On the Efficiency and Informativeness of Deep Conformal Classifiers using the Penalised Inverse Probability Nonconformity Function

Paul Melki [1,2], Lionel Bombrun [1,3], Boubacar Diallo [2],
Jérôme Dias [2] & Jean-Pierre Da Costa [1,3]

[1] *IMS, CNRS UMR 5218, Bordeaux INP, Université de Bordeaux, France*
[2] *EXXACT Robotics, France*
[3] *Bordeaux Sciences Agro, France*
paul.melki@u-bordeaux.fr

Résumé. Les prédictions conformes permettent de transformer n'importe quel predicteur ponctuel en un *prédicteur d'ensembles* avec des garanties formelles sur le recouvrement de la vraie classe à un niveau de confiance choisi. Un composant important de la chaîne de prédiction conforme est le *score de non-conformité* qui attribue à chaque observation une mesure "d'étrangeté" par rapport aux données vues précédemment. Plusieurs modèles conformes sont souvent comparés en fonction de leur *efficacité*, généralement mesurée par la taille moyenne des ensembles prédits, et leur *informativité*, le nombre de singletons prédits. Comme cela a été montré dans la littérature, ces deux critères sont influencés par les données, les performances du modèle de base et le score de non-conformité. Leur maximisation conjointe à l'aide d'une fonction de non-conformité ayant de bonnes propriétés est souhaitable. Le travail actuel introduit la fonction de non-conformité "Penalised Inverse Probability" (PIP) inspirée des fonctions de score classiques (*Hinge Loss* et *Margin Score*). À l'aide d'exemples illustratifs et de résultats empiriques en classification d'images de cultures en agriculture de précision, nous montrons que le PIP présente précisément le comportement souhaité, établissant un bon équilibre entre informativité et efficacité.

Mots-clés. Prédictions conformes, score de non-conformité, classification, incertitude.

Abstract. The conformal prediction framework transforms any point predictor into a *set predictor* with formal guarantees on the coverage of the true value at a chosen level of confidence. An important component of the conformal pipeline is the *nonconformity score function* which assigns to each observation a measure of "strangeness" in comparison to the previously seen data points. Multiple conformal models are often compared based on their *efficiency*, usually measured by the average size of the predicted sets, and their *informativeness*, the number of predicted singletons. As shown in the literature, these two criteria are influenced by the dataset, the performance of the base model and the nonconformity score function. The joint maximisation of these criteria using a well-behaved nonconformity function is desirable. The current work presents the "Penalised Inverse Probability" (PIP) nonconformity function inspired from classical score functions (Hinge Loss and Margin Score). Using some illustrative examples and experimental results on the task of crop and weed image classification in precision agriculture, we show that PIP exhibits precisely the desired behaviour, striking a good balance between informativeness and efficiency.

Keywords. Conformal prediction, nonconformity score, classification, uncertainty.

1 Introduction

Let $\mathbf{x} \in \mathcal{X}$ be a vector of features, which we will call an *object*, following the commonly used annotation in the conformal prediction literature (Vovk et al., 2005). For 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 individual 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 $\sum_{k=1}^{K} \hat{p}^k = 1$.

In inductive conformal prediction (Papadopoulos et al., 2002), the trained classifier is then calibrated on a held-out set of n_{cal} calibration examples $\{(\mathbf{x}_i, y_i), i = 1, ..., n_{\text{cal}}\}$ using a well-defined nonconformity score function $\Delta(\hat{\mathcal{B}}(\mathbf{x})) = \Delta(\hat{y}) : \mathcal{Y} \to \mathbb{R}$. The nonconformity function assigns a "strangeness value" to each individual in the calibration set that measures how *conforming* it is to what the model has previously seen. The result of the calibration procedure is usually a quantile value q_{cal} computed on the distribution of nonconformity scores over the calibration set. This quantile is used at the prediction phase to construct prediction sets $\mathcal{C}_{1-\alpha}(\mathbf{x}) \subset \mathcal{Y}$ that guarantee the inclusion of the true class y at least $1 - \alpha$ of the times, for $\alpha \in (0, 1)$ a chosen significance level. This is known as the *marginal coverage* guarantee (Vovk et al., 2005):

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

The strength of the conformal approach is that it does not assume any distribution of the data and is agnostic to the base model $\mathcal{B}$. Indeed, the coverage guarantee in Equation 1 requires only the weak assumption of exchangeability of the data (Aldous, 1985) to be valid over the distribution of all possible calibration sets (Stutz et al., 2022). However, while the coverage is guaranteed at $1 - \alpha$ regardless of $\mathcal{B}$, the nonconformity score function used and the dataset considered, the efficiency and informativeness of the conformal pipeline depend on these components.

