A probabilistic scaling approach to conformal predictions in binary image classification

A probabilistic scaling approach to conformal predictions in binary image classification

Alberto Carlevaro[1,3,†] ALBERTO.CARLEVARO@IEIIT.CNR.IT
Sara Narteni[1,4,†] SARA.NARTENI@IEIIT.CNR.IT
Fabrizio Dabbene[1] FABRIZIO.DABBENE@CNR.IT
Teodoro Alamo[2] TALAMO@US.ES
Maurizio Mongelli[1] MAURIZIO.MONGELLI@CNR.IT

[1] *CNR-IEIIT, Corso Duca degli Abruzzi 24, 10129, Turin, Italy*
[2] *Departamento de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Escuela Superior de Ingenieros, 41020 Seville, Spain*
[3] *Aitek SpA, Funded Research Department, Via della Crocetta 15, 16122 Genova, Italy*
[4] *Politecnico di Torino, Department of Control and Computer Engineering, 10129, Turin, Italy*
† *A. Carlevaro and S. Narteni contributed equally to the development of the article. (Corresponding authors: S.Narteni, A. Carlevaro.)*

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Deep learning solutions for image classification are more and more widespread and sophisticated today, bringing the necessity to properly address their reliability. Many approaches exist in uncertainty quantification, and, among these, conformal prediction is one of most solid and well-established frameworks. In this paper, we study another approach, defined as *deep probabilistic scaling*, based on the notion of scalable classifiers, combined with probabilistic scaling from order statistics. Given a pre-trained neural network for (binary) image classification and a target class on which it is desirable to control the error, this method is able to bound that error to a user-defined level (ε). The method individuates probabilistic safety regions of target class samples correctly predicted in high probability. We show how the proposed method links with conformal prediction, discussing analogies and differences. By considering a (binary) convolutional neural network classifier, experiments on several benchmark datasets show a good overall performance of the methodology in controlling false negatives.

Keywords: uncertainty quantification, image classification, probabilistic scaling, conformal prediction.

1. Introduction

Uncertainty quantification is a crucial topic in machine and deep learning, playing a fundamental role in making models safe and trustworthy (Abdar et al., 2021). The need for reliable predictions is evident in applications where uncertainty can dramatically affect model safety: avionics, bioengineering, finance, autonomous vehicles, and healthcare are just a few examples where unreliable predictions can lead to serious consequences for users of an AI-based system. Although the issue of AI reliability is at the center of international debate and (preliminary) standards are being implemented (e.g., the recent and much-discussed

"EU AI Act" of the European Parliament[1]), most algorithms currently in use do not meet these standards. In the eyes of our research, therefore, there are two questions that need to be addressed: how can a machine learning algorithm be made more reliable and how can it be done without disrupting the architecture of the algorithm (so that it can also be applied to already existing algorithms). With this in mind, we have begun to address the problem of classification error reduction within the classical framework of binary image classification. We propose a probabilistic method, *deep probabilistic scaling (deep PS)*, based on the concept of *Scalable Classifiers* (Carlevaro et al., 2023). This approach can bound the number of

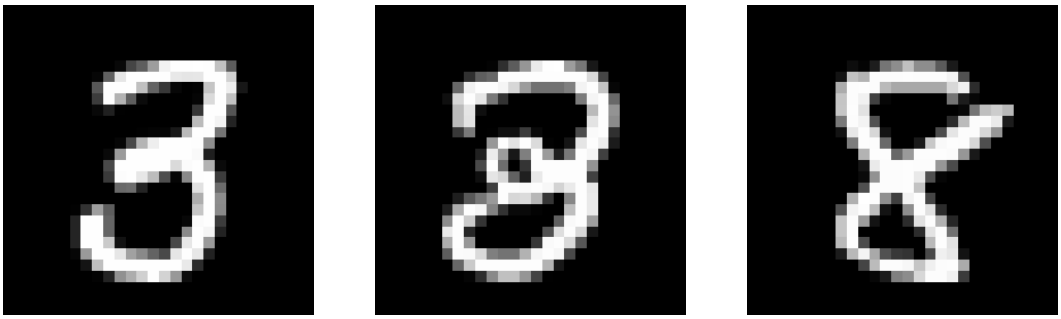


Figure 1: The goal is to understand which samples are "not conformal" with respect to a target class. Specifically, we claim nonconformity is a property of the data that the algorithm has to learn, necessitating an additional (a-posteriori) treatment.

false negatives (or false positives) without changing the structure of the network. In particular, our proposed solution is able to identify a special region in the input space, namely $\mathcal{S}_\varepsilon$, where in high probability the prediction corresponds with a chosen target class. Consider, for example, Figure 1: the numbers 3 and 8 in the digit-MNIST dataset can also be easily confused by the human eye, and this is understandable because the digits have different levels of similitude with each other. We can naively translate this idea of similitude into an "intrinsic probability" that the samples have: the digit on the left edge of the image have a high probability of being a 3 while the middle digit, while labeled as a 3, evidently has more features in common with an 8 and thus a lower probability of being identified as a 3. Our method, which we will describe thoroughly in the next sections, encodes precisely the concept that prediction is only a level of probability. A probability that can be controlled by appropriately shifting the classifier and identifying *regions* where the prediction result is highly "$1-\varepsilon$" certain. Such a design of safety regions with controlled error on a target class proves useful in many practical contexts where errors of the classification could have bad consequences (e.g., missing clinical diagnosis or failing in avoiding collisions): indeed, it will allow the monitoring of the input images and acknowledging when the image classification model will perform well (i.e., when the image falls in the region) or not on those inputs.

