Three-way Decision and Conformal Prediction: Isomorphisms, Differences and Theoretical Properties of Cautious Learning approaches

Andrea Campagner[a], Federico Cabitza[a], Pedro Berjano[b], Davide Ciucci[a]

[a]*Dipartimento di Informatica, Sistemistica e Comunicazione, University of Milano–Bicocca, viale Sarca 336 – 20126 Milano, Italy*
[b]*IRCCS Istituto Ortopedico Galeazzi Via Riccardo Galeazzi, 4 – 20161 Milano, Italy*

Abstract

The aim of this article is to study the relationship between two popular Cautious Learning approaches, namely: *Three-way decision* (TWD) and *conformal prediction* (CP). Based on the novel proposal of a technique to transform three-way decision classifiers into conformal predictors, and vice-versa, we provide conditions for the equivalence between TWD and CP. These theoretical results provide error-bound guarantees for TWD, together with a formal construction to define cost-sensitive cautious classifiers based on CP. The proposed techniques are then applied and evaluated on a collection of benchmark and real-world datasets. The results of the experiments show that the proposed techniques can be used to obtain cautious learning classifiers that are competitive with, and often out-perform, state-of-the-art approaches. Further, through a qualitative medical case study we discuss the usefulness of cautious learning in the development of robust Machine Learning.

Keywords: Three-way Decision, Cautious Learning, Conformal Prediction, Set-valued Prediction, Decision Support

1. Introduction

In this article, we study the problem of *Cautious Learning* [7]. This latter is a generalization of supervised learning in which the Machine Learning (ML) models are allowed to express set-valued predictions. The set-valued predictions allow the ML models to highlight a possible state of uncertainty, that should require further intervention from a human decision maker [5].

Recently, such techniques have been advocated as a promising approach [17] to develop reliable ML-based decision support in so-called *decision-critical domains*, e.g. medicine, social policing. Indeed, in all these settings, errors induced by ML models could have high-impact consequences. Therefore the decision makers could accept less precise, but more reliable predictions. Set-valued predictions could then be used by the decision-maker either to take a decision, if the

risk of doing so is not assumed to be too high; or to prompt the need to collect more information, so as to foster human-in-the-loop decision-making [14, 23].

Cautious learning methods clearly entail a trade-off between different quality dimensions, that should be properly evaluated so as take into account different desirable properties. These may include:

- *Cost-sentitiveness* [10]: that is, whether the model properly takes into account information about the utilities and costs of the different alternative decisions;
- *Validity* [35]: that is, whether the performance of the model can be reliably bounded, usually through a theoretical analysis;
- *Efficiency* [36]: that is, whether the set-valued predictions provided by the model are as *informative* as possible.

In recent years, many different cautious learning techniques have been proposed to strike a balance among these properties. These include models based on imprecise probabilities [41], or belief functions [27]; selective classification [12]; three-way decision [43] (TWD); and conformal prediction [35] (CP).

While all the mentioned models have been successfully employed in empirical settings, their theoretical characterization largely remains an open problem. First, there is a lack of works attempting to characterize the validity of cautious learning methods (with the exception of CP [35]); second, the relationships and similarities among different approaches have not yet been investigated.

In this work, we address these gaps by focusing on two popular approaches, namely *three-way decision* (TWD) and *conformal prediction* (CP):

- TWD, inspired by Rough Set theory and human decision making [43], is a generalization of decision-theory to the setting of set-valued predictions. Intuitively, given a new instance and a loss function, a TWD-based classifier would assign the instance to the set-valued prediction associated with minimal loss;
- CP, by contrast, is a technique to obtain calibrated confidence predictors. For each new given instance, a conformal predictor would return a nested collection of set-valued predictions, each with an associated error probability lower bound [35]. A cautious learning algorithm can then be defined from a conformal predictor by selecting a specific probability threshold.

These differences notwithstanding, the two methods also share some similarities [6]. Indeed, both methods can be applied as a post-processing step to any standard (i.e., non-cautious) learning method [2, 5]; both methods are distribution-free; and both methods make relatively weak assumptions. See Figure 1 for a graphical representation of TWD-based classifiers and CP, in comparison with a standard (i.e., single-valued) ML model.

The aim of this paper, then, is to study the relationships among these two models, and to characterize when these two different approaches can be considered equivalent. To this purpose, we first define techniques to transform a

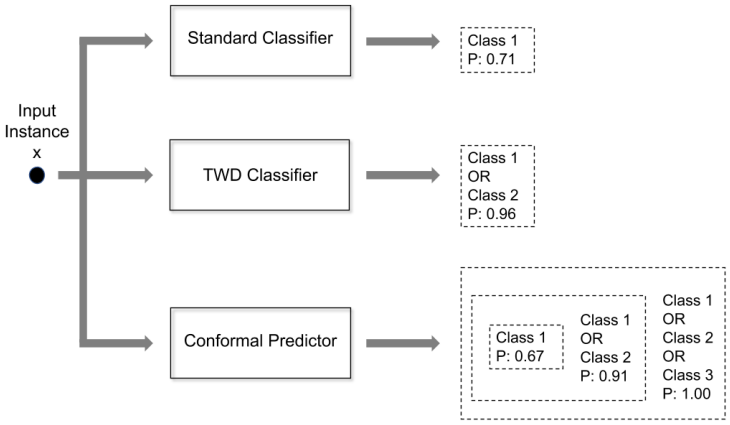

Figure 1: A graphical representation of standard (classification) ML models, TWD models and CP models. Given an input instance x, a standard classifier would provide as output either a single class (in the example, Class 1) together with a confidence score, or a confidence score distribution. By contrast, a TWD classifier would provide as output the set of labels (in the example, Class 1 or Class 2) which is optimal w.r.t. a specific loss function, possibly together with an aggregated confidence score. Finally, a conformal predictor provides as output a nested collection of sets of labels, each with an associated (lower) probability bound.

TWD-based ML model into a CP one, and vice-versa. Harnessing this relationship, we investigate two main theoretical questions:

- Under which conditions a TWD-based model is guaranteed to be valid, and with which error bounds? We answer this question through Theorems 2 and 3, by which it is shown that, under very general assumptions, TWD classifiers are valid;
- Under which conditions TWD and CP methods are equivalent? We answer this question through Theorems 4 and 5, by which conditions for the equivalence between TWD and CP methods are provided.

Moreover, by means of a set of quantitative experiments, we show that, when the above mentioned assumptions do not hold, the proposed techniques can be used to improve the validity of TWD classifiers.

The rest of this article is structured as follows. In Section 2, the necessary technical background on Machine Learning, TWD in the ML domain, and CP is provided. In Section 3, we study the relationship between TWD and CP. Specifically, in Section 3.1, TWD is used as a basis to define a CP. Through this construction, the validity of TWD-based ML models is formally established. Conversely, in Section 3.2 we discuss how TWD can be used to define cost-sensitive cautious learning methods based on CP algorithms. Through these

constructions conditions for the equivalence between TWD and CP are established. In Section 4, the empirical performance of the proposed constructions is investigated, through a set of experiments on real-world datasets. The results of these experiments are then discussed in Sections 4.2 and 4.3; while in Section 5 a short medical case study is discussed. Finally, in Section 6, the obtained results are summarized, and possible future lines of research are outlined.

2. Background

In this section, we recall the necessary background about ML, TWD and CP.

2.1. Supervised Machine Learning

Let X be the input space, i.e., a set of objects described as vectors of feature values. Let Y be the target space, i.e., the set of classes. Then, a classification algorithm, w.r.t. a sample space Z and a hypothesis space $\mathcal{H}$, is a function $\mathcal{A} : 2^Z \mapsto \mathcal{H}$. When $Z = X \times Y$, $\mathcal{A}$ is denoted as a *supervised* classification algorithm; by contrast, when $Z = X \times 2^Y$, $\mathcal{A}$ is a *weakly-supervised* [15] classification algorithm.

Let $S \subseteq Z$ be a sample drawn i.i.d. from an unknown distribution $\mathcal{D}$; $\mathcal{H}$ be an hypothesis space; and $l : \mathcal{H} \times Z \mapsto \mathbb{R}^+$ be a loss function. Then, the goal of the *machine learning problem* is to find a hypothesis $h \in \mathcal{H}$ with minimal (or small) *true risk*:

$$Risk_{\mathcal{D}}(h, l) = \int_{z \in Z} l(h, z) d\mathcal{D}(z). \tag{1}$$

Since $\mathcal{D}$ is unknown, the true risk cannot be computed. Hence, the aim is to minimize a proxy of the true risk, such as the *empirical risk*, based on the finite sample S:

$$Risk_S(h, l) = \frac{1}{|S|} \sum_{z \in S} l(h, z). \tag{2}$$

Empirical Risk Minimization (ERM) is the algorithm that, given $\mathcal{H}$ and a training set S, selects one of the $h \in H$ s.t. $Risk_S(h, l) = \min_{h' \in \mathcal{H}} Risk(h', l)$. We denote any such h as h_S. The ERM learning paradigm has been generalized to the setting of weakly supervised learning [15, 16] by means of *generalized loss functions*. These latter are usually expressed in the form $l^S(h, \langle x, Y_x \rangle) = A(\{l(h, \langle x, y \rangle) : y \in Y_x\})$; where $A \in \{\min, \max, \text{mean}\}$.

In the following, we assume that $\mathcal{H}$ is a class of scoring classifiers. A scoring classifier is a function $h : X \mapsto Y$ s.t. $h = dec \circ s$, where $s : X \mapsto \mathbb{R}^{|Y|}$ is a scoring function (mapping an instance $x \in X$ to a distribution of scores); and $dec : \mathbb{R}^{|Y|} \mapsto Y$ is a decision function (mapping a distribution $s(x)$ to a single label). The decision function dec is usually defined as $dec(s(x)) = argmax_{y \in Y} s(x)_y$, where $s(x)_y$ denotes the score assigned to label y.

A *cautious* classifier [11] is a function $h : X \mapsto 2^Y$. Thus, a cautious classifier maps instances to *sets* of labels. The semantics attached to set prediction $h(x) \subseteq Y$ is that the correct label $\hat{y}$ is *likely* to be in $h(x)$.

2.2. Three-way Decision

Three-way decision (TWD) [43] is a framework for information and uncertainty management, inspired by human decision-making and rough set theory [42], that generalizes standard decision theory.

In the binary setting, one considers three regions: a positive, or acceptance, region; a negative, or rejection, region; and a boundary, or non-commitment, region. This latter region, in particular, represents lack of knowledge, or (temporary) abstention, in regards to the status of the objects it contains.

