The Penalized Inverse Probability Measure for Conformal Classification

Paul Melki
IMS, CNRS, University of Bordeaux
EXXACT Robotics
paul.melki@u-bordeaux.fr

Lionel Bombrun
IMS, CNRS, University of Bordeaux
Bordeaux Sciences Agro
lionel.bombrun@ims-bordeaux.fr

Boubacar Diallo, Jérôme Dias
EXXACT Robotics
boubacar.diallo@exxact-robotics.com
jerome.dias@exxact-robotics.com

Jean-Pierre Da Costa
IMS, CNRS, University of Bordeaux
Bordeaux Sciences Agro
jean-pierre.dacosta@ims-bordeaux.fr

Abstract

The deployment of safe and trustworthy machine learning systems, and particularly complex black box neural networks, in real-world applications requires reliable and certified guarantees on their performance. The conformal prediction framework offers such formal guarantees by transforming any point into a set predictor with valid, finite-set, guarantees on the coverage of the true at a chosen level of confidence. Central to this methodology is the notion of the nonconformity score function that assigns to each example a measure of "strangeness" in comparison with the previously seen observations. While the coverage guarantees are maintained regardless of the nonconformity measure, the point predictor and the dataset, previous research has shown that the performance of a conformal model, as measured by its efficiency (the average size of the predicted sets) and its informativeness (the proportion of prediction sets that are singletons), is influenced by the choice of the nonconformity score function. The current work introduces the Penalized Inverse Probability (PIP) *nonconformity score, and its regularized version* RePIP*, that allow the joint optimization of both efficiency and informativeness. Through toy examples and empirical results on the task of crop and weed image classification in agricultural robotics, the current work shows how PIP-based conformal classifiers exhibit precisely the desired behavior in comparison with other nonconformity measures and strike a good balance between informativeness and efficiency.*

1. Introduction

The development and deployment of machine learning-based autonomous systems has been a flourishing field of research in both academia and, relatively more recently, in the industry [1, 2]. While machine learning models often exhibit high performance "in the lab", they often face much more difficulty when deployed in the real world, for a number of reasons that are not yet fully clear [3]. Indeed, when faced with a new observation, the model will produce a new prediction whose quality is often related to the similarity of this new observation to what the model has previously seen. When the new observation is quite anomalous with respect to the previously seen data or even slightly perturbed, most models will produce wrong predictions [4], with often dire and intolerable consequences in safety critical applications such as autonomous driving [5–7] and medical diagnosis [8–10], to name a few.

The safe deployment of machine learning systems in the real world is therefore incumbent upon the integration of at least two main important features into them [2, 11]: (1) the ability to provide valid and trustworthy guarantees on the quality of predictions in "normal" conditions, and (2) the ability to reliably detect and signal anomalies when faced with them.

Conformal prediction is a method that provides formal statistical guarantees on the predictive quality of any black box model [12, 13]. It has recently gained in popularity due to the minimal assumptions required for its deployment. Without imposing explicit conditions on the data distribution, any base point predictor can be transformed using the conformal approach into a set predictor with formal guarantees on the coverage of the true value at confidence level $1-\alpha$, where α is a chosen level of tolerance to error. Formally, in a supervised learning context, whereby for each object $\mathbf{x} \in \mathcal{X}$ is assigned a label $y \in \mathcal{Y}$, a conformal model produces prediction sets $\mathcal{C}_{1-\alpha} \subset \mathcal{Y}$ that satisfy the *marginal*

coverage guarantee [13, 14]

$$\mathbb{P}\big(y \in \mathcal{C}_{1-\alpha}(\mathbf{x})\big) \geq 1 - \alpha \tag{1}$$

whenever the test data follow the same distribution as the data on which the model was calibrated. Under this condition, the coverage guarantee is satisfied marginally over all possible calibration sets. Additionally, the study of the structure and the size of the predicted sets allows us to quantify the uncertainty of the base model, and to detect examples on which the model is highly uncertain [15]. As such, the conformal approach can be used to satisfy the two conditions for safe deployment of machine learning systems as it has been shown in a number of applications [16] ranging from railway signaling [17], medical imaging [18], to nuclear fusion [19].

Three main components are needed to conduct inductive conformal prediction [20]: a base predictor $\mathcal{B}$ (which can be any machine learning point predictor), a dataset on which to calibrate the model so that it becomes a conformal predictor, and a nonconformity score function Δ that assigns a "strangeness value" to each example in the calibration set. This value measures how *conforming* each individual is to what the model has previously seen. While the marginal coverage guarantee is satisfied by construction, the quality of the predicted sets is influenced by these three components. For example, a neural network $\mathcal{B}$ with low accuracy can still be calibrated to achieve $1 - \alpha = 0.9$ coverage, but will tend to predict much larger sets, since it is uncertain about the true class and thus needs to predict many to guarantee the inclusion of the true one.

The object of interest in this work is the nonconformity score function Δ. In particular, we are interested in studying the influence of different nonconformity functions on two of the most commonly used metrics for the evaluation of conformal classifiers [21]: *efficiency*, the average size of the predicted sets, and *informativeness*, the proportion of predicted singleton sets. These two metrics measure, in some sense, the "usability" of the conformal approach when needed to take decisions under normal condition, and may be useful to signal high uncertainty conditions. The context of the study is automated precision weeding in agriculture [22], whereby a robotic system is embedded on a tractor to detect and spray herbicides on undesirable weeds in real-time, under real-world conditions. The precision agriculture sector is an interesting test-bed for safe AI methodologies since they are indeed needed in agriculture, but do not directly threaten human lives in case of failure.