1. https://artificialintelligenceact.eu/the-act/

1.1. Contribution

The paper studies a new approach to probabilistically bound the model error in image binary classification and shows how it relates with conformal prediction. The main achievements relate to: error bound within algorithm design, finding proper definitions of score functions as well as exploiting probabilistic scaling in conformal prediction.

1.2. Structure of the paper

The remaining of the paper is structured as follows: Section 2 reports the main literature approaches in the field of uncertainty quantification in deep learning; Section 3 introduces the fundamentals of scalable classifiers and probabilistic scaling; Section 4 presents the idea of deep probabilistic scaling and describes the algorithmic steps to perform it; Section 5 describes the theoretical links of the proposed approach with conformal prediction theory; finally, the application of deep probabilistic scaling on benchmark image classification datasets is presented and discussed in Section 6, while Section 7 concludes the paper.

2. Related Work

In recent years, the boosting of highly performing deep learning methods has achieved great results in solving a large number of real-word problems and applications; however, these models generally do not provide information about the reliability of their predictions. For this reason, an important research subject consists in elaborating methods to quantify the epistemic uncertainty of deep models: recent works by Gawlikowski et al. (2023); Abdar et al. (2021) review the main current methods, that can be broadly grouped in Bayesian techniques (e.g., Monte Carlo dropout, Markov Chain Monte Carlo, or Variational Inference), Ensemble techniques, and test-time augmentation (Shanmugam et al., 2021). Bayesian theory is exploited in Sensoy et al. (2021) to design risk-calibrated evidential classifiers that allow to incorporate the misclassification error in the loss function while training deep image classifiers, thus accurately quantifying the uncertainty of the predictions. An innovative approach with respect to the mentioned well-established techniques was presented in Baek et al. (2023), where uncertainty is studied from a metrological perspective that considers neural networks as measurements tools and uses probabilistic robustness theory to provide safety guarantees to a robotic system. Moreover, Ghobrial et al. (2023) introduced a trustworthiness score quantifying the reliability of a deep neural network prediction by checking for the existence of given features in the predictions made by the model. Authors also design a suspiciousness score in the overall input images to help in the detection of those frames where false negatives existed. Yue et al. (2022) designed EviDCNN-3WC, a new methodology that combines deep convolutional neural networks (DCNNs) for feature learning and Dempster–Shafer (D-S) evidence theory as uncertainty measure to implement a three-way method for image classification. This technique was experimented in several medical classification scenarios, showing reduced classification risks. Speaking of trustworthiness, Mackowiak et al. (2021) represents a pioneering work on the explainability and robustness of generative classifiers (i.e., algorithms that classify by maximizing the probability of a sample given a class, as opposed to the standard procedure of discriminative classifiers). Although this classification approach is different from the one

used in this research (classical classifiers, like softmax classification, are discriminative), the work remains relevant in that it addresses the concept of uncertainty quantification as a metric of a model's explainability and not just as an assessment of its robustness. Finally, special attention should be paid to the use of conformal prediction in image classification, see Angelopoulos et al. (2020). The fact that conformal prediction has only recently been worked on for the evaluation of uncertainty quantification in images is due solely to the youth of the method (it has been developed starting in the late nineties and early two thousands by V. Vovk., the definitive reference is Vovk et al. (2005)) and not to its ability to make reliable predictions. Conformal predictors are indeed being used in several computer vision applications. As an example from smart agriculture, Farag et al. (2023) presents the adoption of inductive conformal predictors in a vision-based harvest-readiness classification of cauliflower plants, and compare this method with more traditional softmax outputs and uncertainties derived from Monte Carlo dropout. Another interesting work, Gendler et al. (2021), introduces a novel algorithm, Randomly Smoothed Conformal Prediction, that makes conformal prediction sets valid even in case of adversarial attacks that make the i.i.d. hypothesis untrue.

3. Background

3.1. Scalable Classifiers

Scalable Classifiers (SCs) were introduced in Carlevaro et al. (2023) as a family of (binary) classifiers parameterized by a scale factor $\rho \in \mathbb{R}$

$$\phi(\boldsymbol{x}, \rho) \doteq \begin{cases} 0 & \text{if } f(\boldsymbol{x}, \rho) < 0, \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

where the function $f : \mathcal{X} \times \mathbb{R} \longrightarrow \mathbb{R}$ is the so-called *classifier predictor*. Although this is a classical binary classifier, and thus the role of the labels is "interchangeable", we want to focus attention on class 0, being the one on which our method is able to provide confidence bounds in prediction. For example, to give the classifier a meaningful interpretation, we refer to the class 0 as a "safe" situation we want to target and the other class 1 as an "unsafe" situation[2].

Some examples might be differentiating between a patient's condition in developing or not developing a certain disease (Lenatti et al., 2022), or understanding what input parameters lead an autonomous car to a collision or non-collision (Carlevaro et al., 2022), and many others can be listed.

