	**2.08**	**1.53**
ADDE	**94.57**	2.40	2.88	**89.40**	3.96	3.69	**89.29**	3.42	3.08

In bold are the best achieved performance scores while underlined ones are the values within 2 *p.p.* from the best value.

AID dataset. The accuracy advantage of EfficientNet-B0 becomes even more pronounced when various methodologies are applied. However, considering calibration, ResNet-50 - when paired with temperature scaling, ODIR calibration, and ADDE - can achieve comparable or even superior results to EfficientNet-B0. Despite this, the other two ensemble strategies (Deep Ensemble and ECE Ensemble) -which show decent calibration results with EfficientNet-B0 - perform significantly worse with ResNet-50.

Discussing now the differences found across the different datasets, we start by observing that the stronger results are achieved in the EuroSAT dataset, especially when considering accuracy. One important observation to take from this comparison is that the behaviour of the different methodologies is generally consistent across different datasets (i.e, there is no evidence of dataset-dependent behaviour), despite some observable differences.

Finally, we now discuss the performance of the proposed ADDE algorithm against the different baselines. In terms of accuracy, ADDE is consistently the best performing method, alongside Deep Ensemble; however, ADDE systematically outperforms Deep Ensemble in terms of calibration (in both ECE and mc-ECE evaluation). Temperature scaling and ODIR calibration are the methods that best perform in terms of calibration; nonetheless, these methods are fairly bellow ADDE when evaluating classification accuracy (temperature scaling does not alter the accuracy of a vanilla approach by design). We can finally observe that ADDE reaches an interesting balance between the improvements on accuracy and calibration, since it is in most cases substantially close to the best