Let us first define these two notions. Vovk et al. (2016) propose different criteria for measuring the efficiency of conformal predictors. In the current work, we consider the two most commonly used measures, studying the average size of predicted sets and the proportion of predicted singletons. For a test set $\{(\mathbf{x}_i, y_i), i = 1, ..., n_{\text{test}}\}$ of examples different than those taken for training the base classifier and conformal calibration, we can define:

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

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

  which we would generally like to minimise without violating the coverage guarantee. Hence, for two conformal predictors guaranteeing coverage at the same $1-\alpha$ confidence level, the predictor producing smaller prediction sets is preferred, in general, and is considered more efficient.

- **Informativeness**, as the percentage of predicted sets of size 1 (often called *oneC* in the literature):

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

  where $\mathbb{1}_{\{\,\}}$ is the indicator function. Singletons are considered the most informative predictions and are the most useful in practice to construct decision rules. A most informative model would be one that predicts singletons only while guaranteeing marginal coverage.

The aim of this current work is to propose a novel nonconformity measure that optimises these two, often conflicting, criteria. Indeed, the experimental work by Johansson et al. (2017) studying the impact of different model-agnostic nonconformity functions – in particular, the Hinge Loss (Eq. 4) and the Margin Score (Eq. 5) – on single and ensembles of neural network classifiers has shown that neither of these score functions permits the joint maximisation of informativeness and efficiency. Their empirical results show that the Hinge Loss minimises the size of prediction sets, while the Margin Score maximises the number of singletons. In an attempt to reconcile these two nonconformity functions and optimise the two criteria, Aleksandrova and Chertov (2021) propose a conformal prediction algorithm that computes a prediction set using both nonconformity scores. A decision is then taken: if the Margin Score-based prediction set produces a singleton, it is taken as the final prediction. Otherwise, the prediction set obtained using the Hinge Loss is predicted, as it usually has the smaller size. The empirical evidence presented by the authors shows that their approach does not always provide better results than using any of the two classical nonconformity functions. Moreover, the algorithm requires repeating the calibration and the prediction steps twice using each of the score functions, which can be quite inefficient.

The current work introduces the "Penalised Inverse Probability" (PIP) nonconformity function computed using the estimated probabilities $\hat{p}^k$, $k \in \mathcal{Y}$. It can be interpreted, as will be shown, as both a regularised version of the Hinge Loss function and of the Margin Score. The aim of using this function is striking the balance between maximisation of informativeness and minimisation of inefficiency. Simple illustrative examples are presented to show the behaviour of this function under different possible configurations of output probabilities. Experimental results comparing the proposed measure with other functions from the literature, on the task of crop and weed image classification using deep neural networks for precision agriculture, show that PIP manifests precisely the type of balanced behaviour it is intended for.

2 Nonconformity Score Functions

2.1 Review of Some Nonconformity Scores

The nonconformity score function quantifies the "strangeness" of each new observation by implicitly comparing it to the "old" observations, previously seen during the training and

calibration of the predictive model (Shafer and Vovk, 2008). For the same base model $\mathcal{B}$, different nonconformity functions produce different conformal predictors. In this work, the following nonconformity score functions from the literature (Johansson et al., 2017; Romano et al., 2020; Angelopoulos et al., 2021) are compared. Let y be a class of interest and $\hat{p}^y$ its estimated probability:

- **Hinge Loss (IP)** This score also known as *Inverse Probability* measures how far the estimated probability of y is from the perfect score of 1:

$$\Delta^{\text{IP}}(y) = 1 - \hat{p}^y \tag{4}$$

  The score function measures the model's certainty in the class of interest. A minimum value is assigned to a class with maximum probability, while less probable classes are assigned higher values. The Hinge Loss is, in a sense, a very natural measure of nonconformity. However, it is important to notice that it does not take the scores assigned to other classes into consideration.