SCs rely on the main assumption that for every $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^N$ (where N is the dimension of the feature space) $f(\boldsymbol{x}, \rho)$ is continuous and monotonically increasing in ρ and that $\lim_{\rho \to -\infty} f(\boldsymbol{x}, \rho) < 0 < \lim_{\rho \to \infty} f(\boldsymbol{x}, \rho)$, (Carlevaro et al., 2023, Assumption 1). These assumptions imply that there exists a unique solution $\bar{\rho}(\boldsymbol{x})$ to the equation

$$f(\boldsymbol{x}, \rho) = 0 \tag{2}$$

2. The choice of labels is independent of the properties of scalable classifiers, and a remapping of the output does not affect the performance of the classification. In this case, we chose the labels 0 and 1 to establish a link with the deep learning framework, where usually the output of a (classification) network is interpreted as a probability level.

and the proof is available in (Carlevaro et al., 2023, Property 2). In words, a scalable classifier is a classifier that satisfies some crucial properties: *i*) given $\boldsymbol{x}$, there is always a value of ρ, denoted as $\bar{\rho}(\boldsymbol{x})$, that establishes the border between the two classes, *ii*) the increase of ρ forces the classifier to predict the 1 class and *iii*) a decrease of ρ maintains the target 0 class. Moreover, (Carlevaro et al., 2023, Property 3) shows how any standard binary classifier can be rendered scalable by simply including the scaling parameter ρ in an additive way with the classifier predictor, i.e. given the function $\hat{f} : \mathcal{X} \longrightarrow \mathbb{R}$ and its corresponding classifier $\hat{\phi}(\boldsymbol{x})$ then the function $f(\boldsymbol{x}, \rho) = \hat{f}(\boldsymbol{x}) + \rho$ provides the scalable classifier $\phi(\boldsymbol{x}, \rho)$. Thus, examples of classifiers that can be rendered scalable are support vector machines, support vector data descriptions, logistic regressions but also artificial neural networks.
Different values of the parameter ρ correspond to different classifiers that can be considered as the level sets of the classifier predictor with respect to ρ. In particular, since the class 0 encodes a safety condition, we introduce the scalable set

$$\mathcal{S}(\rho) = \{\, \boldsymbol{x} \in \mathcal{X} \;:\; f(\boldsymbol{x}, \rho) < 0 \,\}, \tag{3}$$

that is the set of points $\boldsymbol{x} \in \mathcal{X}$ predicted as 0 ("safe") by the classifier with the specific value ρ, i.e. the *safety region* of the classifier f for given ρ. The definition of scalable classifier allows to make the classifier more tractable and flexible without the necessity of a retraining. Specifically, special values of the parameter ρ can be defined in order to achieve target tasks of the classifier (like the minimization of the number of misclassified points). To achieve this, it is necessary to introduce methodologies capable of handling the statistical properties of the classified samples: among them, we introduce the *probabilistic scaling* (Mammarella et al., 2022) that, as explained in the next section, provides a readly way to tune ρ.

3.2. Probabilistic Scaling

Following the procedure of (Carlevaro et al., 2023, Theorem II.1) (that is reported in details in the next section), it is then possible to define the $\rho_\varepsilon-$safe set ($\mathcal{S}_\varepsilon$, equation (9) of the next section), in which the probability of unsafe instances ($y = 1$) belonging to the safety region is guaranteed to be less than ε. Before starting, however, we need to introduce the concept of "generalized maximum" that is at the basis of the whole procedure:

Definition 1 (Generalized Max) *Given a collection of n scalars $\Gamma = \{\gamma_i\}_{i=1}^n \in \mathbb{R}^n$, and an integer $r \in [n]$, we denote by*

$$\max^{(r)}(\Gamma)$$

the r-greatest value of Γ, so that there are no more than $r-1$ elements of Γ strictly larger than $\max^{(r)}(\Gamma)$.

This definition is interesting because of the "scaling factor" property in Alamo et al. (2018) that, in words, shows that, if the number of points is chosen large enough, the generalized max constitutes with very high probability a good approximation of the $\frac{r}{n}$-th quantile of the distribution of Γ. This makes it possible to limit a-priori the probability of observing values greater than the generalized maximum, making it possible, in this application on classification learning, to control the misclassification error of the prediction.

4. Deep Probabilistic Scaling

In the context of classification, the classifier's prediction function of a (convolutional) neural network is constructed as a weighted composition of activation functions applied to an affine transformation of the input $\boldsymbol{x}$, specifically

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &= \sigma(\boldsymbol{w}_L g_{L-1}(\boldsymbol{w}_{L-2} g_{L-2}(\dots g_1(\boldsymbol{w}_0 \boldsymbol{x} + b_0)\dots) + b_{L-1}) + b_L) - \frac{1}{2} \\ &= \sigma(h(\boldsymbol{x})) - \frac{1}{2}\end{aligned} \tag{4}$$

where L is the depth of the network, $\boldsymbol{w}_\ell$ and $g_\ell(\cdot)$ are respectively weights and activation function of the layer $\ell \in [L]$, $h(\cdot)$ is the function representing the composition of all hidden layers and $\sigma(\cdot)$ is the sigmoid function. After training, a new instance $\boldsymbol{x}_{\text{test}}$ is then associated to a class according to the rule

$$\hat{\varphi}(\boldsymbol{x}_{\text{test}}) = \begin{cases} 0 & \text{if } \hat{f}(\boldsymbol{x}_{\text{test}}) < 0, \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

In this formulation, (Carlevaro et al., 2023, Property 3) can be applied to obtain a scalable version of the above deep classifier by defining the new prediction function

$$f(\boldsymbol{x}, \rho) = \hat{f}(\boldsymbol{x}) + \rho, \quad \rho \in \mathbb{R}, \tag{6}$$

and by punctually following the procedure explained in (Carlevaro et al., 2023, Theorem II.1) it is therefore possible to deal rigorously with the uncertainty naturally carried by the model and correct it **without retraining the model itself**. In fact, we recall that the whole procedure outlined below is performed after training, as in the spirit of conformal prediction. However, probabilistic scaling allows to provide an operational methodology to control the model's prediction error rather than a quantification of the uncertainty. This methodology has been shown to work for any type of classifier (and recent insights suggest that it might work for regressors as well), but since this is the very first application to deep learning, for the sake of comprehensibility, we summarize the entire procedure here:

