Proceedings of Machine Learning Research 230:1–20, 2024 Conformal and Probabilistic Prediction with Applications

Multi-class Classification with Reject Option and Performance Guarantees using Conformal Prediction

Alberto García-Galindo AGARCIAGALI@UNAV.ES
Marcos López-De-Castro MLOPEZDECAS@UNAV.ES
Rubén Armañanzas RARMANANZAS@UNAV.ES
Institute of Data Science and Artificial Intelligence (DATAI), Universidad de Navarra, Ismael Sánchez Bella Building, Campus Universitario, 31009 Pamplona, Spain
TECNUN School of Engineering, Universidad de Navarra, Donostia-San Sebastián, Spain

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

Abstract

Beyond the standard classification scenario, allowing a classifier to refrain from making a prediction under uncertainty can have advantages in safety-critical applications, where a mistake may hold great costs. In this paper, we extend previous works on the development of classifiers with reject option grounded on the conformal prediction framework. Specifically, our work introduces a novel approach for inducing multi-class classifiers with reliable accuracy or recall estimates for a given rejection rate. We empirically evaluate our suggested approach in six multi-class datasets and demonstrate its effectiveness against both calibrated and uncalibrated probabilistic classifiers. The results underscore our method's capability to provide reliable error rate estimates, thereby enhancing decision-making processes where erroneous predictions bear critical consequences.

Keywords: Conformal Prediction · Multi-class Classification · Classification with Reject Option · Performance Guarantees

1. Introduction

Over recent years, the field of Machine Learning (ML) has undergone an exponential growth due to the increase of computational resources and the availability of massive databases. This has led to the proliferation of data-driven applications and the use of predictive algorithms to enhance automatic decision-making. As a consequence, ML has rapidly infiltrated critical areas of society, impacting many aspects of people's lives and influencing crucial decisions. In such settings, an erroneous decision by an ML model may have dire consequences, hence estimating the likelihood of incorrect predictions at inference-time becomes a desirable feature to gain trust in the technology.

Beyond the most common predictive scenario where an ML model makes a prediction for all the available samples, an interesting alternative is the classification with reject option (Chow, 1957; Herbei and Wegkamp, 2006), or selective classification (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017). The main motivation behind classification with reject option lies in developing suitable mechanisms to determine whether a prediction should be made or not depending on some measure of confidence. Allowing an ML model to refrain from making a prediction for uncertain samples has several advantages. For example, in medical settings, the reject option can be used to refer ambiguous cases to an expert clinician, so that an automatic evaluation is performed only when the model is confident enough (Kompa et al., 2021). In problems where each measurement has an associated cost for the patient, such abstention can also be exploited by designing sequential approaches that efficiently make use of the available predictor features, thus avoiding unnecessary harm (Garcia-Galindo et al., 2023).

The task of computing a suitable measure of confidence for a new test sample can be viewed through the lens of uncertainty quantification approaches. In this sense, the development of classifiers with reject option can be grounded on the conformal prediction framework (Vovk et al., 2005; Shafer and Vovk, 2008; Angelopoulos and Bates, 2023), which provides distribution-free statistical methods to generate calibrated prediction sets with coverage guarantees for a user-specified confidence level. The size of the prediction sets generated by a conformal predictor can be directly interpreted as uncertainty estimates, suggesting them as a promising choice for the reject option scenario. In practice, a conformal predictor would predict the samples with singleton sets, whereas refusing to make a prediction in the case of empty or multiple sets. If used in this fashion, the accuracy of the singleton predictions can not be estimated in a straightforward way. This is because the validity of a conformal predictor holds exclusively a priori, where coverage guarantees are on average, before producing any prediction. As a result, once a particular set-valued prediction is made, no assertion about the probability of containing the ground truth label can be stated. It is well known that achieving such per-sample guarantees is impossible without making further distributional assumptions (Vovk, 2013).

However, conformal predictors offer an interesting property in classification tasks: they are able to estimate error rates with guarantees in a batch of new test samples. This is achieved by means of the confidence-credibility measures (Saunders et al., 1999; Gammerman and Vovk, 2007), which accompany the most plausible predicted label with two additional values that represent the model's belief in such prediction. Specifically, recent works have shown that the confidence measure can be employed as a basis to induce binary classifiers with accuracy and precision guarantees for different rejection rates (Linusson et al., 2018; Johansson et al., 2023a,b). However, in classification problems where a false negative can have critical consequences (e.g., an undetected disease), these guarantees may not be sufficient.

Contribution: Our work draws inspiration from Johansson et al. (2023b), which proposed the use of conformal prediction to build binary classifiers with reject option. Specifically, they suggested employing a rejection mechanism based on the confidence score, resulting in classifiers with both accuracy and precision guarantees for different abstention rates. In this paper, we formalize such approach and extend it to multi-class classification with reject option. Our study presents procedures that accurately estimate the performance of any black-box rejection-based classifier grounded on conformal prediction. We also propose a novel procedure aiming to control the per-class true positive rate, particularly useful in classification problems where a class-specific false negative have critical consequences.

Outline: The rest of the paper is organized as follows. Section 2 introduces the classification with reject option task, probabilistic classification, the Platt scaling method, and the conformal prediction framework. In Section 3, we describe the proposed method to induce multi-class classifiers with reliable accuracy and recall estimates. Section 4 covers the empirical experimentation, including results and discussion. Finally, Section 5 presents the main conclusions and provides potential lines for future research.

2. Background

2.1. Classification with reject option

For the sake of this study, we consider a general multi-class classification task. Let $\mathcal{X}$ the input feature space, and $\mathcal{Y} = \{c_1, c_2, \ldots, c_J\}$ the target space with $J \in \mathbb{N}$ possible classes. Let us assume that the inputs and targets are generated by a data generating process with $P_{\mathcal{X}\mathcal{Y}}$. In the classification with reject option setting, the target space is augmented to allow an abstention alternative, i.e., not making a prediction, that is $\mathcal{Y}' = \{c_1, c_2, \ldots, c_J\} \cup \{\mathsf{reject}\}$. Diverging from the usual classification problem, we introduce a *classifier with reject option* (h, g) (Geifman and El-Yaniv, 2017), which is composed of a standard *classification model* $h : \mathcal{X} \to \Delta^J$, where

$$\Delta^J = \left\{ \pi_j \in [0,1] \,\middle|\, \sum_{j=1}^{J} \pi_j = 1 \right\} \tag{1}$$

is the J-dimensional probability simplex, and a *selection function* $g : \mathcal{X} \to \{0, 1\}$, formally:

$$(h,g)(x) = \begin{cases} \arg\max_{\bar{c}_j} \pi_j & \text{if } g(x) = 1 \\ \mathsf{reject} & \text{if } g(x) = 0 \end{cases} \tag{2}$$

For a given fixed classifier h, the central task when constructing a classifier with reject option lies in defining a suitable selection function g. Intuitively, a suitable selection function should rank samples based on how challenging they are to predict accurately. In practice, the performance of a classifier with reject option is quantified using prediction coverage[1] and selective risk. The prediction coverage $\mathbb{E}_{\mathcal{X}}[g(x)]$ is defined as the expected rate of the non-rejected samples, whereas, for a given real-valued loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, the selective risk is the expected loss for the non-rejected samples, formally:

$$R(h,g) = \frac{\mathbb{E}_{\mathcal{X}\mathcal{Y}}[\ell(h(x), y) g(x)]}{\mathbb{E}_{\mathcal{X}}[g(x)]} \tag{3}$$

For the remainder of this paper, the loss function we consider is taken to be the standard misclassification loss function: $\ell(h(x), y) = \mathbb{I}\,(h(x) \neq y)$, where $\mathbb{I}$ is the indicator function that is 1 when its argument is true and 0 otherwise. In practice, the risk of a selective classifier trades off for prediction coverage: the higher the number of rejected samples is, the lower the selective risk is. In that sense, it becomes interesting to evaluate the performance of a classifier with reject option for different abstention rates, which is usually done by means of the risk-coverage curve.