- **Margin Score (MS)** This score measures the difference between the estimated probability of y and the highest estimated probability among the other classes:

$$\Delta^{\text{MS}}(y) = \max_{k \neq y} \hat{p}^k - \hat{p}^y \tag{5}$$

  In a sense, this score function measures the model's lack of confidence in class y. An implicit hypothesis assumed when using this score function is that good predictive models tend to assign the highest probability estimate to the true class. However, this is not always the case in many practical situations.

- **Regularised Adaptive Prediction Sets (RAPS)** This score function is first introduced by Romano et al. (2020) as part of their APS approach with the aim of constructing valid prediction sets that adapt to the difficulty of each observation. It was further extended by Angelopoulos et al. (2021) with the addition of a regularisation term to reduce the size of the predicted sets. This score function takes into consideration all the classes that have higher estimated probability than the class of interest. It is defined as the cumulative probability of y plus a regularisation term.

  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:

$$\Delta^{\text{RAPS}}(y) = \underbrace{\sum_{r=1}^{R(y)-1} \hat{p}^{[r]} + u \cdot \hat{p}^{[R(y)]}}_{\text{APS}} + \underbrace{\lambda \cdot \Big(R(y) - k_{reg}\Big)^+}_{\text{regularisation}} \tag{6}$$

  Here, u is uniform random variable on $(0, 1)$ for tie-breaking, while λ (the penalisation amount) and k_{reg} (the rank at which to start penalising) are regularisation parameters that can be fixed *a priori*, or optimised on a held-out dataset. Notice that for $\lambda = 0$ we have the non-regularised APS score. It is important to note that while RAPS aims for adaptivity and improved efficiency, it is not clear that it was intended to take the informativeness criterium into consideration.

2.2 Penalised Inverse Probability

This work introduces a new nonconformity function that combines ingredients of the three previously presented measures: *Penalised Inverse Probability* (PIP). Following the same notation presented previously, it is defined as:

$$\Delta^{\text{PIP}}(y) = \begin{cases} 1 - \hat{p}^y & \text{if } R(y) = 1 \\ \underbrace{1 - \hat{p}^y}_{\Delta^{\text{IP}}(y)} + \underbrace{\sum_{r=1}^{R(y)-1} \frac{\hat{p}^{[r]}}{r}}_{\text{regularisation}} & \text{otherwise} \end{cases} \tag{7}$$

The first part of the function is simply the well-known Hinge Loss function defined previously, which is blind, by default, to the estimated probabilities of the other classes. A regularisation term that consists of the cumulative probability of all the classes with a higher estimated probability than the class of interest, weighted by the inverse rank of the class, is added. This term alleviates one of the Hinge Loss' shortcomings, namely the fact that it does not take into consideration the estimated probabilities of other classes, without only considering the maximal class as in Margin Score. The penalisation term does assign the highest penalty to the maximal class, but also considers the classes in-between.