0. Take a pre-trained model and set the value of $\varepsilon \in (0,1), \delta \in (0,1)$.
1. Sample a calibration set $\mathcal{Z}_c \doteq \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_c}$ with size n_c such that $n_c \geq \frac{7.47}{\varepsilon} \ln \frac{1}{\delta}$[3] and consider $\mathcal{Z}_c^1 \doteq \{\ (\boldsymbol{x}, y) \in \mathcal{Z}_c \ : \ y = 1\ \}$.
2. Compute for all the samples belonging to $\mathcal{Z}_c^1$ the $\rho-$scores as follows

$$\bar{\rho}(\boldsymbol{x}) = \frac{1}{2} - \sigma(h(\boldsymbol{x})) \quad \forall \boldsymbol{x} \in \mathcal{Z}_c^1 \tag{7}$$

3. Compute $r = \left\lceil \frac{\varepsilon n_c}{2} \right\rceil$ and take the *probabilistic scaling of level* ε as

$$\rho_\varepsilon \doteq \max^{(r)} \left(\{\bar{\rho}(\tilde{\boldsymbol{x}}_j^1)\}_{j=1}^{n_1} \right) \tag{8}$$

3. This bound is well explained in (Carlevaro et al., 2023, Corollary I.1).

4. Define the ρ_ε-safe set

$$\mathcal{S}_\varepsilon \doteq \begin{cases} \mathcal{S}(\rho_\varepsilon) & \text{if } n_1 \geq r \\ \mathcal{X} & \text{otherwise} \end{cases}$$

Then, with probability no smaller than $1-\delta$,

$$\Pr\Big\{y = 1 \text{ and } \boldsymbol{x} \in \mathcal{S}_\varepsilon\Big\} \leq \varepsilon \tag{9}$$

In few more words, the above procedure defines an algorithm to statistically bound the number of *false negatives*[4] in the prediction of the network, with high probability. This is still an open topic in machine learning.

5. Link with Conformal Prediction

Although developed separately, our approach based on probabilistic scaling has much to share with conformal prediction. Both methods have their roots in quantile regression theory (Hao and Naiman, 2007) and are used in the post-processing phase of an algorithm (i.e., both techniques require having a pre-trained model on which to be applied) but there are aspects that are substantially different: of all of them, in the context of classification, CP provides only a qualitative measure of model uncertainty, while the combination of PS and SC provides a way to *correct* the performance of the algorithm itself, thereby improving the prediction result. Moreover, the approach can be useful when the statistical characteristics of the data have changed after training. Namely, the scalable classifier becomes more resilient while exploiting the re-calibration process on new data against those used for learning.
In this section, we highlight the similarities and differences between the two approaches, thus deriving a common link between the two, with a focus on image classification.

5.1. Score function for scalable deep neural networks

The idea of a score function is at the heart of conformal prediction theory. This special function encodes the agreement between a sample $\boldsymbol{x} \in \mathcal{X}$ and a proposed label $\tilde{y}$, namely $s(\boldsymbol{x}, \tilde{y})$, as well as a statistical information of the model. There is no unambiguous definition, although there are properties that it is preferable to satisfy (see Angelopoulos and Bates (2023)), such as the fact that the greater the value of the score function, the worse the agreement between the sample and the proposed label (in jargon, *negatively oriented*). For this reason, when dealing with image classification, one minus the softmax output of the class is usually adopted. Since we are focusing on binary classification, the softmax is replaced by the sigmoid, and the "standard" score function takes this form:

$$s_{\text{st}} = s(\boldsymbol{x}, \tilde{y}) = \tilde{y}(1 - \sigma(h(\boldsymbol{x}))) + (1 - \tilde{y})\sigma(h(\boldsymbol{x})) \tag{10}$$

considering that, with the notation introduced in (4), $\sigma(h(\boldsymbol{x}))$ denotes the probability of observing $y = 1$ given the sample $\boldsymbol{x}$. The definition of a scalable classifier, however, embodies a natural definition of a score function that generalizes to any type of classifier and thus to deep neural networks (Carlevaro et al., 2024):

4. In this formulation we refer to a false negative when an instance labelled as 1 is predicted as 0.

Definition 2 (Score Function for Scalable Classifiers)
Given a scalable classifier $\phi(\boldsymbol{x}, \rho)$ with classifier predictor $f(\boldsymbol{x}, \rho)$, given a point $\boldsymbol{x}$ and an associated candidate label $\hat{y}$, the score function associated to the scalable classifier is defined as

$$s(\boldsymbol{x}, \hat{y}) = \hat{y}\bar{\rho}(\boldsymbol{x}) + (1 - \hat{y})(-\bar{\rho}(\boldsymbol{x}))$$

with $\bar{\rho}(\boldsymbol{x})$ such that $f(\boldsymbol{x}, \bar{\rho}(\boldsymbol{x})) = 0$.