Related work A good body of research is dedicated to the development of useful and efficient nonconformity score functions [23–25]. For classification, the first comprehensive work is that of Johansson *et al.* [21] in which the authors study the impact of different model-agnostic nonconformity functions – in particular, the Hinge Loss and the Margin Score – on neural network classifiers. The authors find that neither of these score functions allows the joint maximization of informativeness and efficiency. Their empirical results show that the Hinge Loss minimizes the size of prediction sets, while the Margin Score maximizes the number of singletons. These results are further confirmed by Aleksandrova and Chertov [26, 27] on most of the datasets they tested, in their work aiming at reconciling the two scores by computing, for a new observation, two conformal sets using both the Hinge and Margin scores, then choosing the Margin-based set as the final prediction if it is a singleton, or the Hinge set otherwise. Unfortunately, this approach may be quite inefficient as it requires repeating the calibration step for each nonconformity function. Fisch *et al.* [28] propose an efficient conformal classification approach based on an expansion of the notion of validity to include the concept of *admissible* labels, which are semantically plausible class labels for a given example. Such an expansion may lead to highly inefficient prediction sets in learning tasks with a large number of classes. As such, the authors develop an efficient cascaded inference algorithm that reduces the size of the prediction set by progressively filtering the number of candidates via a sequence of increasingly complex classifiers. Other works have explored ways to combine multiple conformal models in such a way as to preserve the validity guarantee while producing sets that are as efficient as possible [25, 29, 30].

Contributions In direct continuation of these previous works, and for the expansion of the still meager body of work on conformal prediction in precision agriculture [31–33], our work proposes the following contributions:

1. The proposal of a new model-agnostic nonconformity function that strikes a good balance between optimizing both efficiency and informativeness: the *Penalized Inverse Probability* (PIP);
2. The proposal of a simple regularized version of PIP, RePIP, inspired by [34] for improved efficiency in use cases with a large number of classes;
3. The comparison of PIP with other nonconformity measures from the literature on toy examples, showing the balanced and adaptive behavior of this measure under different settings;
4. The comparison of PIP and RePIP with other nonconformity measures from the literature based on efficiency and informativeness through rigorous empirical experiments on an image dataset for crop and weed classification taken under real-world conditions with the aim of providing valid guarantees on the performance of a precision weeding system.

2. Definitions & Mathematical Setup

Let $\mathbf{x} \in \mathcal{X}$ be a vector of features, which we will call an *object* [12]. To each object is associated a class label $y \in \mathcal{Y} := \{1, ..., K\}$ to form what we call an *example* $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. A black-box classifier $\mathcal{B}$ is trained on a set of n_{train} examples to output for an object a class prediction $\hat{\mathcal{B}}(\mathbf{x}) = \hat{y} \in \{1, ..., K\}$ and an associated estimated probability $\hat{p}^{\hat{y}} \in [0, 1]$, such that $\sum_{k=1}^{K} \hat{p}^k = 1$.

The inductive conformal approach consists of a calibration step in which the trained classifier is calibrated on a set of n_{cal} calibration examples $\{\mathbf{z}_i = (\mathbf{x}_i, y_i), i = 1, ..., n_{\text{cal}}\}$ using a real-valued nonconformity score function $\Delta(\mathbf{z}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The output of the calibration step is usually a quantile value $q_{\text{cal}} \in \mathbb{R}$ computed on the distribution of nonconformity scores over the calibration set.

This quantile is then used to produce prediction sets $\mathcal{C}_{1-\alpha}(\mathbf{x}) \subset \mathcal{Y}$ on the remaining n_{test} test examples. For each class, its score Δ is computed based on the probability estimated by $\mathcal{B}$, then compared to q_{cal} in a hypothesis test of whether the class is considered "conforming" enough or not. The produced prediction sets are *valid* in the sense that they satisfy the marginal coverage guarantee defined in Equation (1). This property is verified empirically by computing the **empirical marginal coverage**, which is simply the proportion of prediction sets that cover the true label:

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\{y_i \in \mathcal{C}_{1-\alpha}(\mathbf{x}_i)\}} \tag{2}$$

The quality of the prediction sets can then be evaluated using these two metrics:

- **Efficiency**, defined as the average size of the predicted sets:

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}_{1-\alpha}(\mathbf{x}_i)| \tag{3}$$

  where $|\ .\ |$ is the set cardinality, the number of classes in the predicted set.
- **Informativeness**, defined as the percentage of predicted sets of size 1 (often called *oneC* in the literature [21, 26]):

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\{|\mathcal{C}_{1-\alpha}(\mathbf{x}_i)|=1\}} \tag{4}$$

Clearly, conformal predictors that have both high efficiency and high informativeness are the preferred models in practice, at a fixed coverage level of $1 - \alpha$. Smaller set sizes are easier to manipulate and be used to construct decision rules. Singleton predictions are the most informative predictions since they do not manifest any "uncertainty" about the predicted class. A most informative, and efficient, conformal model would be one that predicts only singletons while guaranteeing marginal coverage. Unfortunately, such an optimal conformal model is impossible to attain in practice [14].

3. Nonconformity Score Functions

3.1. Review of Some Nonconformity Scores

The nonconformity measure quantifies the "strangeness" of a given object by comparing it to the objects previously encountered by the model during training and calibration [35]. For the same base predictor $\mathcal{B}$, different nonconformity functions lead to different conformal predictors. Here, we review commonly used nonconformity score functions for classification from the literature [14, 21]. Since the estimated probabilities $\hat{p}^k$ are fixed for a given object $\mathbf{x}$, the nonconformity score function $\Delta(\mathbf{z})$ will simply be denoted $\Delta(y)$ in the following for ease of understanding. Note also that during the calibration step of the conformal procedure, y is the true class of object $\mathbf{x}$, while during the prediction phase, y is the tested class to be included or not in the prediction set.