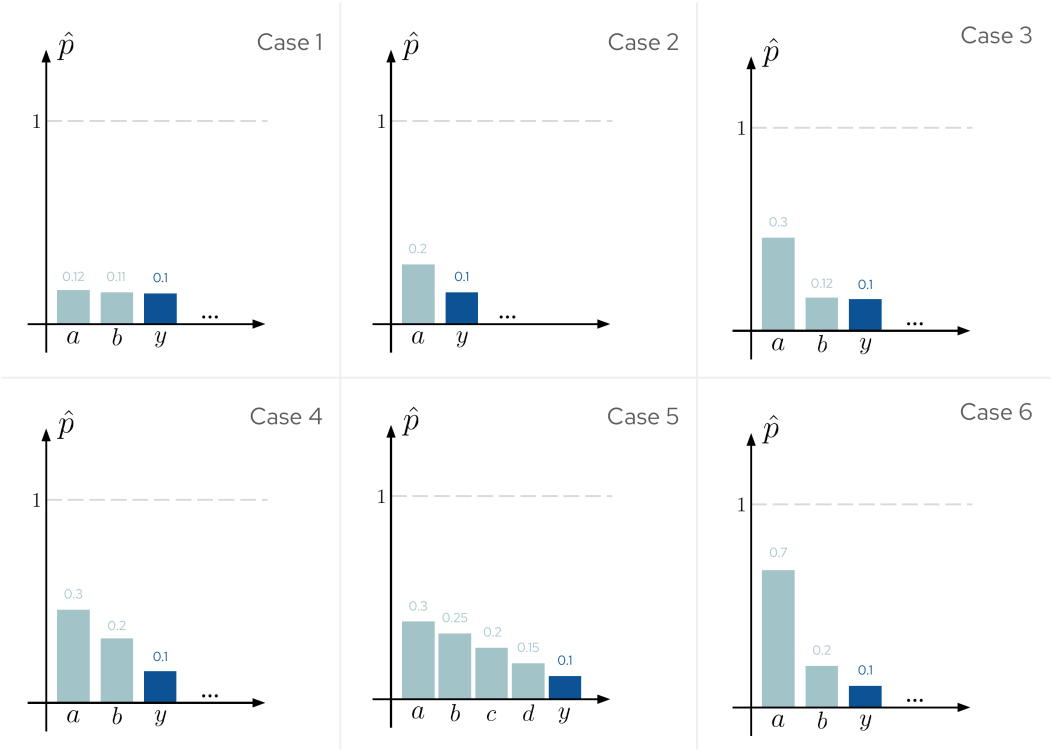


Figure 1: Six different potential configurations of model outputs sorted in decreasing order of $\hat{p}$. Only the classes until reaching the class of interest y are shown. Computed nonconformity scores for each case can be seen in Table 1.

Consider the six different possible configurations shown in Figure 1. The class of interest being y, only the classes having estimated probabilities bigger than $\hat{p}^y = 0.1$ are shown, since they are the only ones taken into consideration for the computation of the different scores. Table 1 shows the different scores assigned by the IP, MS and PIP functions to the class y, sorted in increasing order (a higher score indicates a more uncertain class). The aim of this visualisation is to show how the proposed PIP function exhibits a more adaptive and balanced behaviour in different situations, unlike the classical, more rigid, IP and MS approaches that would assign the same score for quite distinct configurations.

Notice that the IP function, that takes into consideration only the estimated probability of the class of interest, assigns the same score to class y in all configurations. The MS function assigns the smallest score to Case 1, since the difference between the maximum estimated probability and $\hat{p}^y$ is negligible, and attributes the highest score to Case 6 where the difference between these two values is quite large, although y has the same rank in both cases. Case 2 is considered a bit "stranger" than Case 1 by the MS function because the difference between the maximal class and y is a bit larger, which is an expected and desirable behaviour by this score function. However, Cases 3 to 5, albeit exhibiting quite different configurations, all have the same score. This is not surprising, but not desirable either: it would be important to distinguish between Case 3 where the difference between class b and class y is so negligible that class y may very well have been predicted as the second class; and Case 4 where the difference between the second and third classes is more considerable, indicating that the model is even less confident about class y in this case. Case 5, where the class of interest y is quite far away from being in the top classes is, however, still assigned a similar MS value as Cases 3 and 4.

A more flexible behaviour is exhibited by the proposed PIP score function. A first glance at $\Delta^{\text{PIP}}(y)$ shows that it does not assign the same score to any of the six different cases, showing more versatility than the other two score functions. Case 1 is assigned the lowest score, because the difference in estimated probabilities between the three classes is considered negligible. Class y in this case is not very "strange" because it could have been the first predicted class. Case 2 is slightly stranger because the difference away from the first class is larger and cannot be simply attributed to noise. Case 3 is assigned a higher score than Case 2 but a slightly smaller score than Case 4: y is equidistant from the maximum class in both cases, but in Case 2 the difference between b and y is negligible and attributable to insignificant factors. Case 5 is further penalised because more significant classes have higher estimated probabilities than y. Case 6 is attributed the highest score because class y can be considered highly "strange" in the current configuration, having a much lower $\hat{p}^y$ than the maximum class and being preceded by a class with significant difference. In brief, the PIP score function exhibits the following desirable behaviour:

- In all situations, the IP is a baseline value for the PIP function. As such, it will assign high scores for low probability classes, and lower scores for high probability classes. This kind of behaviour leads to lower average size of predicted sets since it tends to exclude the classes with low probability estimates.
- PIP allows to take into consideration the probability estimates of other classes, including the maximum probability class. In cases where y has a low value compared to the

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

maximum class, the observation will be heavily penalised: a behaviour similar to that of MS. This behaviour may lead to more predicted singletons: in cases where one class is assigned a very high probability compared to others, all the other classes will be heavily penalised and thus excluded from the predicted sets.