Note that this definition preserves the property of being negative oriented, i.e. smaller (negative) values mean better. Then, considering the predictor function as in (4), the score function associated is:

$$s_{\text{ps}} = s(\boldsymbol{x}, \hat{y}) = \hat{y}\left(\frac{1}{2} - \sigma(h(\boldsymbol{x}))\right) + (1 - \hat{y})\left(\sigma(h(\boldsymbol{x})) - \frac{1}{2}\right) \tag{11}$$

From now on, we will refer to the "standard" score function defined in (10) as s_{st} and to the "probabilistic scaling" one from (11) as s_{ps}. Without entering in the details of CP (for which we refer to Angelopoulos and Bates (2023)), the very starting point of the procedure is the computation of the (almost) $(1-\varepsilon)-$quantile, $\hat{q}_\varepsilon$, of the score values on the calibration set. The computation of the quantile is clearly not affected by linear transformations or composition with functions that preserve the order of the values. So, given a score function s, a strictly monotone function $\psi : \mathbb{R} \mapsto \mathbb{R}$ and α, β real values, the function $\psi(\alpha s + \beta)$ will generate the same conformal prediction sets of s. Considering then s_{st} and s_{ps} as defined above, the following property holds for the two quantities:

$$s_{\text{ps}} = s_{\text{st}} - \frac{1}{2} \tag{12}$$

This implies that the two score functions generate the *same* conformal sets. This result allows the adoption of probabilistic scaling also for CP. A more in-depth discussion on the topic follows in Section 6.4.

6. Experiments and Results

6.1. Data preparation

The performance of deep probabilistic scaling technique was assessed by considering several open-source datasets[5]. These included two benchmarks for image classification such as the MNIST database of handwritten digits (Deng, 2012) and CIFAR10 (Krizhevsky et al., 2009), and also other two ones, namely pneumoniaMNIST from medMNIST-v2 (Yang et al., 2023), and the WikiArt Art Movements/Styles dataset (Saleh and Elgammal, 2016)[6]. PneumoniaMNIST is the only originally binary classification dataset: it contains 5856 chest X-ray images labelled as *'normal'* or *'pneumonia'*. MNIST dataset collects 60000 training images from 10 classes (one per each digit 0-9), thus we defined three binary classification subtasks: class *'1'* versus class *'7'*, class *'3'* versus class *'8'*, and class *'2'* versus class *'5'*; to this end,

5. Code and data associated to the experiments is available at the following link: https://github.com/AlbiCarle/Deep-Probabilistic-Scaling
6. https://www.kaggle.com/datasets/sivarazadi/wikiart-art-movementsstyles

Dataset	Class 0	Class 1
MNIST1-7	'7'	'1'
MNIST3-8	'3'	'8'
MNIST2-5	'2'	'5'
CIFAR10	'truck'	'automobile'
PneumoniaMNIST	'normal'	'pneumonia'
WikiArt	'renaissance'	'baroque'

Table 1: Labels mapping in class 0 and class 1, for each dataset. The error of a trained image classification model, on class 0 images, will be bounded by ϵ thanks to the deep probabilistic scaling.

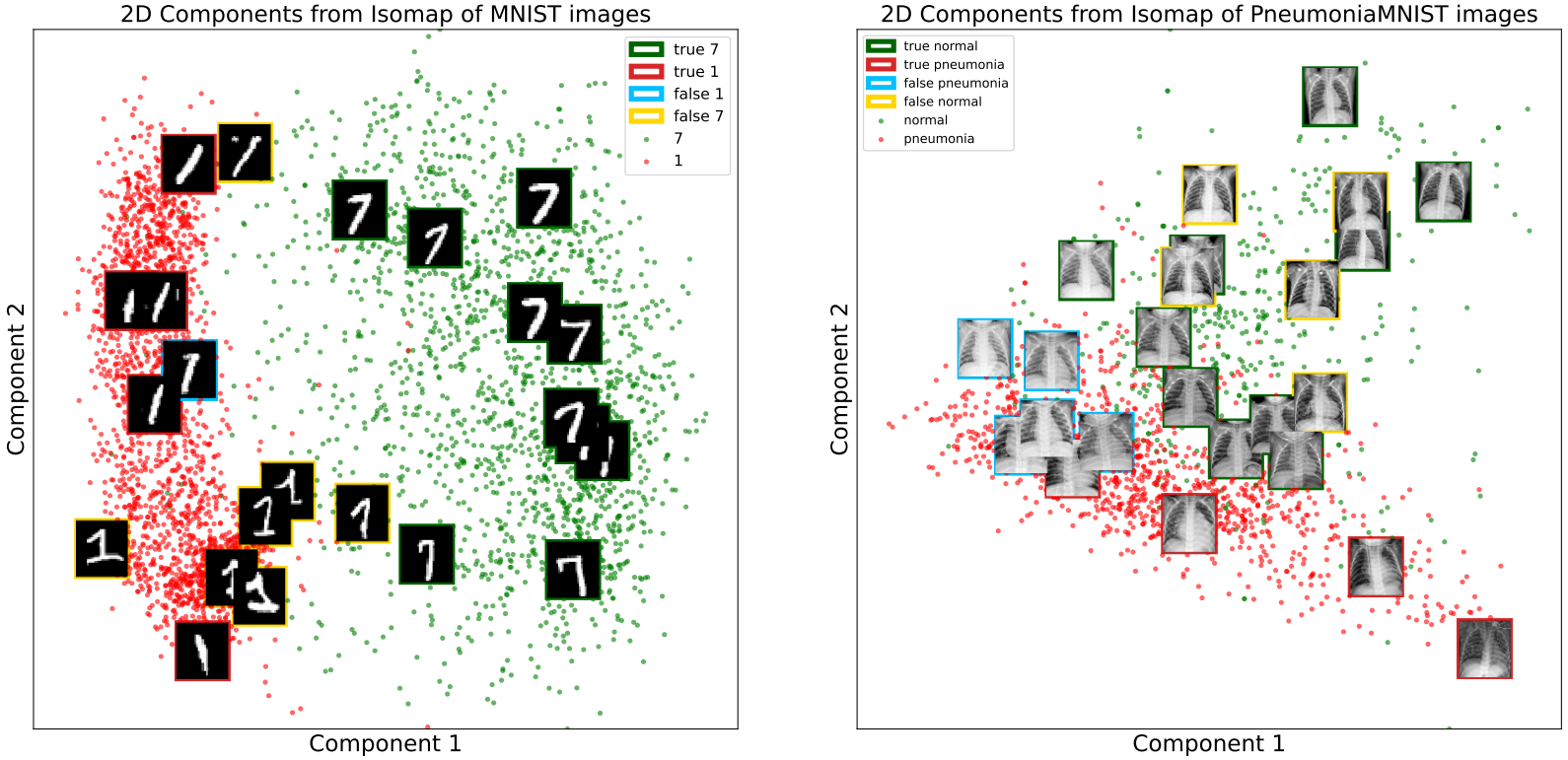
we only picked the images from the mentioned classes, for a total number of more than 10000 images in all cases. Similarly, we selected the *'automobile'* and *'truck'* images from the 10 classes of CIFAR10 dataset. WikiArt dataset contains images of artworks from a set of 13 different artistic movements, from which we chose *'renaissance'* and *'baroque'* as the classes to analyse. We remark that we selected the two classes to work on by thinking at scenarios where the classification by a model could have been more tricky (e.g., digits '1' and '7' from MNIST can look similar in some cases). Since our method asks to define a specific class on which to guarantee the model performance, and we assume this class to be labelled with 0 (with the other being denoted as 1), Table 1 shows which classes we assume as 0 and which as 1, for each dataset.