2.2. Probabilistic classification and calibration

Most ML classifiers are probabilistic predictors, as they are able to produce a posterior score π_j[2] for each label $c_j \in \{c_1, \ldots, c_J\}$. In this way, a probabilistic predictor can be used

[1] The reader should note that this notion of coverage differs from the ground truth coverage definition usually employed to refer to the validity guarantees of a conformal predictor.

[2] We assume without loss of generality that $\pi_j \in [0, 1]$. If this is not the case, each π_j should be properly scaled.

not only to produce a point prediction, but also accompany this prediction with a measure of confidence, specially useful in the reject option scenario. However, to be valuable in decision-making, these posterior scores need to be calibrated, i.e., reflect true probabilities. A probabilistic predictor is considered well calibrated if, given a predicted score for a certain label, the probability of that label matches the predicted score, formally:

$$\mathbb{P}(y = c_j \,|\, \pi_j) = \pi_j \tag{4}$$

It should be noted that calibrated predictors present an interesting property when constructing classifiers with reject option. If a classifier is well calibrated, and we evaluate, for example, a set of predictions whose highest posterior score among all the classes is around 0.8, we would expect an 80% accuracy in such predictions. Hence, we could define an abstention mechanism which ranks samples according to their highest posterior score, and induce a rejection model.

However, despite the fact that the majority of classification algorithms output posterior scores, these are normally poorly calibrated and do not represent true probabilities, including tree-based models (Provost and Domingos, 2003; Boström, 2008; Johansson et al., 2021), support vector machines (Acevedo et al., 2007), logistic regression (Bai et al., 2021), and neural networks (Guo et al., 2017; Minderer et al., 2021).

In the literature, the induction of well-calibrated probabilistic classifiers has been approached through post-hoc procedures that modify the posterior score distribution to compute more accurate probability estimates, usually employing a hold-out dataset, namely the *calibration set*. Some of the most well-known standard methods for calibrating trained classifiers include Platt scaling (Platt et al., 1999), isotonic regression (Zadrozny and Elkan, 2002), and Venn (-Abers) predictors (Vovk et al., 2005; Vovk and Petej, 2014), the latter being the only alternative with strong theoretical guarantees. In this work, we use the Platt scaling method as a standard benchmark for comparing our methodological proposal.

2.3. Platt scaling

Platt scaling (Platt et al., 1999), originally proposed in the context of support vector machines, is a calibration technique that modifies the posterior scores π_j from a classifier by fitting a parametric sigmoid function to these π_j on the calibration dataset. Formally:

$$\hat{\mathbb{P}}(y = c_j \,|\, \pi_j) = \frac{1}{1 + e^{w_0 + \pi_j w_1}} \tag{5}$$

where $\hat{\mathbb{P}}(y = c_j \,|\, \pi_j)$ is the estimated probability that a sample belongs to class c_j given the posterior score π_j, and w_0 and w_1 are parameters learned using maximum likelihood estimation.

2.4. Conformal Prediction

Given a supervised learning task, the conformal prediction framework provides statistical procedures to quantify reliable levels of individual uncertainty by means of valid set-valued predictions (Vovk et al., 2005; Shafer and Vovk, 2008; Angelopoulos and Bates, 2023).

Let us initially assume a multi-class classification setting and an available set of training samples $\{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in \mathcal{X} \times \mathcal{Y}$ drawn exchangeably from $P_{\mathcal{X}\mathcal{Y}}$. The general goal of conformal prediction is to construct a conformal predictor $\Gamma_{\mathsf{set}} : \mathcal{X} \to 2^{\mathcal{Y}}$ with ground truth coverage guarantees at a desired confidence level $1 - \alpha \in [0, 1]$. To this end, a central component of conformal prediction is the definition of a non-conformity measure $\mathcal{S} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, used to quantify the degree of relative strangeness of a new sample z with respect to a collection of samples $\{z_1, \ldots, z_k\}$. The non-conformity measure is a design choice and, although it can be any measurable function, it is usually based on the output (in fact, the posterior scores π_j) of a predictive model, formally:

$$\mathcal{S}(z, \{z_1, \ldots, z_k\}) = \Xi\,(y, h(x)), \tag{6}$$

where $h : \mathcal{X} \to \mathcal{Y}$ is, in our case, a multi-class classifier learned on $\{z_1, \ldots, z_k\}$ and $\Xi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a function of dissimilarity between the true class y and the prediction $h(x)$. In this way, how strange is a sample z is quantified through the ability of a model learned on $\{z_1, \ldots, z_k\}$ to precisely predict its true class.

Despite being initially proposed within a transductive scheme, learning a conformal predictor using this approach is frequently computationally infeasible since it requires fitting the classifier for each new sample. To address this challenge, inductive (or split) conformal prediction (Papadopoulos et al., 2002; Lei et al., 2018) was devised to offer a more efficient variant for conformal calibration. As with classical calibration algorithms, the inductive conformal prediction procedure makes use of a hold-out dataset to calibrate a trained classifier. For this reason, the original training samples $\{z_i\}_{i=1}^{m}$ are typically partitioned into a proper training set $\{z_i\}_{i=1}^{n}$, used to learn a single classifier h, and a calibration set $\{z_i\}_{i=n+1}^{m}$, on which to compute the non-conformity scores. For a new test sample x_{new}, conformal prediction mainly relies on comparing the strangeness (i.e., the non-conformity) of a tentative label $\bar{c}_j \in \mathcal{Y}$ with respect to the calibration set and computes a *p-value* to evaluate the hypothesis that $\bar{c}_j$ is the actual label y_{new}. Hence, given

$$\begin{aligned} s_i &= \mathcal{S}(z_i, \{z_1, \ldots, z_n\}) \qquad i = n+1, \ldots, m \\ s_{\mathsf{new}} &= \mathcal{S}(z_{\mathsf{new}}, \{z_1, \ldots, z_n\}), \end{aligned} \tag{7}$$

we can compute the *p-value* for $\bar{c}_j$ as follows:

$$p_j = \frac{|\{i = n+1, \ldots, m : s_i \geq s_{\mathsf{new}}\}| + 1}{m - n + 1} \tag{8}$$