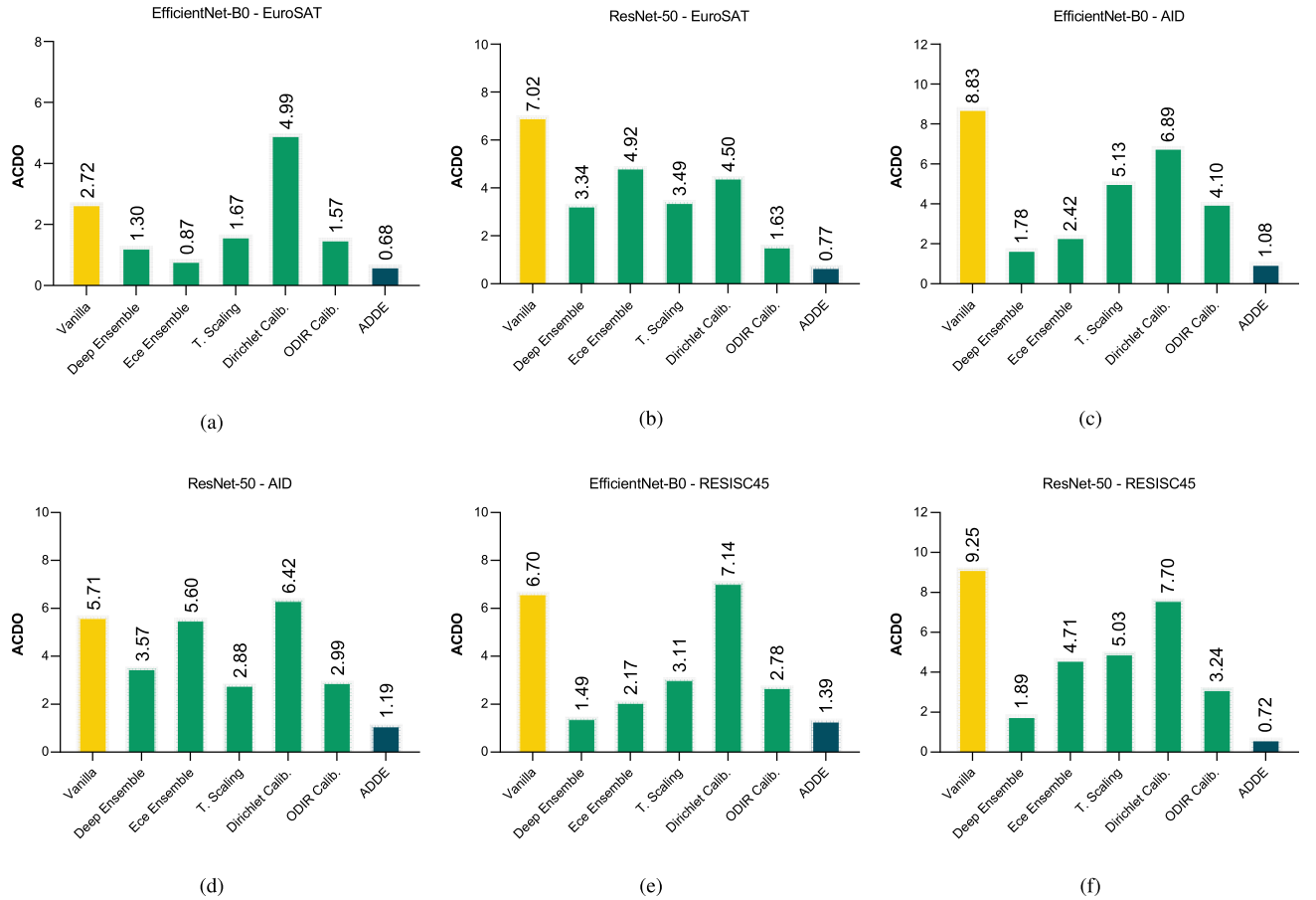


FIGURE 3. **Bar plots comparing the ACDO results for ADDE and multiple baselines, using both EfficientNet-B0 and ResNet-50 DNN architectures, on the test sets of the EuroSAT, AID, and RESISC45 datasets.**

calibration values, while always showing optimal accuracy. This balance between accuracy and calibration improvements is closely evaluated in following subsection.

C. ACCURACY-CALIBRATION DIVERGENCE FROM OPTIMUM

In this subsection we evaluate how the methods perform when considering both accuracy and calibration under the same metrics. For this purpose we propose the use of *accuracy-calibration divergence from optimum* (ACDO) that, for a given method, averages both the differences of accuracy and calibration from the optimal value (obtained on that experiment). Specifically, we first calculate how the accuracy result diverges from the best accuracy value in the experiment

$$\text{ADO} = \text{Acc.}_{best} - \text{Acc.}, \tag{18}$$

an then we perform analogous calculation for calibration, this time considering both ECE and mc-ECE,

$$\text{CDO} = \frac{(\text{ECE} - \text{ECE}_{best}) + (\text{mc-ECE} - \text{mc-ECE}_{best})}{2}. \tag{19}$$

Finally we average both these divergences

$$\text{ACDO} = \frac{\text{ADO} + \text{CDO}}{2}. \tag{20}$$

In this way, we evaluate the methods on their overall relative performance regarding both accuracy and calibration.

Fig. 3 presents the results for ACDO with EfficientNet-B0 (Fig. 3(a), (c) and (e)) and ResNet-50 (Fig. 3(b), (d) and (f)), on each of the three datasets. The value for MC-dropout is not presented in Fig. 3(a), (c) and (e) for the sake of symmetry, given that it is far from being a competitive result.

When using the EfficientNet-B0 architecture, it is observable that the ensemble-based methods achieve the best performance, with ADDE always obtaining the stronger results. On the other hand, with ResNet-50, ODIR calibration is capable of outperforming Deep Ensemble and ECE Ensemble (on most cases), but ADDE can still consistently obtain the best ACDO value by a comfortable margin. With these results, it is possible to observe the robustness of ADDE in terms of overall performance improvements, related to both accuracy and calibration.