6.2. Model architecture

After properly defining the binary scenarios for our tests, the first step of the experiments consisted in training deep image classifiers that will constitute the basis (i.e., pre-trained models) of the deep PS algorithm.
In particular, we considered a Convolutional Neural Network (CNN) characterized by three convolutional layers with 32, 64 and 128 hidden neurons, respectively, a fully-connected layer with 512 neurons, and the output layer with the sigmoid activation function. For MNIST digits, we removed the last layer to prevent overfitting. The training was performed, in all cases, with the binary cross-entropy loss by using Adam optimizer (Kingma and Ba, 2014) with learning rate fixed to 0.001. Due to the simplicity of the test scenarios, the number of epochs was kept low, between 3 and 5. This network had a good performance, with false negative rate $\leq 5\%$ for all the datasets except CIFAR10 (38%). However, the advantage of our proposed technique lies in the possibility of tuning the classifier so to provide high guarantees that this error is bounded by choosing a pre-fixed ε level.

6.3. Impact of deep probabilistic scaling

Deep probabilistic scaling is a tool that has potentially great impact in developing useful and effective reliability in trustworthy AI for images. This is not just a way to check for false negatives, but rather a region where the reliability of the prediction is probabilistically guaranteed, leaving no room for uncertainty. This is clear in Figure 2(a) that shows the



(*a*) MNIST1-7. (*b*) PneumoniaMNIST.

Figure 2: Distribution of uncertainty of MNIST17 and PneumoniaMNIST. Shown in blue and yellow, respectively, are the uncertain classifications for class *'1'* and class *'7'* for MNIST1-7 and for class "*normal*" and class "*pneumonia*" for PneumoniaMNIST.

probabilistic correction of a network used to classify the digits '1' and '7' in the MNIST dataset. To visualize the classes as a 2D scatter plot, we applied the Isomap method (Tenenbaum et al., 2000), that is one of the earliest dimensionality reduction techniques, seeking a lower-dimensional embedding while maintaining the geodesic distances among all points. As shown in Table 1, the target class (0) is the digit '7'. The probabilistic safety region contains digits that look like well handwritten '7', impossible to confuse with '1'. Some of these are shown with a green box around them, over an Isomap embeddings scatter plot representing the real '7' class (green cluster in the figure). Outside the region, instead, we can find samples that evidently share more characteristics with the digit '1': in a straightforward way (in the same spirit of Figure 1 from the introduction) there is a sample labeled as '7' that is more like a '1' (sample with the blue box in the middle left part of the graph). This methodology has potentially a great impact in real safety-critical applications,

where being able to predict safe conditions with high probability guarantees is of paramount importance. Our test case on pneumoniaMNIST dataset is a first example where a safe class can be really defined. Deep probabilistic scaling allowed us to enclose inside the PSR all images that correspond to truly healthy patients, and by setting $\varepsilon = 0.01$ we minimized the probability that a pneumonia diagnosis is missed when the scalable classifier is adopted. The Isomap scatter plot of Figure 2(b) shows the real normal and pneumonia classes through their isomap embeddings. Colored boxes around sample images show the outcomes of the deep PS classifier. As for MNIST1-7, green boxes represent the images composing the safety region. Also, we can observe that false pneumonia images (i.e., normal class images that were left outside the safety region by the PS, depicted with light blue boxes) have isomap embeddings lying close to the pneumonia red cluster. In the same way, a few (at most 1%) pneumonia images are still wrongly labelled as normal (yellow boxes), and also their isomap components are located within the normal class. Current diagnostic America Thoracic Society (Metlay et al., 2019) guidelines for pneumonia diagnosis recognize chest radiography as the gold standard imaging modality, and the signs of this disease in the images are opacities in the thoracic area, including interstitial infiltrates or lobal consolidations. If we better look at the images of false pneumonia, we can note that they are characterized by wider white shades in the lungs, within the ribs, which may have been caused by artifacts. Conversely, the appearance of false normal images is closer to healthy radiographs, with the thorax area looking darker.