Hinge Loss (IP) [21] Also known as *Inverse Probability*, this score function measures how far the estimated probability of y (where y is the true class label) is from the perfect score of 1:

$$\Delta^{\text{IP}}(y) = 1 - \hat{p}^y \tag{5}$$

Indeed, a perfect classifier should always assign a probability of 1 to the true class label, which would have a Hinge score of 0. For smaller probability estimates of y, a higher Hinge score is assigned since the model is deemed more uncertain about y. The Hinge Loss can thus be considered a very "natural" measure of nonconformity. Unfortunately, it suffers from a major shortcoming: it does not take the probability estimates of the other classes into consideration.

Margin Score (MS) [21] Assuming an implicit hypothesis that good predictive models should assign the highest probability estimate to the true class, the MS measures the difference between the estimated probability of y and the highest estimated probability among the other classes:

$$\Delta^{\text{MS}}(y) = \max_{k \neq y} \hat{p}^k - \hat{p}^y = \Delta^{\text{IP}}(y) + \underbrace{\max_{k \neq y} \hat{p}^k - 1}_{\text{penalization}} \tag{6}$$

A large positive value of this score indicates that the estimated probability assigned to y is distant from the class of highest confidence. It means that class y is considered highly strange in comparison to the class the model considers as the true one. Notice that y is *always* penalized, even when it is the most probable class, which is not ideal. Another shortcoming of the MS is that it only takes the maximum probability into consideration, why not take the probabilities of the other classes directly into consideration? It is important to note that in cases of anomalies, OOD observations or adversarial attacks, neural networks would tend to assign the highest confidence to classes that are completely

wrong [4], thus putting the reliability of the Margin Score into question.

Regularized Adaptive Prediction Sets (RAPS) This nonconformity function was first introduced in [14] as part of the APS approach, with the aim of producing prediction sets whose size adapts to, and reflects, the difficulty of each object. It is the first score function that fully integrates a range of estimated probabilities other than that of y. In particular, the APS score incorporates all the estimated probabilities that are larger than that of the class of interest. Observing that the APS score tends to predict relatively large set sizes in learning problems with a large number of classes, Angelopoulos *et al.* [34] introduced a regularized version of this score, named RAPS.

Let the operator $R(k)$ be the rank of class k after the estimated probabilities $p^1, ..., p^K$ have been sorted in decreasing order, and $\hat{p}^{[r]}$ be the probability estimate of the class having rank r, such that $\hat{p}^k = \hat{p}^{[R(k)]}$, we can define the RAPS score function as:

$$\Delta^{\text{RAPS}}(y) = \underbrace{\sum_{r=1}^{R(y)-1} \hat{p}^{[r]} + u \cdot \hat{p}^{[R(y)]}}_{\text{APS}} + \underbrace{\lambda \cdot \left(R(y) - k_{reg}\right)^{+}}_{\text{regularization}} \quad (7)$$

where u is a uniform random variable in $(0, 1)$ for tie-breaking, λ is the penalization amount and k_{reg} is the rank at which to start penalizing. λ and k_{reg} can be fixed by the user or optimized on a held-out dataset. The penalization is proportional to the how further away is y in the ranking of estimated probabilities from k_{reg}. When y has a very low probability, meaning that it has a very high $R(y)$, its score will be very strongly penalized, leading to the exclusion of y from the prediction set. This will lead, on average, to smaller set sizes, as it excludes from the prediction sets those classes that would have been included by the original APS score (obtained for $\lambda = 0$). While the APS and RAPS have been developed with adaptivity and efficiency in mind, their authors' do not seem to take the informativeness criterion into consideration.

3.2. Penalized Inverse Probability

In this article, we introduce the *Penalized Inverse Probability* (PIP), a new nonconformity score function that integrates components from the three previously presented measures with the aim of optimizing both efficiency and informativeness. Following the same notation presented previously, PIP can be defined as:

$$\Delta^{\text{PIP}}(y) = \underbrace{1 - \hat{p}^y}_{\Delta^{\text{IP}}(y)} + \underbrace{\sum_{r=1}^{R(y)-1} \frac{\hat{p}^{[r]}}{r} \mathbb{1}_{\{R(y)>1\}}}_{\text{penalization}} \quad (8)$$

For $R(y) = 1$, when y is the class with the highest estimated probability, the PIP score is simply the Hinge (IP) Loss. In all other cases, the sum of the estimated probabilities of all the classes with higher probability than y weighted by the inverse of their rank is added as a penalization term. As such, a decreasing weight is associated to each class that is closer to y. This penalization term resembles the APS score, and alleviates the shortcoming inherent by IP of not taking the estimated probabilities of other classes into consideration. Furthermore, for $R(y) = 2$, it should be clear that $\Delta^{\text{PIP}}(y) = 1 + \Delta^{\text{MS}}(y)$. As such, the PIP score exhibits analogous behavior to different nonconformity functions depending on the estimated probabilities by the base model $\mathcal{B}$, leading to better adaptivity, as will be shown in the toy examples below. For more detailed developments on the relationship between the PIP and the other scores, we refer the interested reader to Section 1 in the Supplementary Material.