This process is repeated for each possible label $\bar{c}_j \in \mathcal{Y}$, thus producing a set of *p-values* $(p_1, \ldots, p_J)$ which reflects the model's uncertainty for the new sample. At this point, the set of *p-values* is usually employed to define, given a user-specified significance level $1 - \alpha \in [0, 1]$, a set predictor $\Gamma_{\mathsf{set}} : \mathcal{X} \to 2^{\mathcal{Y}}$ that, for a new test sample x_{new}, outputs all the classes whose *p-value* is greater than α, formally:

$$\Gamma_{\mathsf{set}}(x_{\mathsf{new}}) = \{c_j \in \mathcal{Y} \,|\, p_j > \alpha\} \tag{9}$$

This set predictor is guaranteed to contain the true class with $1 - \alpha$ probability as long as $P_{\mathcal{X}\mathcal{Y}}$ is exchangeable:

$$\mathbb{P}\big(y_{\mathsf{new}} \in \Gamma_{\mathsf{set}}(x_{\mathsf{new}})\big) = 1 - \alpha,\ \ \alpha \in [0, 1] \tag{10}$$

In this work, we focus on a less typical output choice to produce a confidence prediction: the confidence-credibility measures (Saunders et al., 1999; Gammerman and Vovk, 2007). This type of output from a conformal classifier is also based on the set of *p-values* generated by a conformal predictor, but is better suited for scenarios where a point prediction is required. Let $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(J)}$ the order statistics of $(p_1, \ldots, p_J)$ (i.e., the set of *p-values* sorted in decreasing order). Specifically, three different values are reported for a particular sample:

- The forced prediction $f(x)$, which is given by the most likely classification (i.e., the tentative class with the largest *p-value*).

$$f(x) = \underset{\bar{c}_j \in \mathcal{Y}}{\arg\max}\, p_j \tag{11}$$

- The confidence $\lambda(x)$, defined as one minus the second highest *p-value*.

$$\lambda(x) = 1 - p_{(2)} \tag{12}$$

- The credibility $\gamma(x)$, defined as the highest *p-value*.

$$\gamma(x) = p_{(1)} \tag{13}$$

In this way, a confidence-credibility predictor $\Gamma_{\mathsf{cc}} : \mathcal{X} \to \mathcal{Y} \times [0,1] \times [0,1]$ is defined. For a new test sample x_{new}, Γ_{cc} outputs the combination of these three values, formally:

$$\Gamma_{\mathsf{cc}}(x_{\mathsf{new}}) = \big(f(x_{\mathsf{new}}), \lambda(x_{\mathsf{new}}), \gamma(x_{\mathsf{new}})\big) \tag{14}$$

The confidence and credibility measures present a suitable way to augment a point prediction with additional information, serving as indicators of reliability. Intuitively, the higher the values of both confidence and credibility are, the more reliable the prediction is. Additionally, these two measures can be interpreted in the context of set prediction with interesting properties: the confidence corresponds to the highest significance level where the conformal predictor outputs a singleton prediction, whereas the credibility represents the significance level from which all classes are rejected.

Note that the properties of the confidence measure provide a convenient background for the induction of reliable classifiers in the reject option scenario. Specifically, for a given confidence $\lambda \in [0,1]$, a confidence-credibility predictor Γ_{cc} is guaranteed to achieve an expected correct classification rate λ for all the forced predictions whose related confidence is greater or equal to λ, formally:

$$\mathbb{P}(f(x) = y \,|\, \lambda(x) \geq \lambda) = \lambda,\ \ \lambda \in [0,1] \tag{15}$$

In other words, Γ_{cc} can be used as a classifier with reject option. It is worth mentioning that, although the guarantees are theoretically granted for any λ, in practice the accuracy level that Γ_{cc} is able to achieve will depend on several factors, such as its efficiency or the prediction task difficulty.

Both coverage and accuracy guarantees considered so far are only satisfied on average. However, an extension of the standard conformal prediction procedure, the so-called Mondrian conformal prediction (Vovk et al., 2005), was proposed to mitigate this limitation and construct confidence predictors with guarantees in a group-wise style. The main difference relies on splitting the calibration dataset into categories predefined by a taxonomy and separately running the calibration step on each category. Formally, a taxonomy is a measurable function $\kappa : \mathcal{X} \times \mathcal{Y} \to \mathcal{K}$ that maps each sample z_i to a specific category $\kappa_i \in \mathcal{K}$, where $\mathcal{K}$ is normally a discrete space.

Similarly to its standard version, a Mondrian conformal predictor can be learned by induction. Hence, given a taxonomy κ, for a new test sample x_{new} belonging to the category κ_{new}, the *p-value* for each of the tentative class $\bar{c}_j \in \mathcal{Y}$ is given by:

$$p_j = \frac{|\{i = n+1, \ldots, m : \kappa_i = \kappa_{\mathsf{new}}, s_i > s_{\mathsf{new}}\}| + 1}{|\{i = 1, \ldots, m+1 : \kappa_i = \kappa_{\mathsf{new}}\}|}, \tag{16}$$

where s_i and κ_i are the non-conformity score and the category related to the i-th calibration sample, respectively. When using a Mondrian conformal predictor, both guarantees (10) and (15) are satisfied for each category individually. For the sake of this study, we consider the Mondrian taxonomy $\kappa : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$, where the categories are given by the actual label $y_i \in \mathcal{Y}$ of each sample z_i. This particular case is often referred to as class-conditional conformal prediction, and provides a useful extension to induce classifiers with reliable per-class recall estimates.

3. Methods

In this section, we propose a general procedure to induce multi-class classifiers with a reject option whose abstention mechanism is based on conformal prediction. The proposed methodology seeks reliable performance estimates for accuracy or recall on a batch of non-rejected test samples. For comparison purposes, we also examine a benchmark abstention mechanism derived from the posterior scores of a probabilistic classifier.

3.1. Score-based rejection

Let us assume a multi-class probabilistic classifier $h : \mathcal{X} \to \Delta^J$ which generates a posterior score π_j for each label $c_j \in \{c_1, \ldots, c_J\}$. Since the posterior scores can be seen as a measure of confidence for each of the possible labels, it makes sense to propose rejection mechanisms based on them. Specifically, for a rejection threshold $\theta \in [0, 1]$, we define the *maximum score selection function* $g_\pi : \mathcal{X} \to \{0, 1\}$ as follows:

$$g_\pi(x) = \begin{cases} 1, & \text{if } \max_j \pi_j \geq \theta; \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

Hence, given g_π, the classifier with reject option (h, g_π) produces a prediction only for those cases where the maximum posterior score is above the rejection threshold. If h is perfectly-calibrated, the selective risk of (h, g_π) is guaranteed to be

$$R(h, g_\pi) = \frac{\mathbb{E}_{\mathcal{X}\mathcal{Y}}\left[\mathbb{I}\left(\arg\max \pi_j \neq y\right) g_\pi(x)\right]}{\mathbb{E}_{\mathcal{X}}\left[g_\pi(x)\right]} = 1 - \mathbb{E}_{\mathcal{X}\mathcal{Y}}\left[\max \pi_j \middle| \max \pi_j \geq \theta\right] \tag{18}$$

i.e., (h, g_π) achieves accuracy guarantees of $\mathbb{E}_{\mathcal{X}\mathcal{Y}}[\max \pi_j | \max \pi_j \geq \theta]$.