- PIP also allows to distinguish the cases where the difference between the probability estimates of the class of interest and the other classes is significant or not, penalising less when such differences are negligible and can be attributed to some noise.

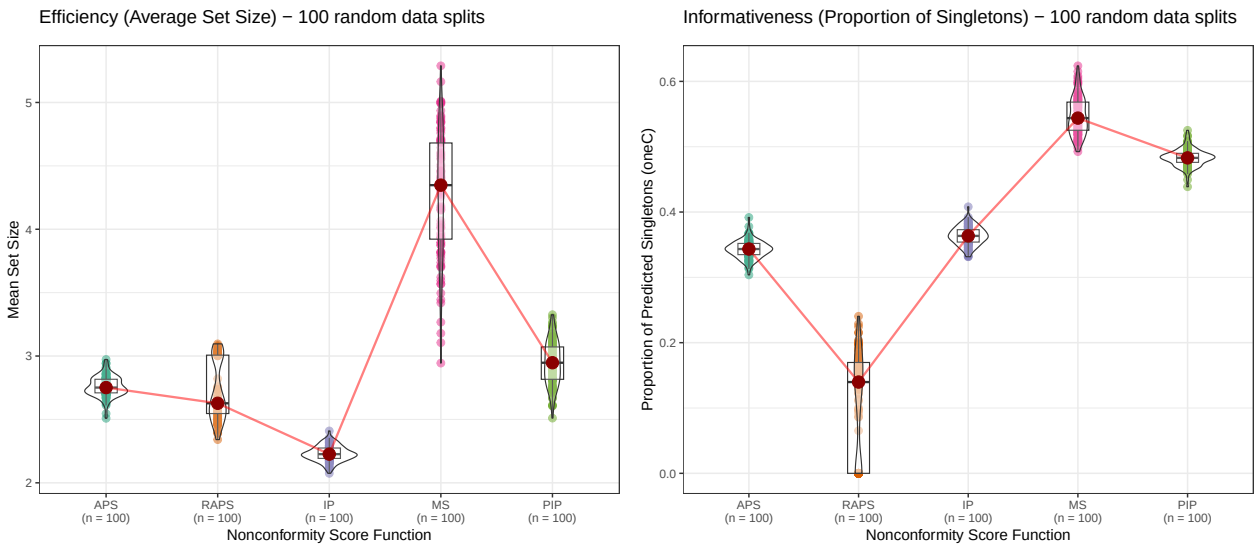
3 Comparison of Nonconformity Scores on Real Data

In this section, some empirical results studying the behaviour of the different nonconformity score functions for image classification are presented. These experiments are conducted in the context of a precision weeding application that aims at classifying images into different crop and weed species using deep neural networks (Melki et al., 2023).

3.1 Experimental Setup

The dataset considered consists of 14,800 RGB images distributed over 13 different classes obtained by dividing the original large images of the publicly available WE3DS dataset (Kitzler et al., 2023) into non-overlapping windows of size 224×224. Since the original large images also come with semantically annotated masks, the ground-truth class of each window is defined as the class with the highest number of pixels. The database is then randomly divided into: (1) a training set (70%), on which a ResNet18 classifier (He et al., 2016) is trained using default hyperparameters, 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 efficiency and informativeness of the conformal approach are evaluated.

After training the deep neural network, for each nonconformity function studied, the neural network is calibrated on the calibration set at the chosen confidence level of $1 - \alpha = 0.9$



(a) (b)

Figure 2: Boxplots of experimental results (each point is a random split): (a) *Efficiency*: PIP shows slightly larger average set sizes than APS, RAPS and Hinge – (b) *Informativeness*: PIP shows a significantly higher proportion of singletons compared to the other methods, and is competitive with the Margin Score.

and then used for constructing prediction sets on the test images. The calibration and prediction steps are repeated 100 times for each nonconformity function with a different random split of the data, in order to study the stability of the results. The RAPS hyperparemeters are fixed at $\lambda = 1$ and $k_{reg} = 3$ (Angelopoulos et al., 2021).