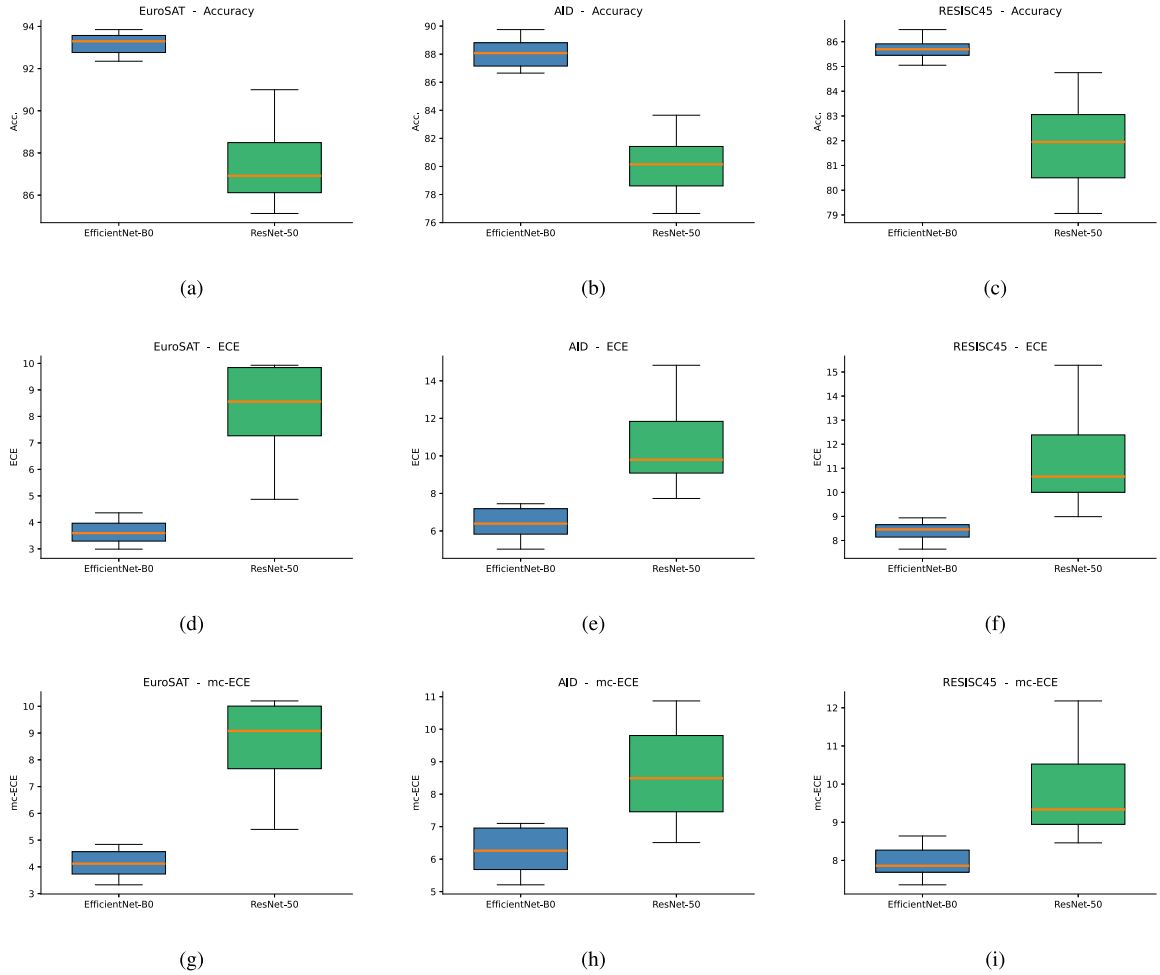


FIGURE 4. Box plots illustrating the distribution of the different ensemble members used, in terms of how they perform under different metrics (accuracy, ECE, and mc-ECE). The results are shown for both DNN architectures (EfficientNet-B0, ResNet-50) as well as the three datasets (EuroSAT, AID, and RESISC45).

D. ADDITIONAL RESULTS

We finalize Section V by outlining some additional results, that provide supplementary insights into our earlier findings and foster a deeper understanding of ADDE through ablation studies.

Fig. 4 illustrates how the DNN ensemble members - that constitute not only ADDE but also the average-based Deep Ensemble and ECE Ensemble benchmarks - distribute in terms of their performance, across the different experimental scenarios. Interestingly, despite the performance disparity between the DNN architectures observed in this plot, ADDE effectively closes this gap, achieving strong results with both architectures (as detailed in the previous subsections).

In Fig. 5, we illustrate the effect that the *threshold* parameter (referred as τ in Section IV), affects de performance of ADDE, by gradually increasing it. We care to note that given the dynamic nature of ADDE, accuracy is not affected by the different *threshold* parameter values and as such only the ECE and mc-ECE are evaluated. Although not consistently the best performing, a choice of $\tau = 2$ shows to be the best all-around choice (specially with ResNet-50), which is in line with the findings outlined in the preliminary work developed in [3] (with respect to ECE-based ensembles).