We also propose a similar rejection mechanism based on the posterior score of a specific label. For a given label c_j, we define the *class score selection function* $g_{\pi_j} : \mathcal{X} \to \{0,1\}$ as follows:

$$g_{\pi_j}(x) = \begin{cases} 1, & \text{if } \pi_j \geq \theta; \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

When considering only the samples belonging to the label c_j, if h is perfectly calibrated, the selective risk of (h, g_{π_j}) on such samples is guaranteed to be

$$R(h, g_{\pi_j} \,|\, y = c_j) = \frac{\mathbb{E}_{\mathcal{X}\mathcal{Y}}[\mathbb{I}\,(\arg\max \pi_j \neq y) g_{\pi_j}(x) \,|\, y = c_j]}{\mathbb{E}_{\mathcal{X}}[g_{\pi_j}(x) \,|\, y = c_j]} = 1 - \mathbb{E}_{\mathcal{X}\mathcal{Y}}[\pi_j \,|\, \pi_j \geq \theta, y = c_j] \tag{20}$$

i.e., (h, g_{π_j}) achieves recall guarantees for the class c_j of $\mathbb{E}_{\mathcal{X}\mathcal{Y}}[\pi_j \,|\, \pi_j \geq \theta, y = c_j]$.

3.2. Conformal-based rejection

Let us now assume a confidence-credibility predictor $\Gamma_{\mathsf{cc}} : \mathcal{X} \to \mathcal{Y} \times [0,1] \times [0,1]$. In this case, a rejection mechanism specified by the confidence score $\lambda(x)$ is proposed. Thus, for a given rejection threshold $\theta \in [0,1]$, we define the *conformal selection function* $g_\lambda : \mathcal{X} \to \{0,1\}$ as follows:

$$g_\lambda(x) = \begin{cases} 1, & \text{if } \lambda(x) \geq \theta; \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Through this mechanism, only the samples whose confidence is above θ are predicted. In this case, θ is reported as the expected error rate of the predicted samples. Using this method, we expect to observe an average accuracy rate of $1 - \theta$ in a batch of test samples.

When seeking reliable recall estimates, we propose a similar procedure that is also based on the confidence measure, but instead the Mondrian taxonomy $\kappa : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ is now used to calibrate the underlying classifier. Here, θ is again reported as the expected error rate, but this time only considering the target class. Hence, we now expect to observe an average recall of $1 - \theta$ in a batch of test samples belonging to this specific class.

4. Empirical experimentation

4.1. Experimental setup

In the experimental phase, we compared the two abstention mechanisms with the goal of developing multi-class classifiers with reject option and performance guarantees. For the score-based rejection, we considered off-the-shelf uncalibrated multi-class classifiers (Uncal) and the Platt scaling method (Platt). In both cases, the posterior scores were used as a basis for refraining from making a prediction. Alternatively, for the conformal-based rejection (Conf), the confidence measure was used to filter uncertain samples. We propose a performance evaluation scheme by setting different rejection rates, ranging from $\tau \in \{0.1, 0.2, \ldots, 0.9\}$.

- For the Uncal and Platt strategies, given a batch of test samples, an initial ranking according to the posterior scores was performed. In the case of the accuracy estimation experiment, the maximum posterior score was used. In the case of the recall estimation experiment, the score for a selected specific class was used. Then, the bottom $(1-\tau)\%$ of the test samples are rejected, leaving the remaining top $\tau\%$ to be predicted and evaluated. In this way, the posterior score averaged across the $\tau\%$ samples was reported as the correct classifications rate that the rejection-based classifier, if well-calibrated, should guarantee.

- For the Conf strategy, a similar policy was adopted. However, in this case, the ranking of the test samples was carried out based on the confidence measure. Hence, the confidence score λ from which the $\tau\%$ predictions were evaluated was reported as the correct classifications rate that the confidence-credibility predictor estimates. In the case of accuracy estimation, a standard conformal classifier was used, whereas in the case of recall estimation, a Mondrian conformal classifier was used.

We restricted the experimental study to inherently multi-class classifiers (i.e., models that automatically produce a probability score of each of the class labels by construction. Specifically, decision tree (*tree*) (Breiman et al., 1984) and random forest (*rfc*) (Breiman, 2001) algorithms were tested. The multi-class classifiers were induced using the scikit-learn library (Pedregosa et al., 2011) with default hyperparameter configurations, except for the number of minimum samples required in a leaf node of the decision trees, which was set to 10. The Platt scaling calibration was carried out using the *CalibratedClassifierCV* method from the same library. The conformal calibration was conducted out using the crepes library (Boström, 2022), setting the hinge loss function (also known as least ambiguous set-valued classifier (Sadinle et al., 2019)) as the non-conformity measure. Formally:

$$\Xi\left(y_i, h(x_i)\right) = 1 - h(x_i)_{y_i}, \tag{22}$$

where $h(x_i)_{y_i}$ is the estimated posterior score provided by the multi-class classifier h for the sample x_i and the label y_i.

To evaluate the rejection classifiers, the testing protocol was a repeated hold-out procedure with 100 repetitions of a 75/25 % random split into training and test sets, respectively. For calibration purposes, each training set was further partitioned into 2/3 of proper training samples and 1/3 of calibration samples. For the uncalibrated models, the multi-class classifiers were induced using the complete training set. Each split was class stratified. To assess the quality of the estimates, we computed the absolute estimation error between expected and observed accuracy or recall for the subset of non-rejected test samples.

4.2. Datasets

We tested our method on six different multi-class classification datasets, including one synthetic dataset and five publicly available real-world datasets (see Table 1). Each dataset varies in complexity, from the number of samples and features to the cardinality of the label space, providing a diverse test bed.

For the estimation of recall, we only considered one predetermined class within each dataset (see Table 2).

Table 1: Description of datasets, including number of samples (n), number of features (p), cardinality of the label space ($|\mathcal{Y}|$) and source.

Dataset	n	p	$\vert\mathcal{Y}\vert$	Source
adult	49,531	10	3	(Ding et al., 2021)
beans	13,611	16	6	UCI Machine Learning
ocr	5,620	64	9	UCI Machine Learning
cars	1,728	6	4	UCI Machine Learning
synthetic	1,000	8	3	scikit-learn (`make_classification`)
glass	214	9	6	UCI Machine Learning

Table 2: Target class information for each dataset.