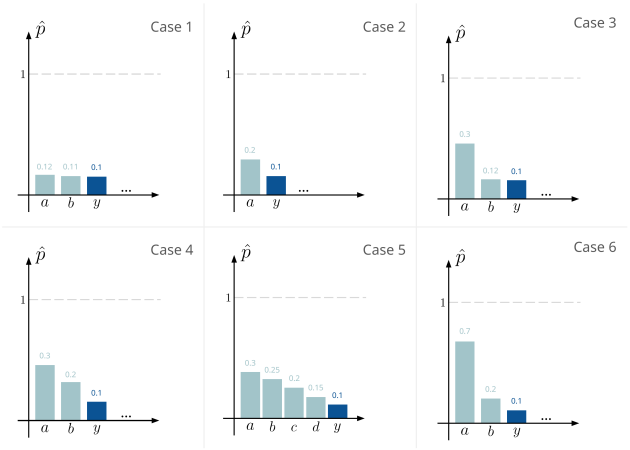


Figure 1. Six different potential configurations of model outputs sorted in decreasing order of $\hat{p}$. Only the classes until reaching the class of interest y are shown. Computed nonconformity scores for each case can be seen in Table 1.

	$\Delta^{\text{IP}}(y)$	$\Delta^{\text{MS}}(y)$	$\Delta^{\text{PIP}}(y)$
Case 1	0.90	0.02	**1.08**
Case 2	0.90	0.10	**1.10**
Case 3	0.90	0.20	**1.26**
Case 4	0.90	0.20	**1.30**
Case 5	0.90	0.20	**1.43**
Case 6	0.90	0.60	**1.70**

Table 1. Computed scores of the example cases shown in Figure 1. The proposed $\Delta^{\text{PIP}}(y)$ manifests a more adaptive behavior for the varying configurations than the classical IP and MS functions.

Toy examples Consider the six different possible output configurations of a neural network classifier shown in Figure 1. The class of interest is y and its estimated probability is fixed to $\hat{p}^y = 0.1$ in all the examples. Only the classes having higher estimated probabilities than y are

shown since they are the only ones that are used in the computations of the different scores. In Table 1 are shown the different scores assigned to class y in each of the cases, sorted in increasing order. A greater score is a sign of greater "nonconformity" – that is, of higher uncertainty – attributed to y.

The first obvious observation is that IP assigns the same score to y in all cases. As $\hat{p}^y = 0.1$ is the same in all cases and IP is, by definition, indifferent to the other classes, all the configurations are reduced to the same score. This rigidity is often undesirable in a nonconformity score function.

The MS measure, on the other hand, manifests a more fluid behavior since it also considers the highest estimated probability. Case 1 is assigned the lowest MS score, since the estimated probabilities of y and a are quite similar. As such, MS considers that class y is as likely a candidate as a to be the first predicted class, and thus assigns it a low nonconformity score. Case 2 is considered a bit "stranger" than Case 1 by the MS function because the difference between the maximal class a and y is a bit larger, which is a desirable behavior by this score function. In Case 6, although y has the same rank $R(y) = 3$ as in Case 1, the MS value is maximal since the *margin* between the $\hat{p}^a$ and $\hat{p}^y$ is large. Cases 3 to 5 show the shortcoming of the MS measure. Since in all these cases the difference between $\hat{p}^a$ and $\hat{p}^y$ is the same, they will all be assigned the same score value, even though it is clear that class y in Case 5 should be assigned a higher nonconformity value than in Case 3 or even in Case 4.

The proposed PIP function exhibits the most versatile behavior since it takes into consideration all the classes having higher estimated probabilities than y. $\Delta^{\text{PIP}}(y)$ is different in all the distinct configurations, manifesting the specificity of each case. Indeed, it can easily be shown that Δ^{PIP} guarantees a different score for every class even in the case of highest uncertainty where all the classes have the same estimated probability $1/K$ (Section 1 in Supplementary Material). Case 1 has the lowest PIP score, since class y is almost as likely as a or b to be predicted as the first class. As such y is not deemed strange in such a condition. The behavior of PIP in such situations is similar to that of MS. Case 2 is considered slightly stranger because the difference between $\hat{p}^a$ and $\hat{p}^y$ is larger and cannot simply be attributed to some "noise." While y has the same estimated probability and rank $R(y) = 3$ in both Case 3 and Case 4, it receives a slightly lower score in Case 3 since the difference with the b is very small: y could very much have been the second class and thus need not be penalized heavily for falling in third place. Class y in Case 5 is further penalized because more significant classes have higher estimated probabilities than y.

Summary of PIP score properties The desirable behavior manifested by PIP can be summarized as:

- In all situations, the Hinge Loss (IP) is a baseline value for the PIP function. Therefore, classes with low probability estimates will tend to be assigned higher nonconformity scores. This kind of behavior leads to a lower average size of predicted sets (higher efficiency) since it tends to exclude the classes with low probability estimates [21].
- PIP takes into consideration all the probability estimates of the other classes with higher probabilities when computing the score for a given class. This includes the maximum probability class. Therefore, when $\hat{p}^y$ has a low value compared to $\max_{k \neq y} \hat{p}^k$, class y will be heavily penalized (just like with the MS measure). This behavior generally leads to more predicted singletons (higher informativeness) because in all cases where one class has a very high probability estimate, all the other classes will be heavily penalized and thus excluded from the predicted set [21].
- Additionally, PIP distinguishes the cases where the difference between $\hat{p}^y$ and the "more probable" classes is significant or not, penalizing less when such differences are negligible and can be attributed to some noise. This leads to scores that are different almost everywhere, permitting better discrimination between the different model outputs.