These are simple but significant examples illustrating that this method can be a powerful and *explainable* approach to uncertainty mitigation in image classification, with applications ranging from deep false detection to supporting tools for medical imaging, perception, surveillance, and any other application where reliable predictions are needed.

6.4. Results with conformal predictions

In Section 5 we defined a score function s_{ps} that is suitable for deep scalable CNN classifiers, which we applied to all the six considered datasets.

As it is common in the evaluation of conformal predictions, we assessed the score function with respect to four metrics, namely the average error and the average rates of empty, singleton (i.e., containing one label) and double-sized (i.e., containing two labels) prediction sets. We compared the results with those obtained using the classical definition of a score function, s_{st}, and, as correctly argued in Section 5, they are equal. So, for the sake of readability and avoiding redundancies, we will only report plots related to the score function s_{ps} defined from the deep probabilistic scaling.
The message we would like to communicate to the reader with the score function defined in (11), and analyzed here, is twofold: first, it allows a bridge to be drawn between two theories unrelated until now, making it possible to derive benefits for improving one and the other (e.g., by exploiting the well-established bounds of probabilistic scaling on the number of samples for the calibration set), and second, the definition of score function for scalable classifiers (11) is quite general and suitable for CP score function as well, whose choice is very peculiar to the field of application and user's experience.

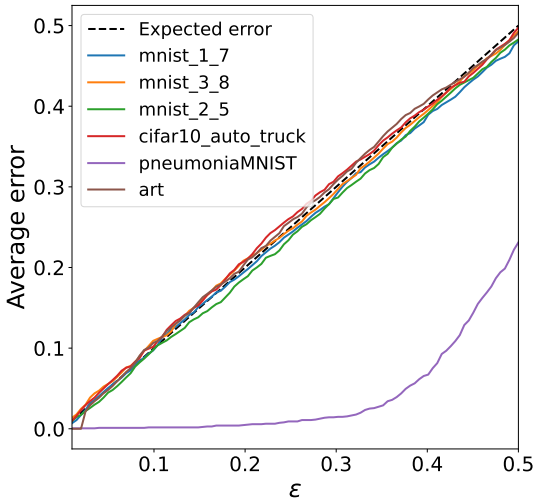
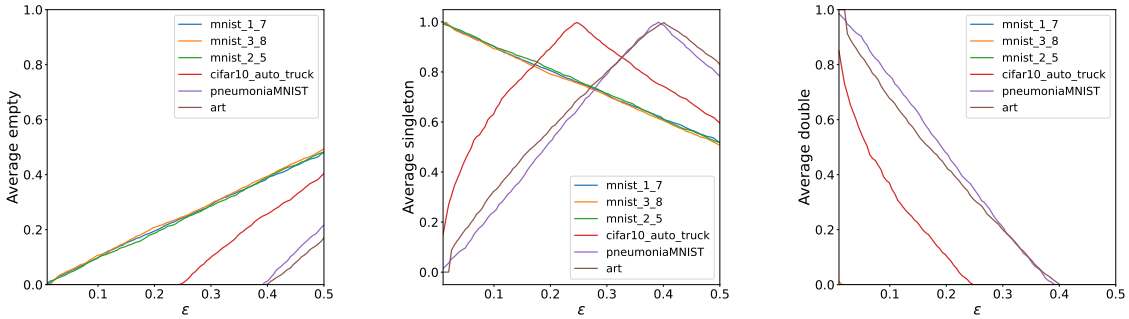


Figure 3: Average error obtained with s_{ps} score function, for $\varepsilon \in [0.01, 0.5]$.



(*a*) Empty. (*b*) Singleton. (*c*) Double.

Figure 4: Plots of the average rate of empty (left), singleton (middle), and double sized (right) prediction sets obtained by using the s_{ps} score function of Eq. 11, for $\varepsilon \in [0.01, 0.5]$

The average error is computed by interpreting an error as the case that the correct class label is not included in the prediction set. As expected, Figure 3 shows that the trend of the average error, for $\varepsilon \in [0.01, 0.5]$, tightly goes with ε for all datasets, with the only exception of pneumoniaMNIST: in this case, it is far lower than the expected error.

On the other hand, evaluating the size of the conformal sets relates to efficiency. The ideal behavior of a highly efficient conformal predictor should evidence large rates of singleton sets, with low rates of empty and double prediction sets. Figures 4(*a*), 4(*b*) and 4(*c*) show the obtained average rates of empty, singleton, and double sized prediction sets, respectively, for our application cases. We can observe that our score function produces a coherent behavior with respect to these metrics. Specifically, results are diversified between MNIST

and other datasets. Concerning all MNIST cases, the average rate of empty sets constantly increase following ε values, but no double sets are generated. As a result, the rate of singleton sets starts with high values and then decreases up to about 0.5 for the maximum ε level. For the other datasets, empty sets are absent up to some ε value (around 0.25 for pneumoniaMNIST, and 0.4 for CIFAR10 and Art), when their rate begins to grow. Correspondingly, we can observe that the rate of double sets is high for lower ε and undergoes a quick descent to zero in the same error level points where the empty rate starts to rise. Also, it is for these ε values that the average singleton curves achieve a maximum peak.