Dataset	Class description	Proportion
adult	Income $\leq$ \$20K	0.399
beans	Sira dry bean type	0.194
ocr	Digit nine	0.100
cars	Car with an acceptable evaluation	0.222
synthetic	Synthetic class ($y = 1$)	0.333
glass	Headlamp	0.136

4.3. Baseline performance

We conducted an initial experiment to set a baseline for the rejection scenarios by predicting all the available test samples (i.e., standard classification) using each strategy. The baseline predictive performance, measured in terms of overall accuracy and recall, is presented in Table 3.

Table 3: Baseline predictive performance.

	tree						*rfc*					
Dataset	Accuracy			Recall			Accuracy			Recall		
	Uncal	Platt	Conf	Uncal	Platt	Conf	Uncal	Platt	Conf	Uncal	Platt	Conf
adult	.675	.764	.671	.758	.671	.760	.677	.758	.675	.762	.674	.756
beans	.909	.906	.905	.845	.846	.844	.923	.921	.921	.861	.862	.862
ocr	.878	.892	.858	.872	.853	.875	.982	.979	.978	.944	.942	.944
cars	.938	.926	.924	.872	.876	.883	.980	.969	.967	.975	.951	.958
synthetic	.755	.734	.734	.734	.717	.708	.863	.850	.852	.849	.842	.835
glass	.945	.931	.935	.968	.969	.958	.974	.967	.966	.921	.910	.904

Note that some datasets were more challenging than others. The uncalibrated classifiers yielded slightly better accuracy results than the Platt-scaled models and the conformal predictors. This is likely due to the fact that the uncalibrated models make use of more training data to induce the classifiers. Additionally, we can observe that, regardless of the strategy, the random forest outperformed the decision trees for all the datasets.

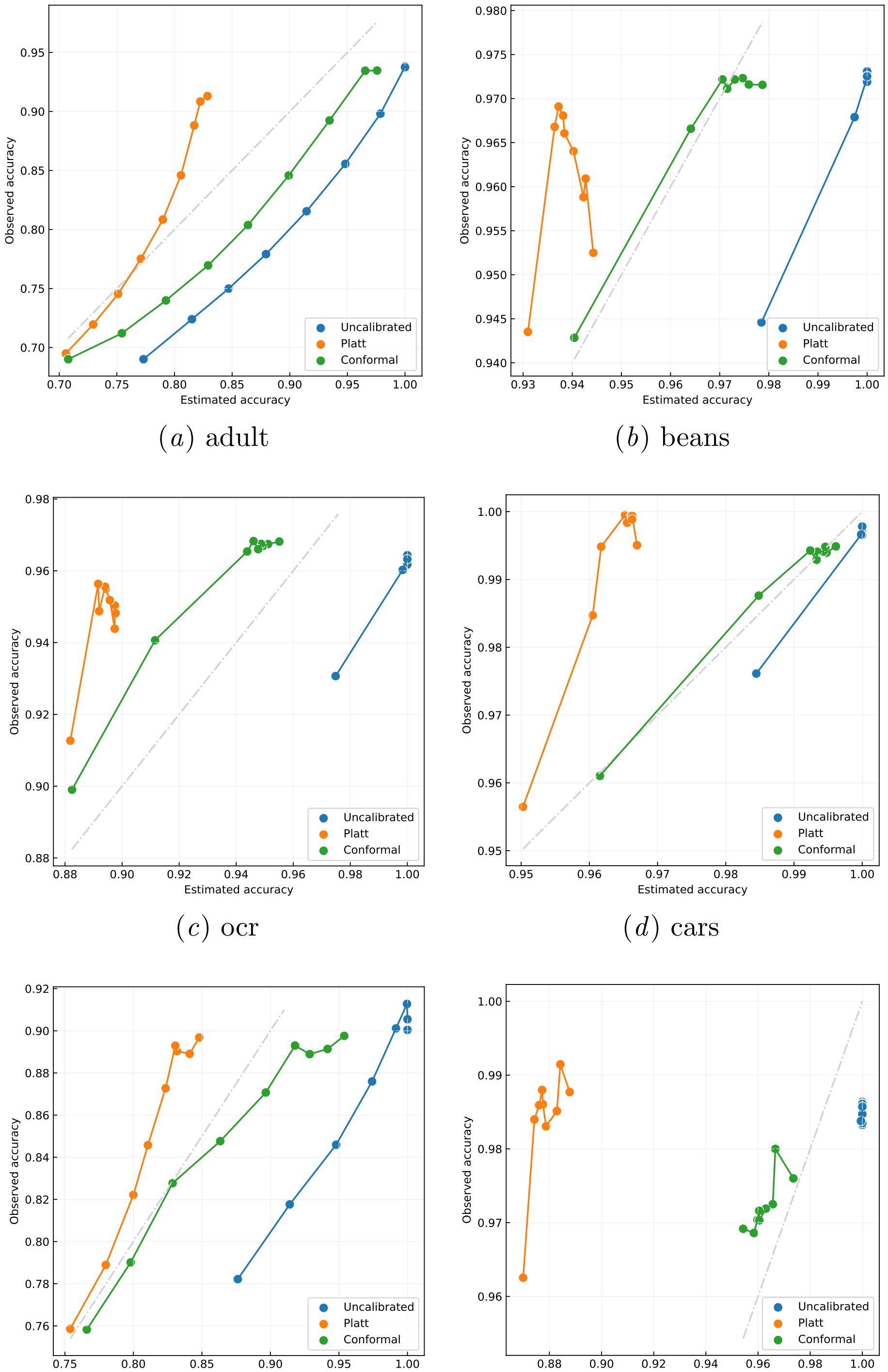
4.4. Accuracy estimation

The quality of the accuracy estimates using decision tree as the underlying classifier is visually presented in Figure 1. Each point represents a certain rejection rate, with lines connecting these points to illustrate the performance progression across successive rates. The estimated accuracy is plotted along the x-axis, whereas the empirical accuracy in the non-rejected samples is plotted along the y-axis. The conformal predictors achieved the best results, showing better accuracy estimates than both the Platt-scaled probabilities and the uncalibrated models. The uncalibrated models tended to be over-confident in their predictions, whereas the Platt scaling method seemed to be excessively cautious when estimating the confidence. The conformal predictors estimates' lied somewhere between those of the uncalibrated models and the Platt scaling method, for all the datasets. It can be noted that, in some cases such as the beans and cars datasets, the accuracy estimates by the conformal classifiers practically matched the empirical rates, achieving the desired statistical guarantees. However, in other cases, namely the adult and the ocr datasets, the accuracy estimates were systematically over-confident and under-confident, respectively. The accuracy improvement achieved from the rejection of the most uncertain samples varied across classification tasks. For example, in the case of the synthetic dataset (Figure 1(e)), the accuracy progressively increased to 90%, which corresponded to abstaining from making a prediction for 60% of the samples. Beyond this rejection rate, there was no additional improvement regardless of the rejection stress. On the other hand, other datasets, e.g., cars (Figure 1(d)), only showed an accuracy improvement up to the rejection of the most 30% uncertain samples.