3.3. Regularized PIP

For learning tasks with a large number of classes, the user may require to preserve the desirable behavior of the PIP score function but with smaller set sizes on average. The same regularization term added to APS [34] can be added to obtain RePIP, a regularized version of the proposed nonconformity measure:

$$\Delta^{\text{RePIP}}(y) = \Delta^{\text{PIP}}(y) + \underbrace{\gamma \cdot \big(R(y) - k_{reg}\big)^+}_{\text{regularization}} \qquad (9)$$

Here, γ is the equivalent of the λ parameter in the RAPS nonconformity score and k_{reg} is, similarly to RAPS, the rank at which to begin penalizing more.

4. Experimental Results

In this section, we study the performance of different conformal classifiers obtained using the nonconformity score functions presented previously on the task of classifying images taken under real world conditions into 13 different plant species. This learning task is part of a precision weeding robotic use case, where an autonomous robot should distinguish weeds from cultivated crops and spray them with herbicide in real-time. Guaranteeing the performance of the weed classifiers is of great importance since missed weeds can multiply quickly and threaten heavily the health of the cultivated crops and the quality of harvest.

Figure 2. Some randomly chosen example images of 6 different classes. *Common buckwheat* and *rye brome* are weeds, while *corn*, *pea* and *sunflower* are cultivated species.

4.1. Experimental Setup

The public WE3DS dataset recently published in [36] is originally a dataset of RGB-D images with semantic segmentation masks densely annotated into 17 plant species classes in addition to the *soil* class for the background. Due to the scarcity of publicly available crop and weed classification datasets, this dataset has been transformed into a classification dataset. Discarding the depth channel, the original RGB images have been divided into non-overlapping windows of size 224×224. To each resulting image is associated a true class label which is defined as the class with the highest number of pixels in the corresponding semantically annotated mask. This results into a dataset of around 14,800 RGB images with 13 different classes, of which six random specimens are shown in Figure 2. We refer the interested reader to Section 2 in the Supplementary Material for a full description of the data preparation procedure.

The database is then randomly divided into: (1) a training set (70%), on which a ResNet18 classifier [37] is trained using default hyperparameters and pretrained weights on ImageNet [38], and fixed for all experiments; the remaining 30% of the data are then split into (2) a calibration set (13.5%) for conformal calibration and (3) a test set (16.5%) on which the conformal classifiers are evaluated. It is important to note that the choice of the base model $\mathcal{B}$ is not of great importance and is not the focal point of this study. It is for this reason, and especially to be able to study the differences among the nonconformity score functions, that we opted for a classical ResNet18 classifier which does not manifest exceptional classification performance on this task. It could have very well been replaced by a newer state-of-the-art deep classifier.

After training the ResNet18 classifier, the neural network is calibrated using each of the previously presented nonconformity score functions at the chosen confidence level of $1 - \alpha = 0.9$. Then, it is used to predict sets of classes for the test images. To make sure that the obtained results are not simply due to having favorable samples of images, the calibration and test steps are repeated 1000 times, each time on a different random split of the data. The random seed of the i^{th} random split, $i = 1, 2, ..., 1000$, is the same across the different nonconformity score functions so as to obtain results that are truly comparable and not simply influenced by the aleatoric uncertainty inherent to the data.

4.2. Setting γ and λ for RePIP and RAPS

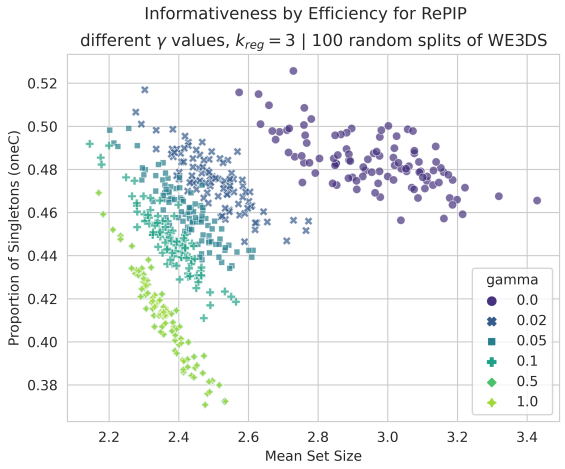
For RAPS and RePIP, k_{reg} is fixed at 3 based on this specific use case's requirements. In general, we prefer not to have prediction sets with more than 3 classes: the cultivated species, a weed species and the soil. In order to choose the regularization amounts λ and γ, we conduct a parameter sweep by testing multiple values from a manually defined grid. For each value and each method, a different conformal classifier is obtained for which we compute the efficiency and informativeness. Similarly to the experimental setup, with the aim of verifying the reliability of the estimated metrics, each conformal classifier is calibrated and tested on multiple random splits of the data.

Figure 3 shows the average set size and the proportion of singletons for each random split of the data and different values of γ (Figure 3a) and λ (Figure 3b). Depending on the user's preferences and the use case requirements, the optimal value can be chosen so as to place more weight on minimizing inefficiency or maximizing informativeness.

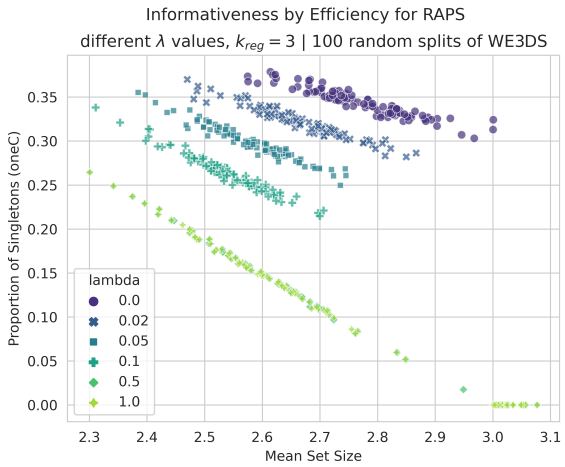
In our case, we deem it more important to maximize the number of predicted singletons while maintaining the coverage guarantee, as it is much easier to construct decision rules when only one class is predicted. Therefore, based on the empirical results in Figure 3, we choose $\gamma = \lambda = 0.02$ as values for the hyperparameters. We also note that for both hyperparameters, a limit seems to be reached at 0.5 whereby any greater value produces the same prediction sets (notice that the data points for the values 0.5 and 1 are overlapping).