With respect to the ML setting [42], according to TWD, every instance can be classified as either belonging to a given class (and thus not belonging to all others); not belonging to a given class; or being in the *boundary*, that is a region that represents lack of knowledge with respect to class assignment. This latter property makes TWD useful for the development of cautious classifiers, by means of a theoretically sound and cost-sensitive approach [6].

Indeed, this approach has been successfully applied in the ML literature for many tasks. Li et al. [22] proposed a cautious classification model for binary classification based on modeling uncertain boundaries; while Xu et al. [39] proposed a generalization of TWD-based cautious classification to multi-class problems, using sequential TWD. Liu et al. [24] proposed a TWD method based on the combination of logistic regression and decision-theoretic rough sets; Zhang et al. [49] proposed an approach for cost-sensitive cautious classification based on TWD and ensemble learning; similarly, Yue et al. [47] and Savchenko [32] proposed computationally efficient techniques for cautious classification based on TWD and deep learning. Min et al. [29] proposed an approach for cautious classification of weakly-supervised data using TWD and active learning; Campagner et al. [5] proposed an approach for weakly supervised learning and multi-class cautious classification based on TWD and statistical learning methods; Gu et al. [13] studied approaches to TWD in group-decision making based on imprecise probabilistic linguistic assessments; Zhou et al. [50] studied and compared different approaches for TWD based on coarse and fuzzy data. More recently, Liu et al. [23] also discussed the interpretability and usefulness of cautious classification methods based on TWD. For a more general discussion about TWD in ML, we refer the reader to the recent surveys by Campagner et al. [6] and Liu et al. [25]. Furthermore, approaches for cautious classification based on TWD have recently been investigated also from a theoretical and conceptual perspective: Liu et al. [26] studied an an alternative model for TWD based on optimization; Yao [46] studied the connections between TWD and set-based approaches; Yao [45] explored the foundations of TWD based on geometrical and numerical concepts; while Xu [38] studied the connections between TWD-based classification and the theory of confusion matrices. We refer the reader to the reviews by Yang et al. [40] and Yao [44] for further details.

In TWD, the loss function is generalized as a set-valued function $l : 2^Y \times Y \mapsto \mathbb{R}$, so as to model the loss w.r.t. a set-valued prediction. In this article, we consider the multi-class formulation of TWD classification [5]. In this latter approach, the loss function l can be decomposed in two parts, namely an *error cost function* and an *abstention cost function*. Formally, let

- $err : 2^Y \times Y \mapsto \mathbb{R}$ be an *error cost function*. Intuitively, $err(S, y)$ represents the cost of predicting S, when $y \notin S$ is the correct label;
- $\alpha : \mathbb{N}^+ \mapsto \mathbb{R}$ be an *abstention cost function*. We assume that

$$\forall i > 1, \alpha(1) = 0 < \alpha(i) \leq \alpha(i+1) \leq \min_{A \in 2^Y, y \in Y} err(A, y). \tag{3}$$

  Inuitively, $\alpha(|A|)$ represents the cost of making a set-valued prediction A that contains the correct label y.

Let h be a scoring classifier, its generalized loss is defined as :

$$Loss_{TWD}(A) = \sum_{y \notin A} h(x)_y \cdot err(A, y) + \alpha(|A|) \sum_{y \in A} h(x)_y. \tag{4}$$

Then, the TWD classifier $\mathcal{W}_h$ is defined, for each $x \in X$, as:

$$\begin{aligned} \mathcal{W}_h(x) = \arg \min_{A \in 2^Y} \{|A| : \\ A \in \arg \min_{B \in 2^Y} Loss_{TWD}(B)\}. \end{aligned} \tag{5}$$

Hence, for each x, the result of $\mathcal{W}_h(x)$ is (one of) the smallest sets having minimal generalized loss. In Example 1 we briefly describe the calculations involved in the definition of a simple TWD classifier.

Example 1. *Let err be the constant* 1 *function, and* $\alpha(|A|) = \frac{|A|-1}{|Y|}$*, with* $Y = \{1, 2, 3, 4, 5\}$.
Let h be a scoring classifier, and x an instance such that

$$h(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle.$$

Since the error cost function err is uniform, the optimization problem in Eq. (5) *can be solved using a greedy algorithm [5]. Thus, the following holds:*

$$\begin{aligned} Loss_{TWD}(\{2\}) &= 0.7 \\ Loss_{TWD}(\{2, 5\}) &= 0.56 \\ Loss_{TWD}(\{1, 2, 5\}) &= 0.55 \\ LossTWD(\{1, 2, 3, 5\}) &= 0.64 \\ Loss_{TWD}(Y) &= 0.8 \end{aligned}$$

Therefore, $\mathcal{W}_h(x) = \{1, 2, 5\}$.

By definition, the TWD classifier $\mathcal{W}_h(x)$ is the cautious classifier with minimal risk, *under the assumption* that the probability scores returned by h approximate the probability of error (i.e. h is calibrated). However, the calibration of h is, in general, only a sufficient condition for the correctness of the set-valued predictions issued by the TWD classifier $\mathcal{W}_h(x)$. In Section 3, based on the relationship between TWD and CP, we study some conditions under which TWD classifiers are guaranteed to be valid.

2.3. Conformal Prediction

Conformal Prediction [35] (CP) is a cautious learning approach that allows to define calibrated classifiers. Since its introduction, the CP framework has been adapted to different settings, including clustering [30], anomaly detection [21], active learning [28], semi-supervised learning [1]. Furthermore, CP has been successfully applied in many empirical settings, including cancer detection [48], cybersecurity [37], drug discovery [4]. See [2] for a recent review on CP.

Conventionally, CP is applied in the *transductive* learning setting [34]. In this latter setting, the instances are assumed to be sampled sequentially. Nonetheless, conformal predictors can be applied also to the standard *inductive* learning paradigm, by using a separate *validation* (or, *calibration*) set [35]. For simplicity of presentation, here and in Section 3, we focus on the transductive setting.

A *non-conformity measure* is a permutation-invariant function $M : 2^{X\times Y} \times (X \times Y) \mapsto \mathbb{R}$, i.e., given $S = (\langle x_1, y_1\rangle ..., \langle x_n, y_n\rangle)$, it holds that $M(S, \langle x, y\rangle) = M(\pi(S), \langle x, y\rangle)$ for every permutation π. Intuitively, a non-conformity measure quantifies how much a new instance $\langle x, y\rangle$ differs from past examples in S. More formally, the value of a non-conformity measure, for a given instance $\langle x, y\rangle$, represents a *statistic* for a non-parametric testing procedure [35].

Let $S_{x_i,x}$ be the result of exchanging $\langle x_i, y_i\rangle$ with $\langle x, y\rangle$ in S. Then, the *conformal predictor* determined by M is a function $\Gamma_M : 2^{X\times Y} \times X \times [0, 1] \mapsto 2^Y$, defined as:

$$\Gamma^{\epsilon}_M(S, x) = \{y | p^{x,y} > \epsilon\}, \tag{6}$$

where $\epsilon \in [0, 1]$ and $p^{x,y}$ is defined as:

$$p^{x,y} = \frac{|\{i = [1, n] : M(S, \langle x, y\rangle) \leq M(S_{x_i,x}, \langle x_i, y_i\rangle)\}| + 1}{n + 1}. \tag{7}$$

Intuitively speaking, relying on the above mentioned interpretation of the non-conformity measure as a testing statistic, the value $p^{x,y}$ is the *p-value* for the null hypothesis that the instance $\langle x, y\rangle$ comes from the same distribution as S [2]. Therefore, the labels in $\Gamma^{\epsilon}_M(S, x)$ are those for which the previously mentioned null hypothesis cannot be rejected (at a threshold confidence value of ϵ).
We denote with $im(\Gamma_M)$ the image of Γ_M, that is:

$$im(\Gamma_M) = \{A \subseteq Y : \exists \epsilon \in [0, 1] \text{ s.t. } \Gamma^{\epsilon}_M(x) = A\}. \tag{8}$$

Thus, $im(\Gamma_M)$ is a nested collection of sets $A_1 = \emptyset \subseteq ...A_i \subseteq A_n = Y$. Each set $A_i \in im(\Gamma_M)$ has an associated ϵ_i s.t. $\epsilon_1 = 1, \epsilon_n = 0$. The map $p(i) = \epsilon_i$ represents the p-value function [2] of the statistical procedure defined by Γ_M.

Notably, a cautious classifier Γ^{ϵ}_M can be constructed from a conformal predictor Γ_M, by selecting an appropriate ϵ.

In Example 2, we illustrate the computations involved in the definition of a conformal predictor, by using an approach based on 1-nearest neighbor [35].

Example 2. *In this example, the non-conformity measure will be defined as:*

$$M_{1NN}(S, \langle x, y\rangle) = \frac{\min_{x' \in S: y_{x'}=y} d(x, x')}{\min_{x' \in S: y_{x'}\neq y} d(x, x')}, \tag{9}$$

where d is a metric. Thus, the similarity of a new example $\langle x, y\rangle$ w.r.t. the training set S is high when x is more similar to the instances in S associated with the same label, than to instances associated with a different label.

Consider the following single-feature training set

$$S = \{i_1 = \langle 0.75, 0\rangle, i_2 = \langle 0.90, 0\rangle, i_3 = \langle 0.48, 1\rangle\},$$

Let $x = 0.615$ be a new instance to be classified. Then $M_{1NN}(S, \langle x, 0\rangle) = M_{1NN}(S, \langle x, 1\rangle) = 1$ and, similarly:

$$\begin{aligned}
&M_{1NN}(S_{i_1,\langle x,0\rangle}, i_1) = 0.5\\
&M_{1NN}(S_{i_3,\langle x,1\rangle}, i_3) = 0.5\\
&M_{1NN}(S_{i_1,\langle x,1\rangle}, i_1) = 1.15\\
&M_{1NN}(S_{i_2,\langle x,0\rangle}, i_2) = 0.36\\
&M_{1NN}(S_{i_2,\langle x,1\rangle}, i_3) = 0.53.
\end{aligned}$$

By contrast, $M_{1NN}(S_{i_3,\langle x,0\rangle}, i_3)$ is undefined, as there is no instance with label 1 in the associated training set. Thus, we set $M_{1NN}(S_{i_3,\langle x,0\rangle}, i_3) = +\infty$. Therefore, $p^{x,0} = p^{x,1} = \frac{1}{2}$ and the corresponding conformal predictor is defined as:

$$\Gamma^{\epsilon}_{M_{1NN}} = \begin{cases} \emptyset & \epsilon > \frac{1}{2} \\ \{0, 1\} & otherwise \end{cases}$$

As previously mentioned, the main advantage of CP, compared to other cautious classification approaches, is that every conformal predictor is *valid*, i.e. the following result holds:

Theorem 1 (Vovk et al. [35])**.** *Let S, x be sampled i.i.d. from the same distribution $\mathcal{D}$, y be the true (but unknown) label associated with x. Let M be a non-conformity measure and $\epsilon \in [0, 1]$. Then, taken Γ_M the conformal predictor based on M, it holds that Γ_M is* conservatively valid*, that is:*

$$Pr[y \notin \Gamma^{\epsilon}_M(S, x)] \leq \epsilon. \tag{10}$$

Thus, the probability of error of $\Gamma^{\epsilon}_M(S, x)$ is no greater than ϵ. Numerous approaches have been proposed in the literature to define conformal predictors, both based on algorithm-specific approaches [33]; and general-purpose ones [18]. One of the most popular general-purpose methods [18] is based on a score-based classifier h (see Section 2.1). In this case, a non-conformity measure based on h can be defined as:

$$M_h(S, \langle x, y\rangle) = \max_{y' \in Y}\{s(x)_{y'}\} - s(x)_y. \tag{11}$$

3. Methods

In this section, we study the relationships between TWD and CP. The main contents of this section, as well as the results of our study, are summarised in Figure 2. As previously mentioned, in the following we focus on the transductive

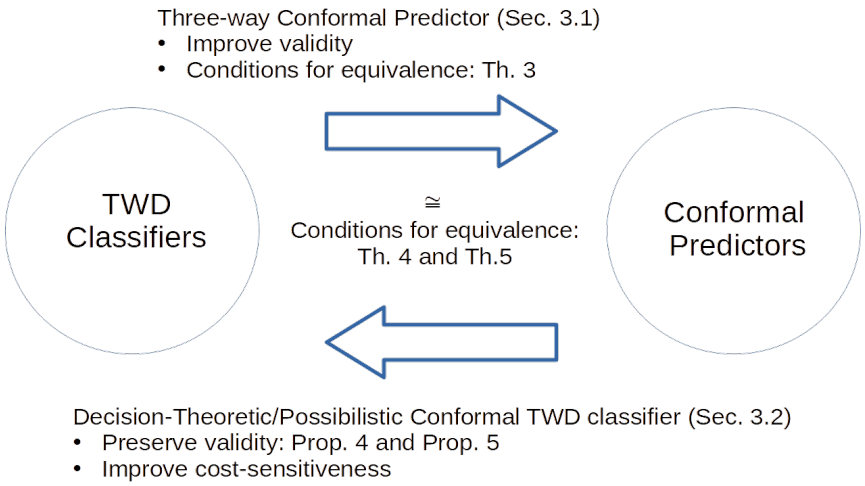

Figure 2: A graphical illustration of the main results in Section 3.

setting. As highlighted in Section 2.3, note, however, that the properties of conformal predictors we study in this section hold also in the inductive setting.

3.1. From Three-way Decision to Conformal Prediction

In this section, we address the first of the research questions mentioned in Section 1. Namely, we study whether, and under which conditions, TWD classifiers are valid. To this purpose, we first show that TWD can be used to design conformal predictors. Then, we provide sufficient and necessary conditions for the validity of TWD classifiers. The connection between TWD and CP is then generalized to the setting of weakly supervised learning.

Let us first consider the standard supervised setting (i.e. $Z = X \times Y$). Let $S = (\langle x_1, y_1 \rangle, ..., \langle x_n, y_n \rangle)$ be the training set. The *three-way non-conformity measure*, based on a given TWD classifier $\mathcal{W}$, can be defined as:

$$M_{\mathcal{W}}(S, \langle x, y \rangle) := l(\mathcal{W}_S, \langle x, y \rangle) = \\ = \begin{cases} \alpha(|\mathcal{W}_S(x)|) & y \in \mathcal{W}_S(x) \\ err(\mathcal{W}_S(x), y) + \alpha(|\mathcal{W}_S(x)|) & otherwise \end{cases} \tag{12}$$

where $l(\mathcal{W}_S, \langle x, y \rangle)$ denotes the loss of the prediction $\mathcal{W}_S(x)$, given $y \in Y$.

Thus, the *three-way non-conformity measure* assigns, to each instance $\langle x, y\rangle$, the loss incurred by using $\mathcal{W}_S$ to predict the label of x. It is easy to observe that, for any $\mathcal{W}$, $M_{\mathcal{W}}$ is indeed a non-conformity measure:

Proposition 1. *Let $\mathcal{W}$ be a three-way classifier, then $M_{\mathcal{W}}$, defined as in Eq. (12), is a non-conformity measure.*

Proof. Let S be a sample, and $\mathcal{W}_S$ the three-way classifier defined by S. Then, by definition, for any permutation S_1 of S, it holds that $\mathcal{W}_S = \mathcal{W}_{S_1}$. Therefore, the training algorithm is permutation-invariant. □

The construction described in Section 2.3, applied to the three-way non-conformity measure $M_{\mathcal{W}}$, allows to define the *three-way conformal predictor* (TWCP) as:

$$\Gamma^{\epsilon}_{\mathcal{W}}(S, x) = \{y | p^{x,y} > \epsilon\}, \tag{13}$$

$$p^{x,y} = \frac{|\{i = 1, ..., n : \text{Pred is verified}\}| + 1}{n + 1}, \tag{14}$$

$$Pred := l(\mathcal{W}_S, \langle x, y\rangle) \leq l(\mathcal{W}_{S_{x_i,x}}, \langle x_i, y_I\rangle). \tag{15}$$

The calculations involved in the definition of the TWCP are briefly illustrated in Example 3.

Example 3. *Let $Y = \{0, 1, 2\}$, and let S be a training set s.t. $S = \{i_1 = \langle x_1, 0\rangle, i_2 = \langle x_2, 1\rangle, i_3 = \langle x_3, 1\rangle, i_4 = \langle x_4, 2\rangle\}$. Let $\mathcal{W}$ be a TW classifier s.t. $\mathcal{W}_S(x_1) = \{0, 2\}$, $\mathcal{W}_S(x_2) = \{0\}$, $\mathcal{W}_S(x_3) = \{1\}$ and $\mathcal{W}_S(x_4) = \{1, 2\}$.*

Let x be a new instance. Assume, for simplicity, that $\forall i_j$, $\mathcal{W}_S = \mathcal{W}_{S_{i_j,x}}$ and that $\mathcal{W}_S(x) = \{0, 1\}$. Let $err = 1$ and $\alpha(|A|) = \frac{|A|-1}{|Y|-1}$.

Then $M_{\mathcal{W}}(S, \langle x, 1\rangle) = 0.5$, $M_{\mathcal{W}}(S, \langle x, 0\rangle) = 0.5$, $M_{\mathcal{W}}(S, \langle x, 2\rangle) = 1.5$, while

$$M_{\mathcal{W}}(S_{i_1,x}, i_1) = 0.5$$
$$M_{\mathcal{W}}(S_{i_4,x}, i_4) = 0.5$$
$$M_{\mathcal{W}}(S_{i_2,x}, i_2) = 1$$
$$M_{\mathcal{W}}(S_{i_3,x}, i_3) = 0$$

.

Therefore $p^{x,0} = p^{x,1} = \frac{4}{5}$, $p^{x,2} = \frac{1}{5}$ and the TWCP $\Gamma_{\mathcal{W}}$ is defined as:

$$\Gamma^{\epsilon}_{\mathcal{W}} = \begin{cases} \emptyset & \epsilon > \frac{4}{5} \\ \{0, 1\} & \frac{1}{5} < \epsilon \leq \frac{4}{5} \\ Y & otherwise \end{cases}$$

Since $M_{\mathcal{W}}$ is a non-conformity measure, as a consequence of Theorem 1, it holds that the TWCP $\Gamma_{\mathcal{W}}$ is conservatively valid:

Corollary 1. *Let S, x be sampled i.i.d. from the distribution, and let $\hat{y}$ be the correct label associated with x. Then, for any ϵ, $Pr[\hat{y} \notin \Gamma^{\epsilon}_{\mathcal{W}}(S, x)] \leq \epsilon$, that is $\Gamma_{\mathcal{W}}$ is conservatively valid.*

Proof. The result follows directly from Theorem 1 and the observation (see Prop. 1) that $M_{\mathcal{W}}$ is a non-conformity measure. □

The previous result holds for any CP algorithm, thus, in particular, for the TWCP. Nonetheless, the previous result does not provide any information about the validity of the original TWD classifier. Then, we ask two main questions: can the validity of a TWCP be used to obtain performance bounds for the corresponding TWD classifier? Under which conditions it holds that a TWD classifier and the corresponding TWCP are equivalent?

In regard to the first question, note that the transformation from a TWD classifier $\mathcal{W}$ to the corresponding TWCP $\Gamma_{\mathcal{W}}$ provides a bound on the probability of error of $\mathcal{W}$. Indeed, if $\mathcal{W}_S(x) \in im(\Gamma_{\mathcal{W}}(S,x))$, then, the following bound follows from Corollary 1:

$$Pr[y \notin \mathcal{W}_S(x)] \leq \arg\min_{\epsilon \in [0,1]} \{\Gamma^{\epsilon}_{\mathcal{W}}(S,x) = \mathcal{W}_S(x)\}. \tag{16}$$

Consequently, a sufficient condition for the validity of the TWD classifier $\mathcal{W}$ would be that $\mathcal{W}_S(x) \in im(\Gamma_{\mathcal{W}}(S,x))$. Then, the next result provides a characterization of this property:

Theorem 2. *The following two conditions are equivalent:*

1. $\mathcal{W}_S(x) \in im(\Gamma_{\mathcal{W}}(S,x))$*;*
2. $\exists \langle x_i, y_i \rangle \in S$ *such that*

$$l(\mathcal{W}_{S_{x_i,x}}, \langle x_i, y_i \rangle) < \min_{y \notin \mathcal{W}_S} err(\mathcal{W}_S, y).$$

Proof. First, we prove that 1 implies 2. Note that $\forall y \in \mathcal{W}_S(x) = A$, then either $l(A,y) = 0$ (when $A = \{y\}$) or $l(A,y) = \alpha(|A|)$. Furthermore, by definition of α and err, it holds that $\forall A, \alpha(|A|) \leq \min_{B \subseteq Y, y \notin B} err(B,y)$. Thus, if 2 does not hold, then it exists $y \notin \mathcal{W}_S(x)$ s.t. $p^{x,y} = 1$. Consequently, the smallest A_i in $im(\Gamma_{\mathcal{W}}(S,x))$ is s.t. $\mathcal{W}_S \cup \{y\} \subseteq A_i$. The proof for the converse implication is analogous. □

Thus, as a consequence of Theorem 2, every non-trivial TWD classifier[1] is valid, and can be associated with an error upper bound. This latter error bound quantifies the probability that the correct label is not contained in the set-valued prediction issued by the TWD classifier.