7. Conclusions and future works

The evaluation of uncertainty quantification in machine learning, and particularly in image analysis, is a topic that never ceases to demand new methodologies, new ideas or new interpretations of existing algorithms. Indeed, with this work on probabilistic scaling we have proposed a different approach in a highly studied field, highlighting the similarities and differences with conformal predictions, one of the most widely used and popular frameworks for uncertainty quantification nowadays. Nevertheless, the work focuses on the possibility of using probabilistic scaling in the deep learning environment, proving that it is possible to control misclassification error in the post-process, without the need to retrain the network. This is a statement contained in the seminal work on probabilistic scaling, the aforementioned Carlevaro et al. (2023), which effectively took shape in this research. In particular, starting from the concept of scalable classifiers, the paper shows the definition of scalable neural networks and consequently the concept of probabilistic safety region for deep algorithms. These special regions are able to enclose samples that represent the target class well with high probability: the smaller the regions, the lower the uncertainty within them. This concept is closely related to the idea of creating a safe environment for classification problems, which can have a significant impact on reliable artificial intelligence for engineering. Moreover, this research contributes also to improve the knowledge on conformal prediction theory, establishing non-trivial relationships between the two approaches (e.g. Definition 2) and allowing to exploit well established results to enhance both the methods. An example is the operational bound on the number of calibration samples (Carlevaro et al., 2023, Corollary I.1) that can be exploited to better understand critical conformal prediction properties like the well-known "$(\varepsilon, \delta)-$validity". Future work will follow, e.g., extension to multi-class and multi-label classification or adaptation of the probabilistic safety region to regression tasks.

Acknowledgments

This work was supported in part by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement ID: 101070028. The work was also supported by Future Artificial Intelligence Research (FAIR) project, Italian National Recovery and Resilience Plan (PNRR), Spoke 3 - Resilient AI. The work of F. Dabbene was supported by the Italian Ministry of Research, under the complementary actions to the PNRR "Fit4MedRob - Fit for Medical Robotics" Grant (PNC0000007). Moreover T. Alamo acknowledges support

from grant PID2022-142946NA-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by ERDF, A way of making Europe.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Teodoro Alamo, Jose M. Manzano, and E. Camacho. *Robust Design Through Probabilistic Maximization: In Honor of Roberto Tempo*, pages 247–274. 01 2018. ISBN 978-3-030-04629-3. doi: 10.1007/978-3-030-04630-9_7.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237.

Anastasios N. Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2020.

Woo-Jeong Baek, Christoph Ledermann, Tamim Asfour, and Torsten Kröger. Combining measurement uncertainties with the probabilistic robustness for safety evaluation of robot systems. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 473–480. IEEE, 2023.

Alberto Carlevaro, Marta Lenatti, Alessia Paglialonga, and Maurizio Mongelli. Counterfactual building and evaluation via explainable support vector data description. *IEEE Access*, 10:60849–60861, 2022.

Alberto Carlevaro, Teodoro Alamo, Fabrizio Dabbene, and Maurizio Mongelli. Probabilistic safety regions via finite families of scalable classifiers. *arXiv preprint arXiv:2309.04627*, 2023.

Alberto Carlevaro, Teodoro Alamo Cantarero, Fabrizio Dabbene, and Maurizio Mongelli. Conformal predictions for probabilistically robust scalable machine learning classification. *arXiv preprint arXiv:2403.10368*, 2024.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Mohamed Farag, Jana Kierdorf, and Ribana Roscher. Inductive conformal prediction for harvest-readiness classification of cauliflower plants: A comparative study of uncertainty quantification methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 651–659, 2023.

Jakob Gawlikowski, Cedrique R. N. Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A

survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Abanoub Ghobrial, Darryl Hond, Hamid Asgari, and Kerstin Eder. *A Trustworthiness Score to Evaluate DNN Predictions*. 2023.

Lingxin Hao and Daniel Q Naiman. *Quantile regression*. Number 149. SAGE, 2007.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, ON, Canada, 2009.

Marta Lenatti, Alberto Carlevaro, Aziz Guergachi, Karim Keshavjee, Maurizio Mongelli, and Alessia Paglialonga. A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations. *PLOS ONE*, 17(11):1–24, 11 2022.

Radek Mackowiak, Lynton Ardizzone, Ullrich Kothe, and Carsten Rother. Generative classifiers as a basis for trustworthy image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2971–2981, 2021.

Martina Mammarella, Victor Mirasierra, Matthias Lorenzen, Teodoro Alamo, and Fabrizio Dabbene. Chance-constrained sets approximation: A probabilistic scaling approach. *Automatica*, 137:110108, 2022.

Joshua P Metlay, Grant W Waterer, Ann C Long, Antonio Anzueto, Jan Brozek, Kristina Crothers, Laura A Cooley, Nathan C Dean, Michael J Fine, Scott A Flanders, et al. Diagnosis and treatment of adults with community-acquired pneumonia. an official clinical practice guideline of the american thoracic society and infectious diseases society of america. *American journal of respiratory and critical care medicine*, 200(7):e45–e67, 2019.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, (2), 2016.

Murat Sensoy, Maryam Saleki, Simon Julier, Reyhan Aydogan, and John Reid. Misclassification risk and uncertainty quantification in deep classifiers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2484–2492, 2021.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. *MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification*, volume 10. Nature Publishing Group UK London, 2023.

Xiaodong Yue, Yufei Chen, Bin Yuan, and Ying Lv. Three-way image classification with evidential deep convolutional neural networks. *Cognitive Computation*, 14(6):2074–2086, 2022.