Finally, Fig. 6 illustrates the effect of reducing the number of DNNs used in ADDE, in terms of accuracy, ECE and mc-ECE, across the different experimental scenarios. Compared to the impact of the *threshold* parameter (Fig. 5) the ensemble size has a relatively weaker effect on calibration metrics, paired with only a minor effect on accuracy, (typically resulting in differences of less than 1 *p.p.*). Thus, although the effect of reducing the ensemble size is shown to be dependent on the data and model architecture, the impact is generally modest (and in some cases even positive), making this a viable option when computational resources are limited.

VI. FINAL REMARKS

This article proposes an ensemble based strategy for DNNs, designated by ADDE, that combines the advantages of both an average-based Deep Ensemble and a calibration-based Deep Ensemble, in terms of their improvements on accuracy and calibration, respectively. Through dynamic behaviour and by leveraging the class-related agreement of those two ensemble

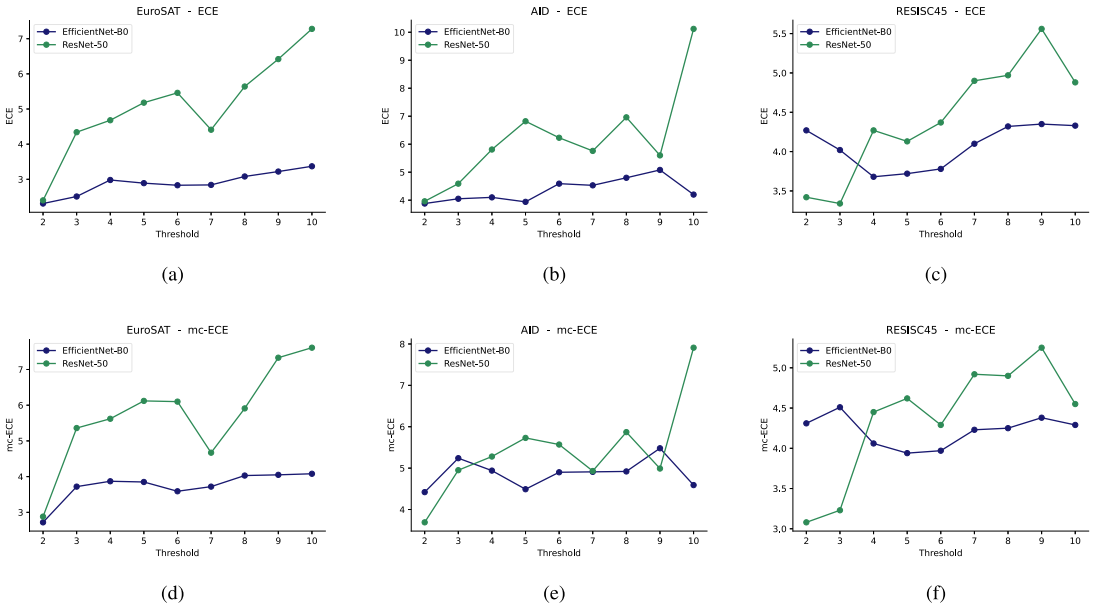


FIGURE 5. Connected scatter plots illustrating the effect that different *threshold* parameter values have on the performance of ADDE in terms of ECE and mc-ECE. The results are shown for both DNN architectures (EfficientNet-B0, ResNet-50) as well as the three datasets (EuroSAT, AID, and RESISC45).