3.2 Results

As expected, all methods are able to maintain the required 90% marginal coverage guarantee on average. We compare the different nonconformity functions based on efficiency and informativeness. In Figure 2(a), we can see the average set size for each of the 100 runs for each nonconformity score function. The Hinge (IP) function leads to the smallest average set size and shows impressive stability over the different runs, which is in accordance with the results reported in (Johansson et al., 2017). RAPS manifests a smaller average set size than its non-regularised version (APS) (Angelopoulos et al., 2021). The Margin (MS) score function exhibits a very unstable behaviour over the different random runs, which manifests its deep dependence on the data considered. It also has a significantly higher average set size than all other methods, which echoes the results in (Johansson et al., 2017). Our proposed PIP score function is slightly less efficient than IP, APS and RAPS, but is still largely more efficient than MS.

A slight decrease in efficiency is a price to pay for the considerable gain in informativeness using PIP, as can be seen in Figure 2(b). Indeed, while MS manifests the highest proportion of singletons predicted, in accordance with Johansson et al. (2017), our proposed PIP

approach is not very away with around 50% of the predicted sets being singletons (all the while maintaining the coverage guarantee). The Hinge (IP) and APS show significantly lower informativeness, and RAPS shows the smallest number of singletons. Indeed, the proposed PIP score shows the intended behaviour of joint maximisation of the two criteria.

4 Conclusion

The current work introduced the "Penalised Inverse Probability" (PIP), a novel nonconformity function for conformal classification that can be used with any base model producing probability estimates for the predicted classes. The motivation behind PIP is the development of an elegant nonconformity function that jointly maximises efficiency and informativeness, while requiring minimal computational overhead. Using illustrative examples, the desirable behaviour of PIP in different conditions has been shown and compared to that of the classical Hinge Loss (IP) and Margin Score (MS) functions. Empirical experiments on the task of crop and weed image classification show promising results: PIP does indeed manifest the kind of behaviour it is intended for, that is, a good trade-off between maximising efficiency and maximising informativeness. During the oral presentation, further results comparing the behaviour of PIP and other nonconformity score functions on different datasets, multiple neural networks exhibiting varying levels of base accuracy, as well as other classification models will be exposed.

References

Aldous, D. J. (1985). Exchangeability and related topics. In Hennequin, P. L., editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aleksandrova, M. and Chertov, O. (2021). 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 and Applications*, pages 151–170. PMLR.

Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. (2021). Uncertainty Sets for Image Classifiers Using Conformal Prediction. In *International Conference on Learning Representations (ICLR)*, volume 2021.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Johansson, U., Linusson, H., Löfström, T., and Boström, H. (2017). Model-Agnostic Nonconformity Functions for Conformal Classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2072–2079.

Kitzler, F., Barta, N., Neugschwandtner, R. W., Gronauer, A., and Motsch, V. (2023). WE3DS: An RGB-D Image Dataset for Semantic Segmentation in Agriculture. *Sensors*, 23(5):2713.

Melki, P., Bombrun, L., Diallo, B., Dias, J., and Da Costa, J.-P. (2023). 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*, pages 614–623.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive Confidence Machines for Regression. In Elomaa, T., Mannila, H., and Toivonen, H., editors, *Machine Learning: ECML 2002*, Lecture Notes in Computer Science, pages 345–356, Berlin, Heidelberg. Springer.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with Valid and Adaptive Coverage. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc.

Shafer, G. and Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12):371–421.

Stutz, D., Dvijotham, K. D., Cemgil, A. T., and Doucet, A. (2022). Learning Optimal Conformal Classifiers. In *International Conference on Learning Representations (ICLR)*, volume 2022.

Vovk, V., Fedorova, V., Nouretdinov, I., and Gammerman, A. (2016). Criteria of Efficiency for Conformal Prediction. In Gammerman, A., Luo, Z., Vega, J., and Vovk, V., editors, *Conformal and Probabilistic Prediction with Applications*, volume 9653, pages 23–39. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York, NY.