The Platt scaling method failed to provide a suitable confidence ranking for some datasets. This was particularly noteworthy in the beans dataset (Figure 1(b)), where from the 30% rejection rate onwards, the accuracy of the evaluated samples began to decrease. However, this method performed well on the adult dataset (Figure 1(a)), specially for lower rejection rates, where the conformal classifier over-estimated the accuracy levels.

The accuracy estimation results using the random forest as the underlying classifier are presented in Figure 2. As with the decision tree, the conformal classifiers also outperformed both the Platt scaling method and the uncalibrated models, except for the adult dataset, where the Platt-scaled probabilities provided the best estimates. In the remaining datasets, the conformal classifiers produced near perfect accuracy estimates for all the rejection levels. Conversely, the uncalibrated models and the Platt-scaled classifiers showed very different behaviors depending on the dataset. In general terms, they tended to under-estimate the actual accuracy rates although, in very specific cases, their estimates were similar to those of the conformal models (see the uncalibrated classifiers in the beans (Figure 2(b)) and in the glass (Figure 2(f)) datasets).

The accuracy estimation errors made by each strategy confirm that the conformal classifiers were the best alternative on average (see Table 4).



(a) adult

(b) beans

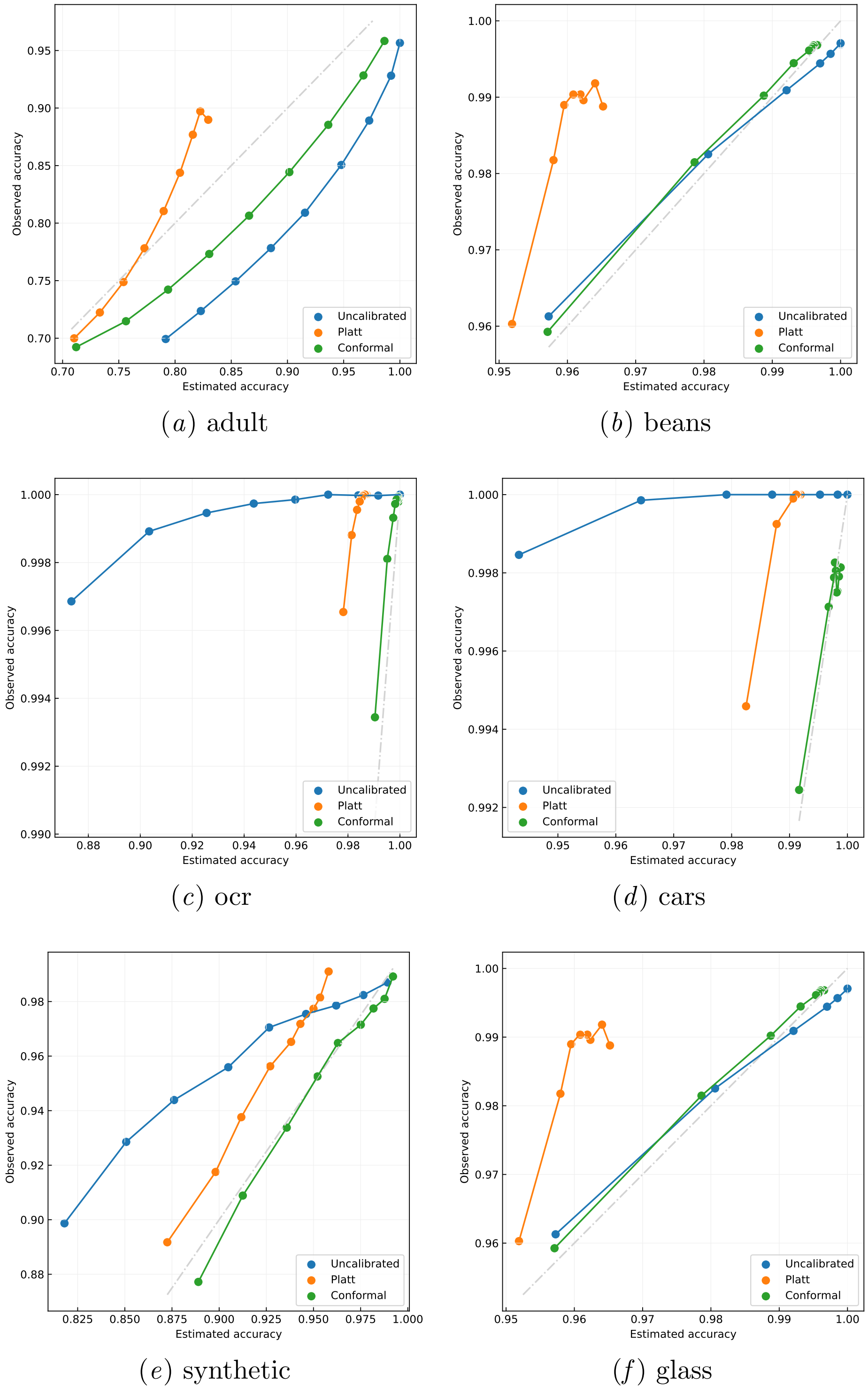
(c) ocr

(d) cars

(e) synthetic

(f) glass

Figure 1: Accuracy estimation using the decision tree as the underlying classifier for each dataset. The dotted gray line represents a perfectly calibrated classifier.



(*a*) adult

(*b*) beans

(*c*) ocr

(*d*) cars

(*e*) synthetic

(*f*) glass

Figure 2: Accuracy estimation using the random forest as the underlying classifier for each dataset. The dotted gray line represents a perfectly calibrated classifier.

Table 4: Accuracy estimation error averaged over all rejection rates for each dataset. Bold numbers highlight estimation errors from the conformal classifiers that were significantly lower with respect to those from both the uncalibrated and the Platt-scaled models (Friedman test of differences ($\alpha = 5\%$) and Nemenyi test to assess pairwise differences).