4.3. Results and Discussion

The comparison of the different models is conducted based on the efficiency and informativeness criteria. A desirable model is one that optimizes both of these criteria by producing prediction sets with small size on average and as many singletons as possible without violating the marginal coverage guarantee. Figure 4 shows the results obtained



(a)

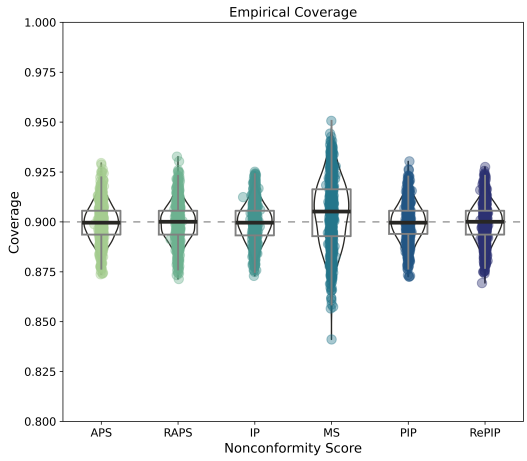


(b)

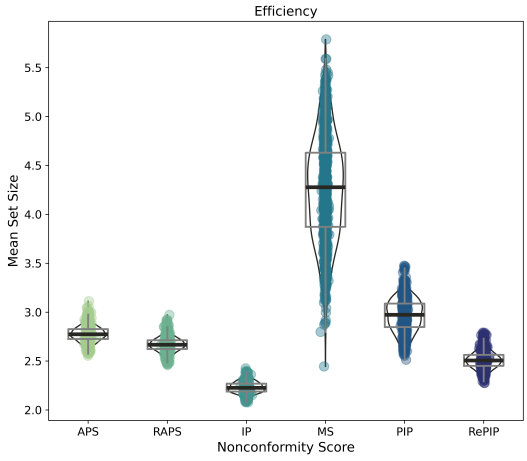
Figure 3. *Efficiency* and *Informativeness* for different values of the regularization hyperparameters. For each value of γ and λ, 100 different splits of the calibration and test sets are considered for more reliable results.

over the 1000 runs for each conformal classifier. Unsurprisingly, all the conformal classifiers are able to maintain the required 90% marginal coverage guarantee on average, with MS showing a comparatively unstable behavior with respect to the other measures (Figure 4a).

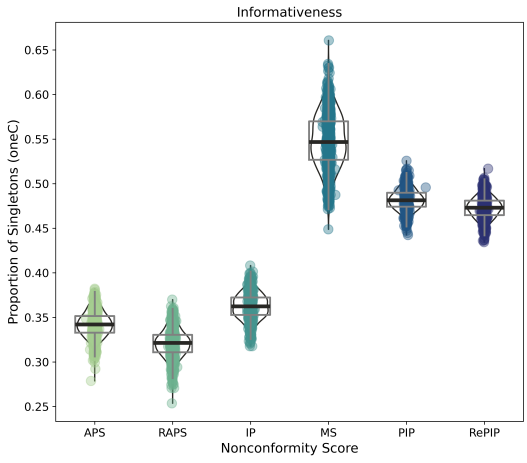
As can be seen in Figure 4b, the Hinge (IP) score leads to the smallest average set size, which is in accordance with the empirical results in [21] showing that IP is the measure to use to maximize efficiency. RAPS and RePIP which are designed with efficiency in mind through the regularization term lead to slightly larger set sizes on average, with RePIP coming in second place after IP. A slight difference between APS and RAPS can be noticed. The Margin (MS) score function shows a significantly unstable behavior over the different random runs. This can be due to its deep de-



(a)



(b)



(c)

Figure 4. Violin plots of experimental results on 1000 random splits of the WE3DS classification dataset (each point is a random split): (a) *Empirical Coverage* – (b) *Efficiency* (Mean Set Size) – (c) *Informativeness* (Proportion of Predicted Singletons).

pendence on the data it faces via the outputs of the base classifiers, an inference that can be made by comparing the divergent results in [21] and [27]. It also manifests a considerably higher average set size on average than all the other methods, a result in agreement with Johansson *et al.* [21]. The proposed PIP score, while exhibiting a slightly larger average set size than the other methods, is still much more efficient than the MS. This slight inefficiency is a price to pay for a considerable increase in informativeness (see Equation 4).

Indeed, MS manifests the highest proportion of predicted singletons, in accordance with the literature [21, 26], with more than 50% of predicted sets being singletons, on average. This result is influenced by the estimated probabilities of the base neural network: when the base classifier assigns a much higher estimated probability to one class in comparison to the others – that is, it is highly "confident" in the class it predicts – all the other classes will be considerably penalized, and thus excluded from the prediction set. This behavior is in agreement with Case 6 in Figure 1 and Table 1. This behavior tends to increase the number of predicted singletons only when the base classifier $\mathcal{B}$ already has a relatively high accuracy. The other nonconformity score functions, APS, RAPS and IP, that are not explicitly concerned with informativeness, have significantly less predicted singletons. On the other hand, our proposed PIP and RePIP scores can be considered quite competitive with MS in terms of informativeness with around 50% of predicted sets having size 1, and manifest better stability with regards to the data in comparison with MS. Interestingly, while the regularization via RePIP leads to considerably smaller set sizes on average, it does not decrease informativeness in any noticeable way, thus striking the required balance between the two evaluation criteria.