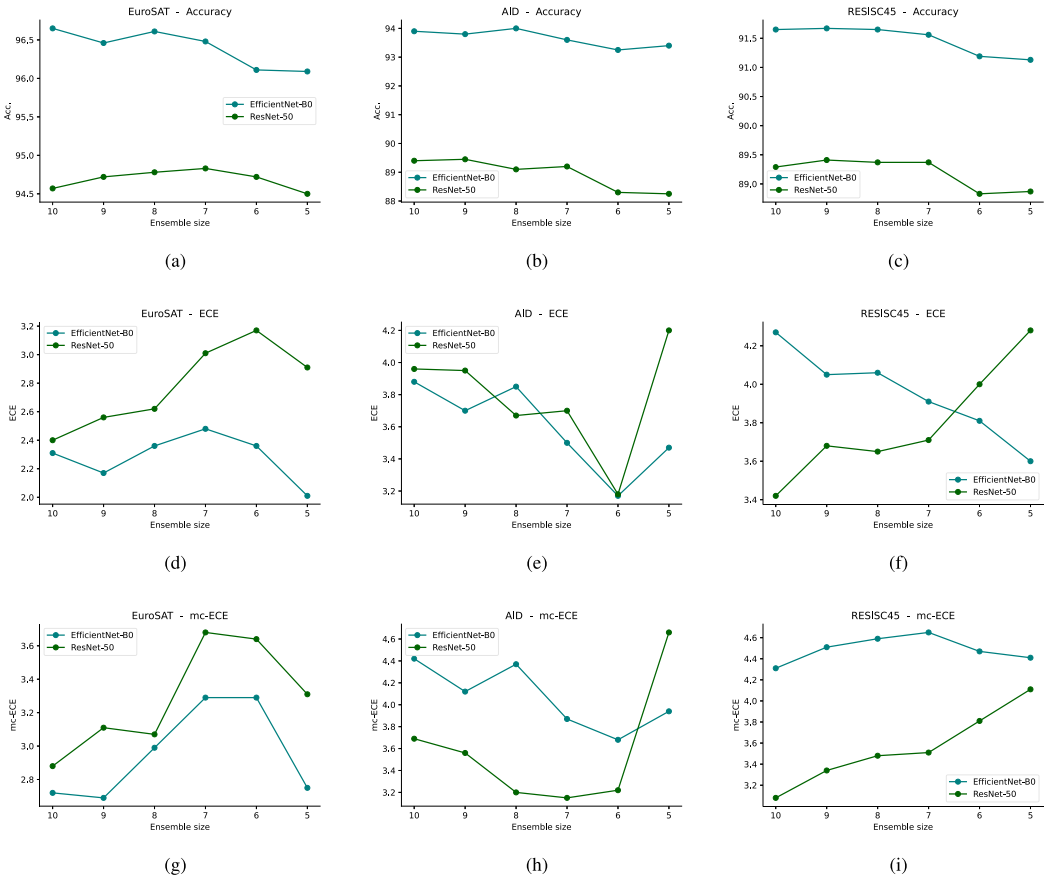


FIGURE 6. Connected scatter plots illustrating the effect that the ensemble size has on the performance of ADDE in terms of accuracy, ECE and mc-ECE. The results are shown for both DNN architectures (EfficientNet-B0, ResNet-50) as well as the three datasets (EuroSAT, AID, and RESISC45).

strategies, ADDE shows good results on calibration-related metrics, like ECE and mc-ECE, while always achieving the best accuracy values against multiple baselines.

From the experimental section, some general observations can be made that are consistent across the different experimental setups:

- In terms of accuracy evaluation, ADDE performs at the same level as average-based Deep Ensemble (by design), with both showing consistently the best results. Nonetheless, the improvements in calibration are significantly higher with ADDE than with the average-based counterpart.
- When solely discussing calibration evaluation, temperature scaling and ODIR calibration obtain the best performance. However, these methods have a significantly lower performance in terms of accuracy, when compared to the aforementioned ensemble strategies.
- When leveraging the ACDO, that evaluates the overall performance in terms of both accuracy and calibration, ADDE shows to be the most robust method, consistently outperforming all the other baselines.

It is important to additionally observe that, despite the overall robust performance of ADDE, this method (like all ensemble-based methods) has bigger computational costs than single-network strategies, since it requires the training of multiple DNNs. Nonetheless, we showed that reducing ADDE size can be a viable option in the case of computational restraints, resulting in small differences in performance.

Finally, it is possible to propose different avenues for future work. ADDE can be adapted to different problems like object detection or semantic/instance segmentation scenarios, by adjusting both of the ensemble strategies that (in cooperation) define the proposed algorithm, taking in consideration the particularities of each of these problems. On the other hand, even in pure classification scenarios, the main ideas behind ADDE can be leveraged to combine the properties of other ensemble strategies and be applied to contexts beyond both accuracy and calibration evaluation.