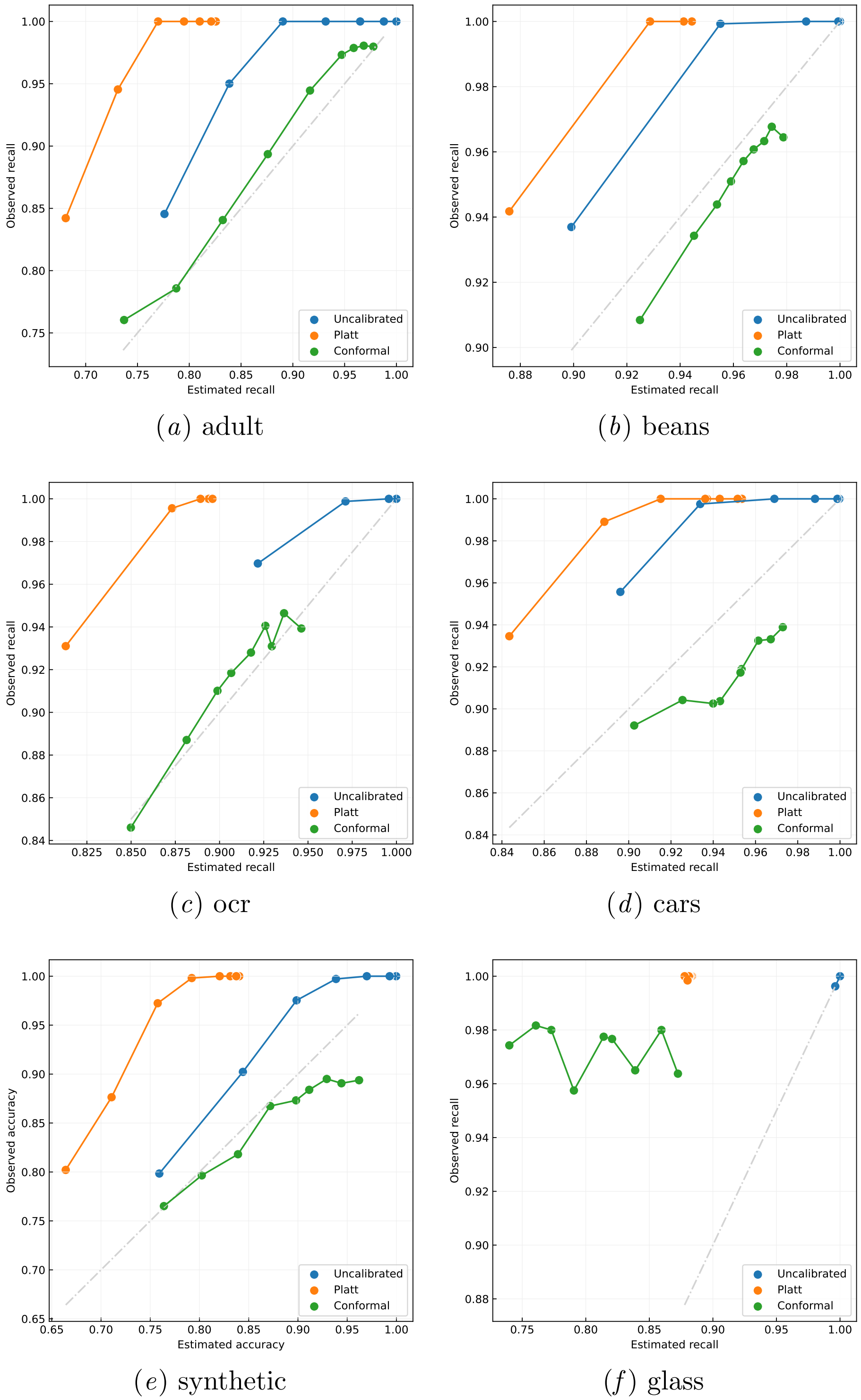
	tree			*rfc*		
Dataset	Uncal	Platt	Conf	Uncal	Platt	Conf
adult	.085	.037	.044	.089	.032	.045
beans	.029	.022	**.002**	.003	.025	**.002**
ocr	.038	.053	**.019**	.049	.015	**.001**
cars	.003	.029	**.001**	.015	.009	**.001**
synthetic	.094	.037	**.025**	.042	.026	**.004**
glass	.015	.105	**.010**	.047	.097	**.005**

4.5. Recall estimation

The performance of the different strategies for the recall estimation task in each dataset using the decision tree and the random forest as the underlying classifiers is presented in Figures 3 and 4, respectively. The uncalibrated and the Platt-scaled probabilistic models consistently under-estimated the number of correctly classified samples belonging to the targeted class in all datasets. In fact, it is interesting to observe that both strategies yielded a 100% recall for rejection levels beyond the 30%, even when the estimates were under 85% (see, for example, the uncalibrated classifiers in Figure 4(e) and the Platt method in Figure 4(a)).

Turning the attention to the Mondrian conformal classifiers, the displayed results clearly highlight the capability of the method to produce reliable recall estimates for the targeted classes. In contrast to the uncalibrated and the Platt-scaled predictors, the Mondrian conformal classifiers avoided underestimating the observed recall, instead yielding better estimates for all rejection rates. The usefulness of our method with respect to the probabilistic predictors was particularly evident in the low rejection rates, in which the uncalibrated models and the Platt-scaled classifiers failed to provide reliable recall by producing large estimation errors. In fact, the Platt-scaled models produced, in most cases, larger estimation errors than the uncalibrated classifiers. The Mondrian models yielded precise levels of recall for large-data regime datasets like beans, ocr, or adult, where the estimates practically matched the empirical rates for all the rejection rates. This was not the case of the glass dataset (Figure 3(f) and Figure 4(f)), for which the extremely low number of samples for the targeted class led to a failure when attempting to calibrate the Mondrian conformal predictors and to properly rank the uncertain samples. In this case, the uncalibrated models achieved the lowest estimation errors, regardless of the underlying model.

The recall estimation errors made by each strategy are reported in Table 5, demonstrating that the Mondrian conformal classifiers were the best option.