Furthermore, this latter error bound is dependent on the predictive performance of the TWD classifier. This dependency is formalized through the following Theorem, which provides a characterization of the nested set structure for any TWCP:

Theorem 3. *Let* $\epsilon \in [0,1]$ *and let* $\mathcal{W}_S(x) = A$*. Then* $A = \Gamma^{\epsilon}_{\mathcal{W}}(S,x)$ *iff both:*

[1] Here, non-trivial refers to any TWD classifier that does not err on all of its predictions.

1. *$\mathcal{W}$ makes at least $\lfloor \epsilon \cdot (n+1) \rfloor$ predictions on S with risk greater than $\alpha(|A|)$;*
2. *$\mathcal{W}$ makes at most $\lceil \epsilon \cdot (n+1) \rceil$ predictions on S with risk greater than $\min_{y \notin A} err(A, y)$.*

Proof. First, note that $\forall y \in A, l(A, y) = \alpha(|A|)$. Thus, if $y \in A$ is in $\Gamma^{\epsilon}_{\mathcal{W}}(S, x)$, then the same holds for all $y' \in A$. Thus, a sufficient (and necessary) condition for $y \in A$ to be included in $\Gamma^{\epsilon}_{\mathcal{W}}(S, x)$ is the existence of at least $\lfloor \epsilon \cdot (n+1) \rfloor$ instances $x' \in S$ s.t. $l(\mathcal{W}_{S_{x',x}}(x'), y') \geq \alpha(|A|)$. Otherwise $\Gamma^{\epsilon}_{\mathcal{W}}(S, x) = \emptyset$.

As for the second condition, note that for any $y \notin A$

$$l(\mathcal{W}_S, \langle x, y \rangle) \geq \min_{y' \notin A} l(\mathcal{W}_S, \langle x, y' \rangle) > \alpha(|A|).$$

Thus a sufficient and necessary condition for excluding any $y \notin A$ from $\Gamma^{\epsilon}_{\mathcal{W}}(S, x)$ is that for at most $\lfloor \epsilon \cdot (n+1) \rfloor$ instances $\langle x', y' \rangle \in S$, it holds that

$$l(\mathcal{W}_{S_{x',x}}, \langle x', y' \rangle) \geq \min_{y \notin A} l(\mathcal{W}_S, \langle x, y \rangle).$$

Thus, the theorem follows. □

Finally, with respect to our second question, we note that in the uniform-cost classification setting, a finer version of Theorem 3 can be derived. This result shows that any TWD classifier and its corresponding TWCP are equivalent (see also Example 4 for a brief illustration of the following Theorem):

Corollary 2. *Let $\epsilon \in [0, 1]$, then in the uniform-cost classification setting it holds that:*

- *If $|\mathcal{W}_S(x)| = 1$, then $\mathcal{W}_S(x) = \Gamma^{\epsilon}_{\mathcal{W}}(S, x)$ iff $\mathcal{W}$ makes at most $\lfloor \epsilon \cdot (n+1) \rfloor$ errors on S (otherwise, $\Gamma^{\epsilon}_{\mathcal{W}}(S, x) = Y$);*
- *Otherwise, $\mathcal{W}_S(x) = \Gamma^{\epsilon}_{\mathcal{W}}(S, x)$ iff $\mathcal{W}$ makes at most $\lceil (1-\epsilon) \cdot (n+1) \rceil$ predictions on S with risk lower than $\alpha(|\mathcal{W}_S(x)|)$ (otherwise, $\Gamma^{\epsilon}_{\mathcal{W}}(S, x) = \emptyset$) and at most $\lfloor \epsilon \cdot (n+1) \rfloor$ errors (otherwise, $\Gamma^{\epsilon}_{\mathcal{W}}(S, x) = Y$).*

Thus, $\Gamma_{\mathcal{W}}(S, x)$ is completely determined by two thresholds $0 \leq \epsilon_1 < \epsilon_2 \leq 1$ s.t.

$$\Gamma^{\epsilon}_{\mathcal{W}}(S, x) = \begin{cases} \emptyset & \epsilon_2 < \epsilon \\ \mathcal{W}_S(x) & \epsilon_1 < \epsilon \leq \epsilon_2 \\ Y & 0 \leq \epsilon \leq \epsilon_1 \end{cases} \tag{17}$$

Proof. The result follows directly from Theorem 3, applying the result to the case of uniform-cost classification. □

Example 4. *Consider the TWCP introduced in Example 3. Then, $|\mathcal{W}_S(x)| > 1$, and $\mathcal{W}$ makes exactly one prediction with risk lower than $\alpha(\mathcal{W}_S(x)) = 0.5$. Hence, by Theorem 2 it holds that $\epsilon_2 = \frac{4}{5}$. Similarly, $\mathcal{W}$ makes exactly 1 error, hence by Theorem 2 it holds that $\epsilon_1 = \frac{1}{5}$.*

We now focus on the more general weakly supervised learning setting (i.e. $Z = X \times 2^Y$). Let $S = (\langle x_1, Y_1\rangle, ..., \langle x_n, Y_n\rangle)$ be the current training set. The three-way nonconformity measure can be generalized as follows:

$$M_{\mathcal{W}}^{\min}(S, \langle x, Y_x\rangle) = \min_{y \in Y_x} l(\mathcal{W}_S, \langle x, y\rangle). \tag{18}$$

where $\mathcal{W}$ is a three-way in/three-way out classifier [5]. Thus, the *superset* TWCP is defined as:

$$\Gamma^{\epsilon}_{\mathcal{W},\min}(S, x) = \{y | p^{x,y} > \epsilon\}, \tag{19}$$

$$p^{x,y} = \frac{|\{i = 1, ..., n : \text{Pred is verified}\}| + 1}{n + 1}, \tag{20}$$

$$Pred = l(\mathcal{W}_S, \langle x, y\rangle) \leq \min_{y' \in Y_i} l(\mathcal{W}_{S_{x_i,x}}, \langle x_i, y'\rangle). \tag{21}$$

The nonconformity measure $M_{\mathcal{W}}^{\min}$ is defined in terms of the *minimum* operator. Thus, it is similar to the optimistic loss minimization [15] approach for weakly supervised learning. Remarkably, however, the role of the minimum operator in the two formulations is different. In the generalized loss minimization framework, the minimum operator selects the instantiation of the set labels that minimizes the empirical loss, over all possible instantiations. On the other hand, in Eq. (19), the minimum operator acts as a conservative bound for the similarity between x and the training set S. Indeed, given $\langle x_i, Y_i\rangle$, the corresponding nonconformity score is

$$\min_{y \in Y_i} M_{\mathcal{W}}(S, \langle x_i, y\rangle) \leq M \leq \max_{y \in Y_i} M_{\mathcal{W}}(S, \langle x_i, y\rangle).$$

Thus, the nonconformity score of x is compared against the most conservative threshold, among those that are considered possible. In this sense, Eq. (19) is more similar to the principle underlying pessimistic loss minimization [16].

As a second remark, we study the efficiency of the superset TWCP. It is not hard to see that changing min, in Eq. (19), with max or *mean* would equally result in a non-conformity measure. However, it is easy to observe that the approach based on the minimum operator is more efficient than those based on, either, the maximum or mean operators. Indeed, denote these latter non-conformity measures as, resp., $M_{\mathcal{W}}^{\max}, M_{\mathcal{W}}^{mean}$. Similarly, denote the corresponding conformal predictors as, resp., $\Gamma_{\mathcal{W},\max}, \Gamma_{\mathcal{W},mean}$. Then, the following result holds:

Proposition 2. *Let $\mathcal{W}$ be a TWD classifier, S a training set and x a new instance. Then, for any $\epsilon \in [0, 1]$, $\Gamma^{\epsilon}_{\mathcal{W},\min}(x) \subseteq \Gamma^{\epsilon}_{\mathcal{W},mean}(x) \subseteq \Gamma^{\epsilon}_{\mathcal{W},\max}(x)$.*

Proof. Note that, for any set of positive numbers $\{n_1, ..., n_m\}$, it holds that $\arg\min_i\{n_i\} \leq \frac{1}{m}\sum n_i \leq \arg\max_i\{n_i\}$. Then, the result easily follows. □

3.2. From Conformal Prediction to Three-way Decision

In this section we address the second of the research questions mentioned in Section 1. Namely, we study conditions under which TWD and CP methods are equivalent. To this aim, we first outline two approaches to define a cost-sensitive cautious classifier from any conformal predictor. These latter approaches can be used to transform any conformal predictor into a TWD classifier. We then study the equivalence between TWD and CP methods, by applying the above mentioned approaches to the case in which the conformal predictor is defined as in Section 3.1.

The first approach to obtain a TWD classifier, starting from a CP algorithm Γ_M, relies on the observation that Γ_M is defined as a collection of nested sets, associated with corresponding (lower) probabilities.

Let M be any non-conformity measure, and let Γ_M be the corresponding conformal predictor. Let $A \subseteq Y$ be s.t. $A \in im(\Gamma_M(x))$, i.e. $\exists \epsilon \in [0,1]$ s.t. $\Gamma^{\epsilon}_M(x) = A$. Denote with ϵ^A, the (unique) solution of the following equality:

$$\epsilon^A := \arg \min_{\epsilon \in [0,1]} \{\Gamma^{\epsilon}_M(x) = A\}. \tag{22}$$

Eq. (22) implies that, given $A \in im(\Gamma_M(x))$, it is known that $Pr_{\langle x,y\rangle \sim \mathcal{D}}[y \notin A] \leq \epsilon_A$. Therefore, the loss $Loss_{\Gamma_M}(x)$ w.r.t. A can be bounded as follows:

Proposition 3. *Let $A \subseteq Y$ be in the image of $\Gamma_M(x)$, and let ϵ^A be the corresponding solution of Eq.* (22)*. Then:*

$$\begin{aligned} \alpha(|A|) \leq Loss_{\Gamma_M(x)}(A) \leq \alpha(|A|) \cdot (1 - \epsilon^A) + \\ + \epsilon^A |Y \setminus A| \cdot \max_{y \notin A} \{err(A, y)\}. \end{aligned} \tag{23}$$

Proof. Let $y \in Y$ be the real label attached to x. Then, by Theorem 1, $Pr[y \notin A] \leq \epsilon^A$. Further, by definition of err and α, it holds that

$$\alpha(|A|) \leq \min_{y \notin A} err(A, y) \leq \max_{y \notin A} \{err(A, y)\}.$$

Note, also, that the rightmost summand in Eq. (23) is monotonically increasing w.r.t. $\epsilon \in [0, \epsilon^A]$, and the left and right side of the inequality chain coincide when $\epsilon = 0$. Then, the result easily follows. □

Denote the right-most side of Eq. (23) as $Loss^*_{\Gamma_M(x)}(A)$. Then, given a conformal predictor Γ^{ϵ}_M, the *decision-theoretic conformal TWD* (DCTWD) classifier $\mathcal{W}_\Gamma$ is defined as follows:

$$\mathcal{W}^{dec}_\Gamma(x) = \arg \min_{A \in im(\Gamma_M(x))} Loss^*_{\Gamma_M(x)}(A). \tag{24}$$

On the other hand, the second approach to transform a conformal predictor into a TWD classifier relies on the observation that a conformal predictor Γ_M defines a *possibility distribution* over Y. Indeed, given $A \in im(\Gamma_M)$, it holds that $1 - \epsilon_A$ is a lower bound on the probability that the correct label y is in A.