In a robotic pipeline, a conformal model that satisfies the condition of guaranteed coverage under normal conditions with such a high level of singletons along with a moderate average set size (such as with PIP or RePIP) is quite attractive. While providing around half of the predictions as singletons that can readily be used to take decisions, the conformal classifier produces the remaining predictions as sets that consist of only 2 or 3 classes, on average, on which adapted decision rules can be constructed easily for autonomous agents [39].

5. Conclusion

Conformal prediction is an important methodology for developing safe, deployable, machine learning systems. As long as the data faced by the model resembles, to a certain extent, the data on which it has been calibrated, the conformal model maintains the marginal coverage guarantee. Even though this marginal warranty can be strengthened, for example to provide class-conditional [13, 14, 40] or group-conditional coverage guaranties [32, 41], it already constitutes a strong gauge of validity for machine learning models, in particular black box neural networks that do not provide such guarantees by default. The conformal envelope around any learning model can be an important step for its certification as a valid model for deployment.

However, while any well-calibrated conformal model can provide coverage guarantees, the utility of the predictive model as a component in a larger decisional pipeline, in fully autonomous systems or human decision support systems, depends heavily on the prediction sets produced [39, 42]. In the current work, we introduced the *Penalized Inverse Probability* (PIP), and its regularized version (RePIP), with the aim of jointly optimizing the efficiency and informativeness of conformal classifiers. PIP and RePIP, mixing elements from other nonconformity score functions, provide a well-balanced hybrid behavior. The empirical results on crop and weed classification using deep neural networks show that PIP-based classifiers lead to relatively efficient prediction sets with significantly higher level of informativeness than their counterparts. Future work will continue this line of research notably by studying the behavior the different nonconformity measures on multiple datasets consisting of varying number of classes. A promising direction of exploration in safe AI is the comparison of the performance and robustness of these different nonconformity score functions under "abnormal" conditions, for example under distribution shifts and with regards to anomalous observations.

References

[1] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in Deploying Machine Learning: A Survey of Case Studies," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–29, Jul. 2023. 1

[2] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability," *Computer Science Review*, vol. 37, p. 100270, Aug. 2020. 1

[3] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, and D. Sculley, "Underspecification Presents Challenges for Credibility in Modern Machine Learning," *Journal of Machine Learning Research*, vol. 23, no. 1, pp. 226:10 237–226:10 297, Jan. 2022. 1

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," no. arXiv:1312.6199, Feb. 2014. 1, 4

[5] B. Spanfelner, D. Richter, S. Ebel, U. Wilhelm, W. Branz, and C. Patz, "Challenges in Applying the ISO 26262 for Driver Assistance Systems," *Tagung Fahrerassistenz, München*, vol. 15, no. 16, p. 2012, 2012. 1

[6] Q. Rao and J. Frtunikj, "Deep Learning for Self-Driving Cars: Chances and Challenges," in *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, ser. SEFAIS '18. New York, NY, USA: Association for Computing Machinery, May 2018, pp. 35–38.

[7] M. R. Alam and C. M. Ward, "Adversarial Examples in Self-Driving: A Review of Available Datasets and Attacks," in *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. DC, USA: IEEE, Oct. 2022, pp. 1–6. 1

[8] E. Petersen, Y. Potdevin, E. Mohammadi, S. Zidowitz, S. Breyer, D. Nowotka, S. Henn, L. Pechmann, M. Leucker, P. Rostalski, and C. Herzog, "Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions," *IEEE Access*, vol. 10, pp. 58 375–58 418, 2022. 1

[9] W. N. Price, II, S. Gerke, and I. G. Cohen, "Potential Liability for Physicians Using Artificial Intelligence," *JAMA*, vol. 322, no. 18, pp. 1765–1766, Nov. 2019.

[10] G. Maliha, S. Gerke, I. G. Cohen, and R. B. Parikh, "Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation," *The Milbank Quarterly*, vol. 99, no. 3, pp. 629–647, Sep. 2021. 1

[11] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved Problems in ML Safety," no. arXiv:2109.13916, Jun. 2022. 1

[12] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY: Springer, 2005. 1, 3

[13] A. N. Angelopoulos and S. Bates, "Conformal Prediction: A Gentle Introduction," *Foundations and Trends® in Machine Learning*, vol. 16, no. 4, pp. 494–591, Mar. 2023. 1, 2, 8

[14] Y. Romano, M. Sesia, and E. Candes, "Classification with Valid and Adaptive Coverage," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 3581–3591. 2, 3, 4, 8

[15] R. Laxhammar, "Chapter 4 - Anomaly Detection," in *Conformal Prediction for Reliable Machine Learning*, V. N. Balasubramanian, S.-S. Ho, and V. Vovk, Eds. Boston: Morgan Kaufmann, Jan. 2014, pp. 71–97. 2

[16] V. Balasubramanian, S.-S. Ho, and V. Vovk, *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Newnes, Apr. 2014. 2

[17] L. Andeol, T. Fel, F. de Grancey, and L. Mossina, "Confident Object Detection via Conformal Prediction and Conformal Risk Control: An Application to Railway Signaling," in *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*. PMLR, Aug. 2023, pp. 36–55. 2

[18] C. Lu, A. Lemay, K. Chang, K. Höbel, and J. Kalpathy-Cramer, "Fair Conformal Predictors for Applications in Medical Imaging," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 12 008–12 016, Jun. 2022. 2