(*a*) adult

(*b*) beans

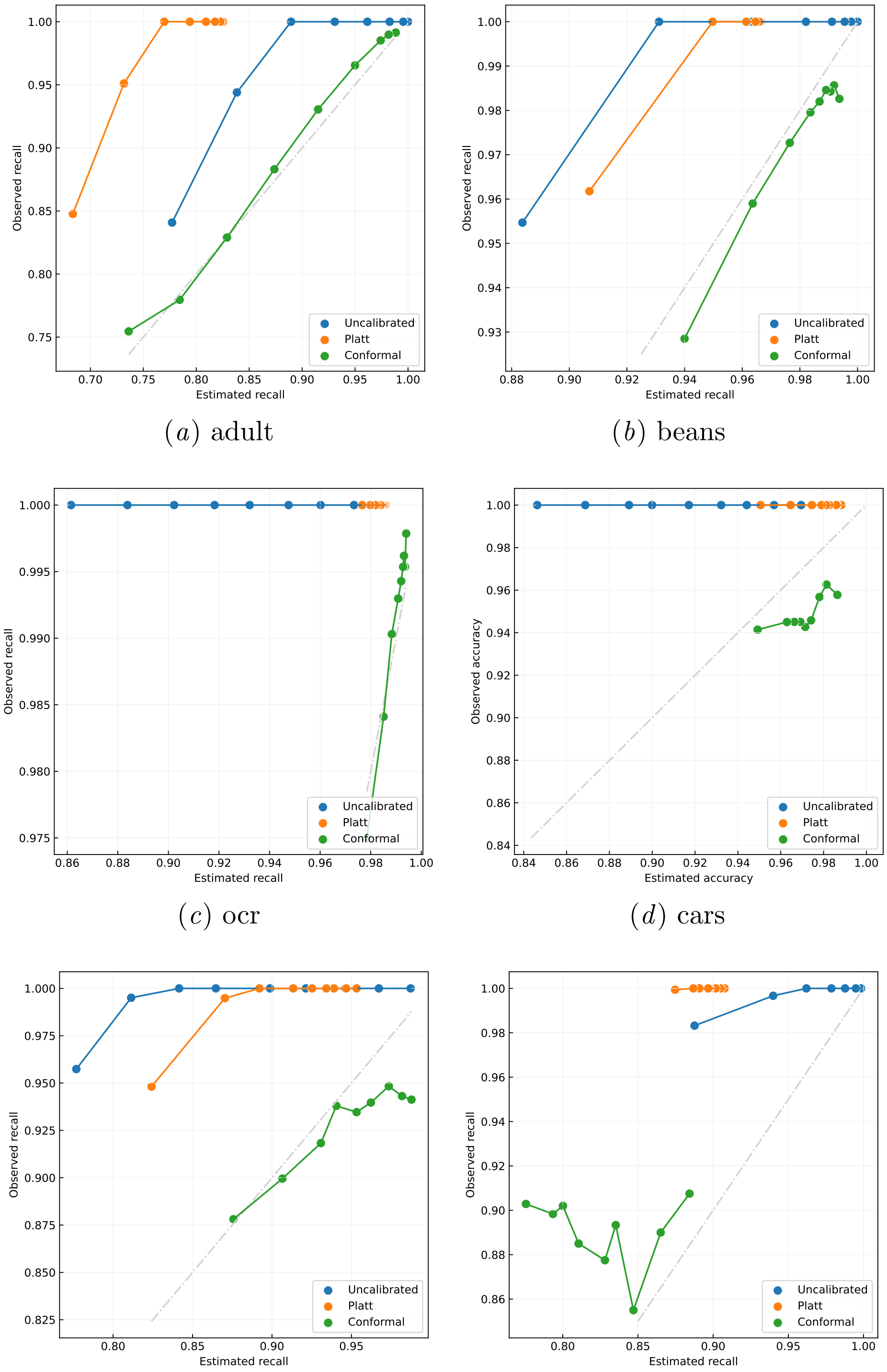
(*c*) ocr

(*d*) cars

(*e*) synthetic

(*f*) glass

Figure 3: Recall estimation using the decision tree as underlying classifier for each of the considered datasets. The dotted gray line represents a perfectly calibrated classifier.



(a) adult

(b) beans

(c) ocr

(d) cars

(e) synthetic

(f) glass

Figure 4: Recall estimation using the random forest as underlying classifier for each of the considered datasets. The dotted gray line represents a perfectly calibrated classifier.

Table 5: Recall estimation error averaged over all rejection rates for each dataset. Bold numbers highlight estimation errors from the conformal classifiers that were significantly lower with respect to those from both the uncalibrated and the Platt-scaled models (Friedman test of differences ($\alpha = 5\%$) and Nemenyi test to assess pairwise differences).

	tree			*rfc*		
Dataset	Uncal	Platt	Conf	Uncal	Platt	Conf
adult	.045	.189	**.015**	.045	.191	**.009**
beans	.011	.060	**.010**	.011	.060	**.006**
ocr	.009	.110	.008	.071	.017	**.002**
cars	.019	.067	.031	.080	.023	.022
synthetic	.030	.173	**.027**	.104	.083	**.020**
glass	.000	.119	.165	.031	.104	.072

5. Conclusions

In high risk domains, the reliable quantification of uncertainty becomes a key feature when assessing whether a prediction from an ML model should be accepted. In this work, we have extended previous research on the development of confidence classifiers with reject option grounded on conformal prediction. Our method, based on the confidence measure, leverages the statistical guarantees of the framework. It enables any multi-class classifier to estimate with precise levels of confidence the expected accuracy or recall in a set of test samples. We evaluated our approach in six different multi-class datasets and compared it with off-the-shelf uncalibrated models and Platt-scaled classifiers. The findings demonstrate that our approach consistently delivers reliable error rate estimates at various rejection thresholds, outperforming other alternatives in most scenarios. In some cases, our approach was able to achieve performance guarantees, particularly when using the random forest as the underlying algorithm. Understanding the reasons behind the failure of the decision tree to achieve the desired guarantees in certain datasets is left for future research.

Future empirical experimentation may also involve the comparison of our proposal with other methods based on the calibration of the posterior scores, such as Dirichlet calibration (Kull et al., 2019) and Venn-Abers prediction (Vovk and Petej, 2014). The main limitation of our proposal lies in its application in low-data regime scenarios, when only a few calibration samples are available and statistical efficiency is sacrificed. An interesting extension in this case would be the construction of confidence classifiers with rejection based on cross-conformal predictors (Vovk, 2015). Finally, we would like to mention that, although this paper is focused on multi-class classification, our method could be easily adapted to tackle binary classification problems.

Acknowledgments

This work was supported by the Gobierno de Navarra through the ANDIA 2021 program (grant no. 0011-3947-2021-000023) and the ERA PerMed JTC2022 PORTRAIT project (grant no. 0011-2750-2022-000000).

References

Francisco Javier Acevedo, Saturnino Maldonado, Elena Dominguez, Arantzazu Narvaez, and Francisco Lopez. Probabilistic support vector machines for multi-class alcohol identification. *Sensors and Actuators B: Chemical*, 122(1):227–235, 2007.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 566–576. PMLR, 18–24 Jul 2021.

Henrik Boström. Calibrating random forests. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 121–126. IEEE, 2008.

Henrik Boström. crepes: a python package for generating conformal regressors and predictive systems. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 24–41. PMLR, 24–26 Aug 2022.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Leo Breiman, Jerome Friedman, Charles J Stone, and RA Olshen. *Classification and Regression Trees*. Routledge, 1984.

Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.

Alberto Garcia-Galindo, Marcos Lopez-De-Castro, and Ruben Armananzas. An uncertainty-aware sequential approach for predicting response to neoadjuvant therapy in breast cancer. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 74–88. PMLR, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

Ulf Johansson, Tuwe Löfström, and Henrik Boström. Calibrating multi-class models. In *Conformal and Probabilistic Prediction and Applications*, pages 111–130. PMLR, 2021.

Ulf Johansson, Tuwe Löfström, Cecilia Sönströd, and Helena Löfström. Conformal prediction for accuracy guarantees in classification with reject option. In Vicenç Torra and Yasuo Narukawa, editors, *Modeling Decisions for Artificial Intelligence*, pages 133–145, Cham, 2023a. Springer Nature Switzerland.

Ulf Johansson, Cecilia Sonstrod, Tuwe Lofstrom, and Henrik Bostrom. Confidence classifiers with guaranteed accuracy or precision. In Harris Papadopoulos, Khuong An Nguyen, Henrik Boström, and Lars Carlsson, editors, *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 513–533. PMLR, 2023b.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Classification with reject option using conformal prediction. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 94–105, Cham, 2018. Springer International Publishing.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning*, ECML '02, page 345–356, Berlin, Heidelberg, 2002. Springer-Verlag.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Machine learning*, 52:199–215, 2003.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019.

Craig Saunders, Alexander Gammerman, and Vladimir Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, volume 2, pages 722–726, 1999.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3):371–421, 2008.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2-3):349–376, 2013.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 829–838, Arlington, Virginia, USA, 2014. AUAI Press.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafe. *Algorithmic Learning in a Random World*. Springer-Cham, 1 edition, 2005.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.