Denote as $\emptyset \subset A_1 \subseteq ... \subseteq A_k$ the nested sets in $im(\Gamma_M)$. Given any $y \in Y$, let $j(y) = \max\{i : y \notin A_i\}$. Then, a possibility distribution π_Γ can be defined as follows [8]:

$$\pi_\Gamma(y) = \begin{cases} 1 & A_1 = \{y\} \\ \epsilon^{A_{j(y)}} & otherwise \end{cases} \tag{25}$$

The possibility distribution π_Γ can then be used to define a TWD classifier by transforming π_Γ into a probability distribution, so that the Loss function in Eq. (5) is well-defined. This transformation is performed by means of the possibility-probability transformation [9]:

$$Pr_{\pi_\Gamma}(y) = \sum_{i=1}^{k} \frac{\hat{\pi}_i - \hat{\pi}_{i+1}}{|B_i|} \mathbf{1}_{y \in B_i}, \tag{26}$$

where $\hat{\pi}$ is the ordering of π_Γ in terms of decreasing possibility value; B_i is the $\hat{\pi}_i$ α-cut (i.e., $B_i = \{y \in Y : \pi_\Gamma(y) \geq \hat{\pi}_i\}$). Then, the *possibilistic conformal three-way*(PCTWD) classifier is defined as:

$$\begin{aligned} \mathcal{W}_\Gamma^{poss}(x) &= Loss^{poss}_{\Gamma_M(x)}(A) \\ &= \arg\min_{A \in 2^Y} \sum_{y \notin A} Pr_{\pi_\Gamma}(y) \cdot err(A, y) + \\ &+ \alpha(|A|) \sum_{y \in A} Pr_{\pi_\Gamma}(y). \end{aligned} \tag{27}$$

Example 5 below provides an illustration of the calculations involved in the construction of a DCTWD and a PCTWD.

Example 5. *Consider the TWCP $\Gamma_\mathcal{W}$ and loss function defined in Example 3.*

Then, considering the instance x, it holds that $Loss^*_{\Gamma_\mathcal{W}(x)}(\{0,1\}) = \frac{4}{5} * \frac{1}{3} + \frac{1}{5} = 0.47$; while $Loss^*_{\Gamma_\mathcal{W}(x)}(Y) = 0.67$. Hence $\mathcal{W}^{dec}_{\Gamma_\mathcal{W}}(x) = \{0,1\}$.

By contrast, the corresponding PCTWD can be defined by noting that $\pi_\Gamma(0) = \pi_\Gamma(1) = 1$ and $\pi_\Gamma(2) = 0.25$, therefore $Pr_{\pi_\Gamma} = \langle 0 : 0.46, 1 : 0.46, 2 : 0.08 \rangle$.

Hence, $Loss^{poss}_\Gamma(0) = 0.54$, $Loss^{poss}_\Gamma(1) = 0.54$, while $Loss^{poss}_\Gamma(\{0,1\}) = 0.39$ and $Loss^{poss}_\Gamma(Y) = 0.67$. Therefore $\mathcal{W}^{poss}_{\Gamma_\mathcal{W}}(x) = \{0,1\}$.

The two above mentioned constructions allow to transform any conformal predictor into a cost-sensitive TWD classifier. Furthermore, it is easy to see that these constructions preserve validity. Indeed, for the case of a DCTWD $\mathcal{W}^{dec}_\Gamma$, the following result holds:

Proposition 4. *$Pr[y_x \notin \mathcal{W}^{dec}_\Gamma(x)] \leq \epsilon^A$, where ϵ^A is defined as in Eq. (22).*

Proof. By construction, it holds that $\mathcal{W}^{dec}_\Gamma(x) = A \in im(\Gamma(x))$. Then, the result follows by Theorem 1, and the definition of ϵ^A □

By contrast, for the case of a PCTWD $\mathcal{W}_\Gamma^{poss}$, there is no guarantee that $\mathcal{W}_\Gamma^{poss}(x) \in im(\Gamma(x))$. Nonetheless, a weaker bound can be obtained through the following result:

Proposition 5. *$Pr[y_x \notin \mathcal{W}_\Gamma^{poss}(x)] \leq \epsilon^{B^*}$, where*

$$B^* = \arg\max_{B \in im(\Gamma)(x): B \subseteq \mathcal{W}_\Gamma^{poss}(x)} |B|. \tag{28}$$

Proof. The result directly follows from the definition of Γ and $\mathcal{W}_\Gamma^{poss}(x)$. In particular, for all $B \subseteq B^*$ it holds that $Pr[y_x \notin \mathcal{W}_\Gamma^{poss}(x)] \leq \epsilon^B$. □

We now consider our main research question: namely, we ask under which conditions a given TWD classifier, and the corresponding DCTWD (resp. PCTWD) classifier, are equivalent. Such conditions would then establish an isomorphism between the class of TWD classifiers and (three-way) conformal predictors.

To this aim, let $\mathcal{W}$ be a TWD classifier, let $\Gamma_\mathcal{W}$ be the TWCP obtained from $\mathcal{W}$ and, finally, let $\mathcal{W}_{\Gamma_\mathcal{W}}^{dec}$ (resp. $\mathcal{W}_{\Gamma_\mathcal{W}}^{poss}$) be the corresponding DCTWD (resp. PCTWD). The following result provides sufficient and necessary conditions for the equivalence between the TWD classifier $\mathcal{W}$ and the DCTWD classifier $\mathcal{W}_{\Gamma_\mathcal{W}}^{dec}$.

Theorem 4. *Let $\mathcal{W}(x) = A$, then $\mathcal{W}_{\Gamma_\mathcal{W}}^{dec}(x) = A$ holds iff the following two conditions are satisfied:*

1. *$\exists \epsilon \in [0,1]$ s.t. $\mathcal{W}$ makes at least $\lfloor \epsilon \cdot (n+1) \rfloor$ predictions on S with risk greater than $\alpha(|A|)$ and makes at most $\lfloor \epsilon \cdot (n+1) \rfloor$ predictions on S with risk greater than $\min_{y \notin A} R(y, A)$;*
2. *$\epsilon^A \leq \min_{B \in im(\Gamma_\mathcal{W}^{dec}(x))} \frac{Loss^*_{\Gamma_\mathcal{W}}(B) - \alpha(|A|)}{\max_{y \not\vdash A} err(A,y) - \alpha(|A|)}$.*

Proof. The first condition, by Theorem 3, ensures that $\mathcal{W}(x) \in im(\Gamma_\mathcal{W}(x))$. The second condition, on the other hand, ensures that the transformation preserves the minimal element w.r.t. the ordering of 2^Y in terms of the *Loss* value. Thus, if both conditions hold, then $\mathcal{W}(x)$ is the unique solution to Eq. (24), and the result follows. □

The following corollary shows that, in the uniform-cost setting, the conditions required by Theorem 4 can be relaxed:

Corollary 3. *Let S be the current training set with $|S| = n$, $\mathcal{W}_S(x) = A$, then $\mathcal{W}_{\Gamma_\mathcal{W}}^{dec}(x) = A$ iff*

$$m \leq \frac{\alpha(|Y|) - \alpha(|A|)}{err - \alpha(|A|)} \cdot n, \tag{29}$$

where m is the number of errors made by $\mathcal{W}_S$.

Proof. The result directly follows from Theorems 2 and 4. □

We now discuss the case of the PCTWD classifier. First of all, irrespective of the non-conformity measure used, $\forall A \in im(\Gamma)$, the following proposition holds:

Proposition 6. *$Pr_{\pi_\Gamma}(A) \geq 1 - \epsilon^A$.*

Proof. Let j be s.t. $B_j = A$ (i.e. A is the i^{th} α-cut). Then:

$$\begin{aligned} Pr_{\pi_\Gamma}(A) &= (1-\epsilon^{B_1}) + (\epsilon^{B_1} - \epsilon^{B_2}) + ... \\ &+ (\epsilon^{B_{j-1}} - \epsilon^{A}) + |A| \sum_{i=j}^{k} \frac{\hat{\pi}_i - \hat{\pi}_{i+1}}{|B_i|} = \\ &= (1-\epsilon^A) + |A| \sum_{i=j}^{k} \frac{\hat{\pi}_i - \hat{\pi}_{i+1}}{|B_i|} \\ &\geq (1-\epsilon^A) + \frac{|A|}{|Y|}\epsilon^A > 1 - \epsilon^A \end{aligned} \tag{30}$$

□

This result implies, in particular, that $Loss^{poss}_{\Gamma}(A) \leq Loss^{*}_{\Gamma}(A)$.

Furthermore, note that if $\Gamma = \Gamma_{\mathcal{W}}$ (i.e. Γ is a TWCP) and the cost function is uniform, then the penultimate inequality in Eq. (30) holds with equality (as a consequence of Theorem 2). Then, the following result provides sufficient and necessary conditions for the equivalence between a TWD classifier and the corresponding PCTWD classifier in the uniform-cost setting:

Theorem 5. *Let S be the current training set. Let $\mathcal{W}_S(x) = A$. Then, $\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}}}(x) = A$ iff all the following conditions hold:*

$$\frac{\alpha(|A|+1)}{\alpha(|A|)} > f(|A|), \tag{31}$$

$$\forall k < |A|,\ 1-\epsilon^A > g(|A|,k), \tag{32}$$

$$\forall k > |A|,\ \epsilon^A \leq g(|A|,k), \tag{33}$$

where

$$\begin{aligned} f(|A|) &= \frac{1 - \frac{|A|}{|Y|}}{(|A|+1)(\frac{1}{|A|} - \frac{1}{|Y|})}, \\ g(|A|,k) &= \frac{\frac{k}{|A|}\alpha(k) - \alpha(|A|)}{D(|A|,k)}, \\ D(|A|,k) &= \frac{k}{|A|}\alpha(k) + \frac{k}{|Y|} + \frac{|A|}{|Y|}\alpha(|A|) \\ &\quad - \alpha(|A|) - \frac{k}{|Y|}\alpha(k) - \frac{|A|}{|Y|}. \end{aligned}$$

and ϵ^A is equal to ϵ_1 in Theorem 2.