[19] J. Vega, A. Murari, A. Pereira, S. González, and I. Pastor, "Accurate and Reliable Image Classification by Using Conformal Predictors in the TJ-II Thomson Scattering," *Review of Scientific Instruments*, vol. 81, no. 10, p. 10E118, Oct. 2010. 2

[20] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive Confidence Machines for Regression," in *Machine Learning: ECML 2002*, ser. Lecture Notes in Computer Science, T. Elomaa, H. Mannila, and H. Toivonen, Eds. Berlin, Heidelberg: Springer, 2002, pp. 345–356. 2

[21] U. Johansson, H. Linusson, T. Löfström, and H. Boström, "Model-Agnostic Nonconformity Functions for Conformal Classification," in *International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 2072–2079. 2, 3, 5, 7, 8

[22] G. R. Y. Coleman, A. Bender, K. Hu, S. M. Sharpe, A. W. Schumann, Z. Wang, M. V. Bagavathiannan, N. S. Boyd, and M. J. Walsh, "Weed Detection to Weed Recognition: Reviewing 50 Years of Research to Identify Constraints and Opportunities for Large-Scale Cropping Systems," *Weed Technology*, vol. 36, no. 6, pp. 741–757, Dec. 2022. 2

[23] H. Boström, H. Linusson, T. Löfström, and U. Johansson, "Evaluation of a Variance-Based Nonconformity Measure for Regression Forests," in *Conformal and Probabilistic Prediction with Applications*, ser. Lecture Notes in Computer Science, A. Gammerman, Z. Luo, J. Vega, and V. Vovk, Eds. Cham: Springer International Publishing, 2016, pp. 75–89. 2

[24] H. Linusson, U. Johansson, H. Boström, and T. Löfström, "Efficiency Comparison of Unstable Transductive and Inductive Conformal Classifiers," in *Artificial Intelligence Applications and Innovations*, ser. IFIP Advances in Information and Communication Technology, L. Iliadis, I. Maglogiannis, H. Papadopoulos, S. Sioutas, and C. Makris, Eds. Berlin, Heidelberg: Springer, 2014, pp. 261–270.

[25] N. Gauraha and O. Spjuth, "Synergy Conformal Prediction," in *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*. PMLR, Sep. 2021, pp. 91–110. 2

[26] M. Aleksandrova and O. Chertov, "How Nonconformity Functions and Difficulty of Datasets Impact the Efficiency of Conformal Classifiers," in *ICML 2021 Workshop on Distribution-Free Uncertainty Quantification*, Aug. 2021. 2, 3, 8

[27] ——, "Impact of Model-Agnostic Nonconformity Functions on Efficiency of Conformal Classifiers: An Extensive Study," in *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction with Applications*. PMLR, Sep. 2021, pp. 151–170. [Online]. Available: https://proceedings.mlr.press/v152/aleksandrova21a.html 2, 8

[28] A. Fisch, T. Schuster, T. S. Jaakkola, and R. Barzilay, "Efficient Conformal Prediction via Cascaded Inference with Expanded Admission," in *International Conference on Learning Representations (ICLR)*, 2020. 2

[29] P. Toccaceli and A. Gammerman, "Combination of Conformal Predictors for Classification," in *Proceedings of the Sixth*

Workshop on Conformal and Probabilistic Prediction and Applications. PMLR, May 2017, pp. 39–61. 2

[30] ——, "Combination of inductive mondrian conformal predictors," *Machine Learning*, vol. 108, no. 3, pp. 489–510, Mar. 2019. 2

[31] M. Farag, J. Kierdorf, and R. 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 (ICCV) Workshops*, 2023, pp. 651–659. 2

[32] P. Melki, L. Bombrun, B. Diallo, J. Dias, and J.-P. Da Costa, "Group-Conditional Conformal Prediction via Quantile Regression Calibration for Crop and Weed Classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023, pp. 614–623. 8

[33] S. Chiranjeevi, M. Sadaati, Z. K. Deng, J. Koushik, T. Z. Jubery, D. Mueller, M. E. O. Neal, N. Merchant, A. Singh, A. K. Singh, S. Sarkar, A. Singh, and B. Ganapathysubramanian, "Deep Learning Powered Real-Time Identification of Insects Using Citizen Science Data," no. arXiv:2306.02507, Jun. 2023. 2

[34] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty Sets for Image Classifiers Using Conformal Prediction," *International Conference on Learning Representations (ICLR)*, vol. 2021, 2021. 2, 4, 5

[35] G. Shafer and V. Vovk, "A Tutorial on Conformal Prediction," *Journal of Machine Learning Research*, vol. 9, no. 12, pp. 371–421, 2008. 3

[36] F. Kitzler, N. Barta, R. W. Neugschwandtner, A. Gronauer, and V. Motsch, "WE3DS: An RGB-D Image Dataset for Semantic Segmentation in Agriculture," *Sensors*, vol. 23, no. 5, p. 2713, Mar. 2023. 6

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 6

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2009, pp. 248–255. 6

[39] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, and A. Majumdar, "Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners," in *Proceedings of the Conference on Robot Learning (CoRL)*. Atlanta, GA, USA: PMLR, Aug. 2023. 8

[40] V. Vovk, "Conditional Validity of Inductive Conformal Predictors," *Machine Learning*, vol. 92, no. 2-3, pp. 349–376, Sep. 2013. 8

[41] I. Gibbs, J. J. Cherian, and E. J. Candès, "Conformal Prediction With Conditional Guarantees," May 2023. 8

[42] E. Straitouri, L. Wang, N. Okati, and M. G. Rodriguez, "Improving Expert Predictions with Conformal Prediction," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 32 633–32 653. 8