REFERENCES

[1] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[2] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[3] P. Conde, T. Barros, C. Premebida, and U. J. Nunes, "ECE-based deep ensemble for neural network calibration in satellite image classification," in *Proc. IEEE Int. Conf. Auton. Robot Syst. Competitions*, 2023, pp. 50–55.

[4] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[5] C. Giardino, M. Bresciani, P. Villa, and A. Martinelli, "Application of remote sensing in water resource management: The case study of lake trasimeno, Italy," *Water Resour. Manage.*, vol. 24, no. 14, pp. 3885–3899, 2010.

[6] A. Graves, "Practical variational inference for neural networks," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, vol. 24, pp. 2348–2356.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[8] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable Bayesian deep learning methods for robust computer vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 318–319.

[9] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct. 1990.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[11] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[12] M. Kull et al., "Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration," 2019, *arXiv:1910.12656*.

[13] F. Kuppers, J. Kronenberger, A. Shantia, and A. Haselhoff, "Multivariate confidence calibration for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 326–327.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6405–6416.

[15] S. Mangalathu and H. V. Burton, "Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions," *Int. J. Disaster Risk Reduction*, vol. 36, 2019, Art. no. 101111.

[16] G. Martinez-Munoz, D. Hernández-Lobato, and A. Suárez, "An analysis of ensemble pruning techniques based on ordered aggregation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 245–259, Feb. 2009.

[17] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 625–632.

[18] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 14003–14014.

[19] J. Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[20] R. Rahaman et al., "Uncertainty quantification and deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 20063–20075.

[21] K. E. Sawaya, L. G. Olmanson, N. J. Heinert, P. L. Brezonik, and M. E. Bauer, "Extending satellite remote sensing to local scales: Land and water resource monitoring using high-resolution imagery," *Remote Sens. Environ.*, vol. 88, no. 1/2, pp. 144–156, 2003.

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[23] S. Voigt, T. Kemper, T. Riedlinger, R. Kiefl, K. Scholte, and H. Mehl, "Satellite image analysis for disaster and crisis-management support," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1520–1528, Jun. 2007.

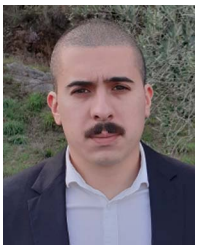
[24] T. Wellmann et al., "Remote sensing in urban planning: Contributions towards ecologically sound policies?," *Landscape, Urban Plan.*, vol. 204, 2020, Art. no. 103921.

[25] D. Widmann, F. Lindsten, and D. Zachariah, "Calibration tests in multiclass classification: A unifying framework," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 12257–12267.

[26] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[27] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers," in *Proc. Int. Conf. Mach. Learn.*, 2001, vol. 1, pp. 609–616.

[28] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2002, pp. 694–699.



PEDRO CONDE received the bachelor's degree in mathematics from the NOVA School of Science and Technology, Caparica, Portugal, and the master's degree in mathematics from the University of Coimbra, Coimbra, Portugal, where he is currently working toward the Ph.D. degree in electrical and computer engineering (specialization in automation and robotics). He is also a Researcher with the Institute of Systems and Robotics, Coimbra. His main research intersets include reliability, uncertainty calibration and probabilistic interpretation of deep neural networks.

RUI L. LOPES received the Ph.D. degree in AI from the University of Coimbra, Coimbra, Portugal, where he specialized in indirect representations for Genetic Programming. He is currently the Principal AI Engineer with DatasetLinks Inc. In 2007, he began his career with Critical Software, then developed end-to-end AI solutions for Operations Research with the European Space Agency. Following the Ph.D., he joined INESC-TEC, creating AI-driven solutions for healthcare, pharmaceuticals, oil and gas, and precision agriculture. After a decade in academia, he returned to Critical Software, working on AI applications for Fintech, Insurtech, Smart Buildings, among others. He leads AI initiatives with DatasetLinks Inc.

CRISTIANO PREMEBIDA (Senior Member, IEEE) is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal, where he is a member of the Institute of Systems and Robotics. He has been working on multimodal and multisensory perception for robotics and autonomous systems applications, developing calibration strategies and probability-prediction approaches to increase robustness of deep models. His research interests include autonomous systems, intelligent vehicles, robotic perception, machine learning, and sensor fusion.