Proof. The result directly follows from standard algebraic manipulations and the observation that, in the uniform-cost setting, the penultimate inequality in (30) holds with equality. □

The generalization of Theorem 5 to general, non-uniform, loss functions is left as an open problem.

In regard to the significance of Theorems 4 and 5, we discussed in Section 2.2 that, while TWD is optimal w.r.t. cost-sensitiveness, its results may in general be not valid. In particular, the latter may happen when the underlying classifier is not calibrated. Therefore, the transformation from a TWD classifier to a CP one (by means of TWCP and then, either, a DCTWD, or a PCTWD, classifier), can be seen as an approach to correct this lack of validity. In particular, then, Theorems 4 and 5 show that calibration is not a necessary condition for a TWD classifier to be valid, and provide conditions for validity.

Indeed, the two Theorems show that, under the condition that the TWD classifier $\mathcal{W}$ is sufficiently accurate, the correction implemented by means of CP has no effect. In this latter case, the set-valued predictions obtained before and after the validity correction are identical. Consequently, Theorems 4 and 5 establish an isomorphism between the class of (non-trivial) TWD classifiers and (three-way) conformal predictor. An illustration of these latter observations is shown in Example 6 and in Figure 2.

Example 6. *Let us refer to the TWCP $\Gamma_{\mathcal{W}}$ defined in Example 3 and the corresponding DCTWD and PCTWD classifiers defined in Example 5. In Example 5, it was shown that the predictions provided by the three TWD classifiers were equivalent, hence Theorems 4 and 5 should hold, as they provide sufficient and necessary conditions for such equivalences.*

Indeed, as regards the DCTWD, we note that $\mathcal{W}$ made exactly $1 < \frac{2/3-1/3}{1-1/3} \cdot |S| = 2$ error and thus the conditions in Theorem 4 are satisfied.

Similarly, with respect to the PCTWD, we note that Eq. (31) reduces to $2 > \frac{1-2/3}{3(1/2-1/3)} = 0.67$, Eq. (32) reduces to $\frac{4}{5} > 0.5$ and Eq. (33) reduces to $\frac{1}{5} < \frac{2}{5}$ which are all obviously satisfied.

4. Results

4.1. Experimental Design

The theoretical study of the previous sections shows some important connections between TWD and CP. In particular, it provides conditions for the equivalence among these two cautious classification methods. Based on these results, in this section, we describe a set of experiments to investigate the relationship between TWD and CP also from an empirical point of view.

More in detail, we address three research questions:

1. In Sections 3.1 and 3.2, we studied conditions for the equivalence between a TWD classifier $\mathcal{W}$ and the corresponding DCTWD (resp., PCTWD) classifier. In particular, we showed that these latter two classes of classifiers are equivalent, provided that the original TWD classifier is sufficiently accurate. Do these conditions hold in real-world datasets?

2. In Section 3.2 we proposed the DCTWD and PCTWD classifiers as techniques to obtain cost-sensitive cautious classifiers, starting from any conformal predictor. Nonetheless, we did not study any difference, in terms of validity or efficiency, between the DCTWD and PCTWD construction. Are there any empirical differences among these two latter methods, in terms of either classification accuracy or efficiency?
3. The proposed constructions can be seen as techniques to both improve the predictive performance of a TWD, as well as objective[2] approaches to obtain a cautious classifier from any CP method. Do these constructions result in an increase in predictive performance compared with other state-of-the-art TWD and CP algorithms?

To this end, we considered a set of experiments, based on 12 datasets from the UCI repository. These datasets are listed in Table 1.

Table 1: List of used datasets

Dataset	Instances	Features	Classes
Digits	1797	64	10
Breast Cancer	569	30	2
Wine	178	13	3
Covertype	581012	54	7
20Newsgroups	18846	130107	20
Diabetes	786	8	2
Epileptic Seizure	11500	179	2
Diabetic Retinopathy	1151	20	4
Hepatitis C virus	1385	29	4
Chronic Kidney Disease	400	25	2
Abalone	4177	8	27
Arrhythmia	452	279	16

We considered two different classes of scoring classifiers, namely Random Forest and k-Nearest Neighbors. For each of these latter two classes, we compared the results of 6 different methods:

- The (standard, single-valued prediction) classifiers h_{RF}, h_{KNN};
- The TWD classifiers $\mathcal{W}_{RF}$, $\mathcal{W}_{KNN}$;
- The TWCP-based DCTWD and PCTWD classifiers $\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}_{RF}}}$, $\mathcal{W}^{dec}_{\Gamma^{\mathcal{W}_{KNN}}}$ based on $\mathcal{W}_{RF}$, $\mathcal{W}_{KNN}$;
- The DTCWD and PCTWD classifier $\mathcal{W}^{dec}_{\Gamma^{h_{RF}}}$, $\mathcal{W}^{dec}_{\Gamma^{h_{KNN}}}$, directly based on h_{RF}, h_{KNN} (see Section 2.3).

[2]Here, by objective it is meant that there is no a-priori selection of a probability threshold

The loss function (used to determine the TWD classifiers and to evaluate the performance of the models) was defined through the following abstention cost function:

$$\alpha(n) = \frac{n-1}{|Y|}, \tag{34}$$

while the *err* function was

- Uniform, with all costs equal to 1, for the Abalone, Digits, Wine, Covertype and 20Newsgroups datasets;
- Equal to 1 when the true class label was associated to healthy status, and equal to 2 otherwise, for the medical datasets.

The CP algorithms were implemented using the inductive approach, i.e., by relying on a validation set. The size of the validation set was set to 20% of the training set. We decided to use the inductive approach, rather than the sequential one, in order to reduce the computational cost of re-training the classification algorithms.

All algorithms were evaluated in terms of the complement of the above mentioned loss function[3], henceforth *accuracy*, as well as in terms of coverage, as a measure of efficiency. This latter measure, in particular, was defined, for a set-valued prediction S, as:

$$coverage(S) = 1 - \frac{|S|-1}{|Y|-1}. \tag{35}$$

All performances were computed using 5-fold cross-validation. Thus, we report the results in terms of both the average performance and the corresponding 95% confidence interval. In order to assess the presence of statistically significant differences, if any, we performed the Friedman rank test. Namely, for each cautious learning approach and each model class (Random Forest, kNN), we computed the ranks with respect to each of the considered datasets; for each cautious learning approach and each dataset we then averaged the Random Forest and kNN ranks.

4.2. Experimental Results

The results of the Experiments, in terms of the average accuracy, are reported in Tables 2, 3 and in Figure 3. The average coverage values are reported in Tables 4, 5 and Figure 4.

As regards the observed accuracies, the average ranks are reported in Table 6: the observed test statistic was $Q = 36.11$, which was significant at the 95% confidence level (p-value < 0.00001). Thus, we also performed a post-hoc pairwise comparison using the Nemenyi test procedure. The critical value of the

[3]Note that when the *err* function was uniform, then the complement of the loss function is equivalent to a penalized accuracy, in which the penalization depends on the size of the set-valued prediction

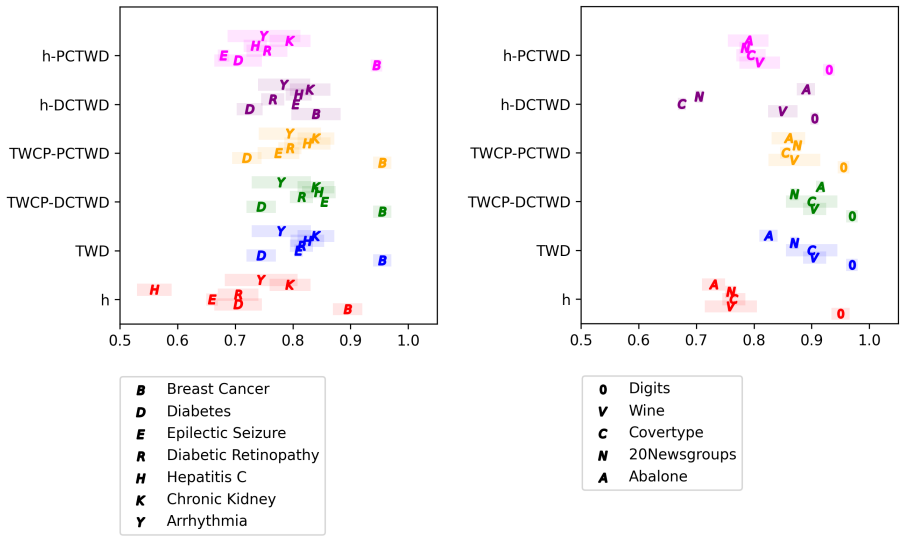

Figure 3: Average accuracy and 95% confidence intervals for each of the evaluated classifiers, on the medical (left) and non-medical (right) datasets. Each marker refers to the average of the corresponding Random Forest-based and kNN-based classifiers. In the legend, h denotes the standard classifiers (h_{RF}, h_{KNN}); TWD the three-way decision classifiers ($\mathcal{W}_{RF}$, $\mathcal{W}_{KNN}$); TWCP-DCTWD (resp., TWCP-PCTWD) the DCTWD (resp., PCTWD) classifier based on the TWCP ($\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}$, $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}$); while h-DCTWD (resp., h-PCTWD) the DCTWD (resp. PCTWD) classifier based on the standard conformal predictors ($\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}$, $\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}$).

test (with 12 datasets and 6 compared methods), at the 95% confidence level, is 2.176. The pairwise comparisons are reported in Table 7.

As regards the observed coverage values, the average ranks are reported in Table 8. The observed test statistic was $Q = 26.33$ which was significant at the 95% confidence level (p-value = 0.00003). Thus, we performed the Nemenyi post-hoc pairwise test. The critical value of the test (with 12 dataset 5 compared methods), at the 95% confidence level, is 1.761. The pairwise comparison are reported in Table 9.

4.3. Discussion

Commenting the results reported in Section 4.2, we can see that the TWD classifiers (i.e., $\mathcal{W}$, $\Gamma^{dec}_{\mathcal{W}}$ and $\Gamma^{poss}_{\mathcal{W}}$) outperformed the corresponding single-valued classifiers in terms of accuracy. This finding should not be surprising. Indeed, the considered cautious classifiers are, by construction, *cost-sensitive*. Hence, they always return the set-valued prediction that maximizes the accuracy. Nonetheless, it shows that both TWD classifiers and the corresponding CP-based corrections can be useful to obtain significantly improved predictive performance (if set-valued predictions are allowed).

More interestingly, we can observe that the standard CP-based classifiers (i.e., Γ^{dec}_{h} and Γ^{poss}_{h}) were not significantly different from the single-valued clas-

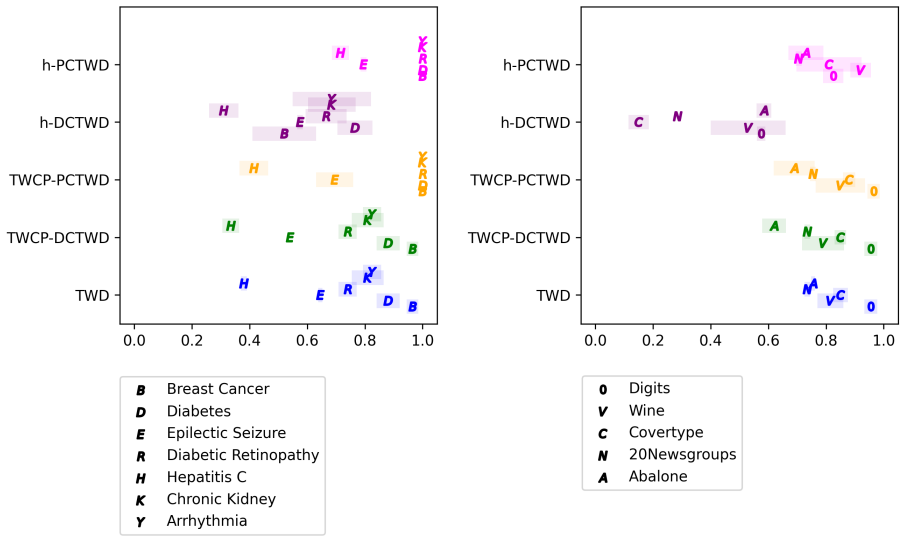

Figure 4: Average accuracy and 95% confidence intervals for each of the evaluated classifiers, on the medical (left) and non-medical (right) datasets. Each marker refers to the average of the corresponding Random Forest-based and kNN-based classifiers. In the legend, h denotes the standard classifiers (h_{RF}, h_{KNN}); TWD the three-way decision classifiers ($\mathcal{W}_{RF}$, $\mathcal{W}_{KNN}$); TWCP-DCTWD (resp., TWCP-PCTWD) the DCTWD (resp., PCTWD) classifier based on the TWCP ($\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}$, $\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}$); while h-DCTWD (resp., h-PCTWD) the DCTWD (resp. PCTWD) classifier based on the standard conformal predictors ($\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}$, $\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}$).

Table 2: Average loss value and 95% confidence intervals for the Random Forest-based classifiers, on all 12 datasets.

Dataset	h_{RF}	$\mathcal{W}_{RF}$	$\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}_{RF}}}$	$\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma_{h_{RF}}}$
Abalone	0.79 ± 0.02	0.91 ± 0.01	$\mathbf{0.92 \pm 0.01}$	0.91 ± 0.01	0.90 ± 0.01	0.85 ± 0.03
Arrhythmia	0.80 ± 0.04	0.81 ± 0.02	0.81 ± 0.02	$\mathbf{0.84 \pm 0.03}$	0.80 ± 0.04	0.81 ± 0.04
Breast Cancer	0.86 ± 0.03	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	0.95 ± 0.01	0.96 ± 0.01
Chronic Kidney	0.85 ± 0.03	$\mathbf{0.87 \pm 0.02}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{0.87 \pm 0.02}$	0.85 ± 0.03	0.85 ± 0.03
Covertype	0.82 ± 0.02	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.93 \pm 0.01}$	0.85 ± 0.01	0.88 ± 0.03
Diabetes	0.73 ± 0.05	$\mathbf{0.77 \pm 0.02}$	$\mathbf{0.77 \pm 0.02}$	0.74 ± 0.02	0.75 ± 0.01	0.73 ± 0.05
Diabetic Retinopathy	0.78 ± 0.04	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.84 \pm 0.02}$	0.81 ± 0.02	0.82 ± 0.02	0.78 ± 0.04
Digits	0.94 ± 0.02	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.95 \pm 0.01}$	0.91 ± 0.01	0.90 ± 0.01
Epileptic Seizure	0.73 ± 0.01	$\mathbf{0.88 \pm 0.00}$	$\mathbf{0.88 \pm 0.00}$	0.77 ± 0.05	0.86 ± 0.00	0.70 ± 0.01
Hepatitis C	0.56 ± 0.03	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.86 \pm 0.04}$	0.79 ± 0.03	0.77 ± 0.02
Wine	0.81 ± 0.05	$\mathbf{0.96 \pm 0.02}$	$\mathbf{0.96 \pm 0.02}$	$\mathbf{0.96 \pm 0.02}$	0.92 ± 0.02	0.91 ± 0.03
20Newsgroups	0.85 ± 0.01	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.91 \pm 0.01}$	0.81 ± 0.01

sifiers in terms of accuracy. In particular, the PCTWD classifier was significantly outperformed by all TWD-based classifiers. Thus, we can provide a positive an-

Table 3: Average loss value and 95% confidence intervals for the kNN-based classifiers, on all 12 datasets.

Dataset	h_{KNN}	$\mathcal{W}_{KNN}$	$\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{KNN}}}$	$\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}_{KNN}}}$	$\mathcal{W}^{dec}_{\Gamma_{h_{KNN}}}$	$\mathcal{W}^{poss}_{\Gamma_{h_{KNN}}}$
Abalone	0.67 ± 0.02	0.74 ± 0.02	$\mathbf{0.91 \pm 0.00}$	0.81 ± 0.04	0.88 ± 0.02	0.73 ± 0.04
Arrhythmia	0.69 ± 0.08	0.75 ± 0.07	0.75 ± 0.07	0.75 ± 0.07	$\mathbf{0.77 \pm 0.05}$	0.69 ± 0.08
Breast Cancer	0.93 ± 0.02	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.94 \pm 0.02}$	0.73 ± 0.06	0.93 ± 0.01
Chronic Kidney	0.74 ± 0.04	$\mathbf{0.81 \pm 0.04}$	$\mathbf{0.81 \pm 0.04}$	$\mathbf{0.81 \pm 0.04}$	$\mathbf{0.81 \pm 0.04}$	0.74 ± 0.04
Covertype	0.71 ± 0.02	$\mathbf{0.87 \pm 0.06}$	$\mathbf{0.87 \pm 0.06}$	0.78 ± 0.01	0.50 ± 0.00	0.71 ± 0.02
Diabetes	0.68 ± 0.03	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.72 \pm 0.03}$	0.70 ± 0.03	0.70 ± 0.03	0.68 ± 0.03
Diabetic Retinopathy	0.63 ± 0.03	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.79 \pm 0.02}$	0.78 ± 0.02	0.71 ± 0.02	0.73 ± 0.03
Digits	0.96 ± 0.01	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	0.96 ± 0.01	0.90 ± 0.00	0.96 ± 0.01
Epileptic Seizure	0.59 ± 0.01	0.74 ± 0.01	$\mathbf{0.83 \pm 0.00}$	0.78 ± 0.01	0.75 ± 0.00	0.66 ± 0.01
Hepatitis C	0.56 ± 0.03	0.79 ± 0.01	$\mathbf{0.83 \pm 0.01}$	0.79 ± 0.04	$\mathbf{0.83 \pm 0.01}$	0.70 ± 0.02
Wine	0.71 ± 0.04	$\mathbf{0.85 \pm 0.02}$	$\mathbf{0.85 \pm 0.02}$	0.78 ± 0.06	0.78 ± 0.03	0.71 ± 0.04
20Newsgroups	0.67 ± 0.01	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$	0.50 ± 0.00	0.76 ± 0.01

Table 4: Average coverage value and 95% confidence intervals for the Random Forest-based classifiers, on all 12 datasets.

Dataset	$\mathcal{W}_{RF}$	$\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}_{RF}}}$	$\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma_{h_{RF}}}$
Abalone	$\mathbf{0.74 \pm 0.01}$	0.64 ± 0.05	0.69 ± 0.01	0.60 ± 0.02	0.69 ± 0.05
Arrhythmia	0.86 ± 0.02	0.86 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.82 ± 0.09	$\mathbf{1.00 \pm 0.00}$
Breast Cancer	0.97 ± 0.02	0.97 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.54 ± 0.10	$\mathbf{1.00 \pm 0.00}$
Chronic Kidney	0.83 ± 0.05	0.83 ± 0.05	$\mathbf{1.00 \pm 0.00}$	0.75 ± 0.06	$\mathbf{1.00 \pm 0.00}$
Covertype	$\mathbf{0.76 \pm 0.03}$	$\mathbf{0.76 \pm 0.02}$	$\mathbf{0.76 \pm 0.03}$	0.30 ± 0.05	0.62 ± 0.16
Diabetes	0.94 ± 0.04	0.94 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.81 ± 0.05	$\mathbf{1.00 \pm 0.00}$
Diabetic Retinopathy	0.73 ± 0.02	0.73 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.74 ± 0.01	$\mathbf{1.00 \pm 0.00}$
Digits	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.93 \pm 0.03}$	0.60 ± 0.02	0.65 ± 0.05
Epileptic Seizure	0.58 ± 0.01	0.58 ± 0.00	0.77 ± 0.09	0.47 ± 0.01	$\mathbf{0.81 \pm 0.01}$
Hepatitis C	0.44 ± 0.01	0.44 ± 0.01	0.44 ± 0.01	0.25 ± 0.06	$\mathbf{0.64 \pm 0.04}$
Wine	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.87 \pm 0.05}$	0.44 ± 0.16	$\mathbf{0.87 \pm 0.05}$
20Newsgroups	$\mathbf{0.64 \pm 0.01}$	$\mathbf{0.64 \pm 0.01}$	$\mathbf{0.64 \pm 0.01}$	0.57 ± 0.00	0.59 ± 0.01

swer to our third experimental research question. Indeed, the proposed methods out-performed the state-of-the-art CP methods, in terms of predictive accuracy, with comparable or even better efficiency.

In regard to our first research question, we can see that there were no significant differences among the three TWD-based cautious classifiers (i.e., $\mathcal{W}$, $\Gamma^{dec}_{\mathcal{W}}$ and $\Gamma^{poss}_{\mathcal{W}}$). This finding lends empirical support to the results proven in Section 3. Indeed, in Section 3, we proved that a TWD classifier is equivalent to the corresponding CP-based model, provided the original TWD classifier is sufficiently accurate. The experimental results, then, show that the conditions of Theorems 3, 4, 4 are usually satisfied in real-world setting. Hence, TWD methods

Table 5: Average coverage value and 95% confidence intervals for the kNN-based classifiers, on all 12 datasets.

Dataset	$\mathcal{W}_{RF}$	$\mathcal{W}^{dec}_{\Gamma_{\mathcal{W}_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma_{\mathcal{W}_{RF}}}$	$\mathcal{W}^{dec}_{\Gamma_{h_{RF}}}$	$\mathcal{W}^{poss}_{\Gamma_{h_{RF}}}$
Abalone	$\mathbf{0.77 \pm 0.00}$	0.60 ± 0.03	0.69 ± 0.10	0.57 ± 0.03	$\mathbf{0.77 \pm 0.07}$
Arrhythmia	0.79 ± 0.04	0.79 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.55 ± 0.17	$\mathbf{1.00 \pm 0.00}$
Breast Cancer	0.96 ± 0.02	0.96 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.50 ± 0.12	$\mathbf{1.00 \pm 0.00}$
Chronic Kidney	0.79 ± 0.06	0.79 ± 0.06	$\mathbf{1.00 \pm 0.00}$	0.62 ± 0.10	$\mathbf{1.00 \pm 0.00}$
Covertype	0.94 ± 0.02	0.94 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$
Diabetes	0.82 ± 0.04	0.82 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.72 ± 0.07	$\mathbf{1.00 \pm 0.00}$
Diabetic Retinopathy	0.75 ± 0.04	0.75 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.59 ± 0.10	$\mathbf{1.00 \pm 0.00}$
Digits	0.98 ± 0.01	0.98 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.55 ± 0.00	$\mathbf{1.00 \pm 0.00}$
Epileptic Seizure	0.71 ± 0.00	0.50 ± 0.00	0.62 ± 0.01	0.68 ± 0.01	$\mathbf{0.78 \pm 0.01}$
Hepatitis C	0.32 ± 0.01	0.23 ± 0.04	0.39 ± 0.07	0.37 ± 0.04	$\mathbf{0.79 \pm 0.01}$
Wine	0.76 ± 0.04	0.71 ± 0.09	0.83 ± 0.11	0.62 ± 0.09	$\mathbf{0.97 \pm 0.01}$
20Newsgroups	0.83 ± 0.01	0.83 ± 0.01	$\mathbf{0.87 \pm 0.01}$	0.00 ± 0.00	0.82 ± 0.01

Table 6: Average ranks of the compared learning algorithms, in terms of observed loss value, according to the Friedman test procedure.

	h	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
Average rank	5.29	2.23	1.83	2.67	4.01	4.91

Table 7: Pairwise differences in ranks, in terms of observed loss values, among the compared learning algorithms. Statistically significant differences (according to the Nemenyi test) are denoted in bold and with an asterisk.

	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
h	$\mathbf{3.06}^*$	$\mathbf{3.46}^*$	$\mathbf{2.62}^*$	1.28	0.38
$\mathcal{W}$	-	0.40	0.44	1.78	$\mathbf{2.68}^*$
$\Gamma^{dec}_{\mathcal{W}}$	-	-	0.84	$\mathbf{2.18}^*$	$\mathbf{3.08}^*$
$\Gamma^{poss}_{\mathcal{W}}$	-	-	-	1.34	$\mathbf{2.24}^*$
Γ^{dec}_{h}	-	-	-	-	0.9

Table 8: Average ranks of the compared learning algorithms, in terms of observed coverage, according to the Friedman test procedure.

	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
Average rank	3.02	3.46	2	4.75	1.90

Table 9: Pairwise differences in ranks, in terms of observed coverage values, among the compared learning algorithms. Statistically significant differences (according to the Nemenyi test) are denoted in bold and with an asterisk.

	$\mathcal{W}$	$\Gamma^{dec}_{\mathcal{W}}$	$\Gamma^{poss}_{\mathcal{W}}$	Γ^{dec}_{h}	Γ^{poss}_{h}
$\mathcal{W}$	-	0.44	1.02	1.73	1.12
$\Gamma^{dec}_{\mathcal{W}}$	-	-	1.46	1.29	1.36
$\Gamma^{poss}_{\mathcal{W}}$	-	-	-	$\mathbf{2.75}^*$	0.10
Γ^{dec}_{h}	-	-	-	-	$\mathbf{2.85}^*$

can usually be expected to have the same level of validity as the corresponding CP-based correction.

More in detail, in the case of Random Forest, the results for the TWD-based cautious classifiers were almost always identical. By contrast, in the case of kNN, there were 4 datasets on which the DCTWD and PCTWD classifiers achieved increased performance. This observation can be explained by noting that kNN classifiers usually have lower accuracy and generalization capability than Random Forest ones. As a consequence of the results in Section 3, this observation implies that TWD classifiers based on kNN are expected to satisfy the conditions of Theorems 3, 4 and 5 less often than classifiers based on Random Forest. Therefore, the proposed TWCP, DCTWD and PCTWD constructions would result in less efficient (but more accurate) predictions, as observed in Tables 3, 5. More generally, the proposed construction can be applied to improve the predictive accuracy of any TWD classifier whose underlying single-valued ML models may be prone to either under- or over-fitting, as in the case of kNN.

In regard to our second research question, we did not find significant differences among the DCTWD classifiers and the PCTWD classifiers in terms of predictive accuracy, though the PCTWD classifiers were on average less accurate than the DCTWD ones. On the other hand, in terms of efficiency, the PCTWD classifiers reported a larger coverage than the DCTWD ones. In particular, the difference between Γ_h^{dec} and Γ_h^{poss} was statistically significant.

Thus, the DCTWD and PCTWD classifiers offer a trade-off between greater accuracy (for the DCTWD classifier) and greater efficiency (for the PCTWD classifier). The selection among the two methods should then be made by the decision-maker, based on the quality dimension which is deemed most important for the specific decision-making task at hand.

As a general final remark, we focus on the TWD-based classifiers, i.e., the TWD classifier $\mathcal{W}$, the DCTWD classifier $\mathcal{W}_{\Gamma_{\mathcal{W}}}^{dec}$ and the PCTWD classifier $\mathcal{W}_{\Gamma_{\mathcal{W}}}^{poss}$. Compared with $\mathcal{W}$, the DCTWD classifier reported, on average, improved predictive accuracy but slightly lower efficiency. By contrast, the PCTWD reported, on average, improved efficiency with comparable but slightly reduced accuracy. Therefore, the application of the proposed CP-based corrections could be useful not only for classifiers whose predictions are insufficiently accurate, or for classifiers that are known to be prone to over-fitting, but also for more general TWD classifiers. Indeed, in the worst case situation, the set-valued predictions provided by TWD and the corresponding DCTWD and PCTWD would be equivalent, as a consequence of the results in Section 3. In all other cases, however, the proposed constructions would allow to achieve either more accurate (using the DCTWD classifier) or more specific (using the PCTWD classifier) predictions, compared with a standard TWD classifier.

5. A Medical Case Study

Up to now, we have discussed the relationship between TWD and CP. Through this relationship we studied some formal property of TWD, by introducing the TWCP, DCTWD and PCTWD classifiers as a means to both

study validity bounds for TWD and to improve the validity of standard TWD classifiers. In this section, we address the potential of the proposed approach for human decision making in classification tasks and, therefore, for its integration into Decision Support Systems.

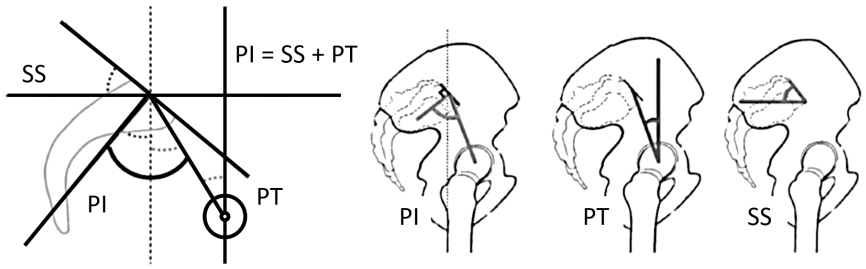

Figure 5: The main angles considered in the sagittal imbalance classification.

As we argued in the Introduction, cautious learning approaches could be useful to develop valid and reliable decision support in human decision making. Nonetheless, to the knowledge of the authors, no previous study evaluated the usefulness of such set-valued advice compared to standard support. Indeed, even though the recent study by Liu et al. [23] assessed the effectiveness of TWD from the perspective of interpretability, the authors did not specifically evaluate the usefulness of set-valued advice.

In order to understand whether set-valued advice could be supportive in naturalistic decision making [19], we tested this approach in the case of the assessment of sagittal misalignment. This latter is a kind of spine deformity regarding an imbalance along the front-to-back direction of the outward curve of the middle spine called kyphosis.

We chose this case for three main reasons. First, there is a lack of standard criteria to classify imbalance [20], as this is characterized in terms of a number of angles, among which the main ones are called pelvic tilt (PT), sacral slope (SS) and pelvic incidence (PI, which can be defined as the sum of PT and SS - see Figure 5). Second, it fits well a set-valued output. Indeed, real cases form a continuous range, where specific instances of pathological shape of the spine might be borderline, sharing characteristics of two "adjacent" patterns. On the other hand, existing classification schema provide discrete and mutually exclusive categories by which to characterize spine misalignment. Lastly, and more importantly, the diagnosis of this kind of spine deformity is strongly related to treatment. That is, recognizing a kyphosis type, and therefore classifying sagittal misalignments into a specific pattern, provides spine surgeons with a range of treatment guidelines to restore a physiological profile and reduce the odds of adverse events or of poor outcome [3].

To this aim, we considered a dataset of 120 patients (26 male subjects), whose imaging and sets of 14 spine angles were analyzed and annotated by two

senior expert spine surgeons. The two surgeons annotated each case with one out of 7 mutually exclusive labels, namely: normal and 6 different types of kyphosis. The normal cases (N) were 14% of the sample; of the abnormal cases, 36% were affected by lumbar kyphosis (L), 23% suffered from thoracic kyphosis (T), 21% from global kyphosis (G), 17% from thoracolumbar kyphosis (TL), while the other disorders (Lower Lumbar (LL), Cervical (C)) accounted for the remaining 9%.

As proof that the classification task was not a trivial one, although the two expert surgeons shared a taxonomic framework that they had jointly published [20], they could only agree on slightly more than two thirds of the cases (68.3%) and only exhibited a moderated agreement (Cohen's Kappa and Krippendorff's Alpha both equal to 0.62). Thus, the considered dataset was a natural example of the weakly supervised learning setting, discussed in Section 2. So, for model training, we applied the techniques proposed in Section 3.1, using as base TWD classifier the state-of-the-art TW Random Forest method [5].

One of the authors, an expert spine surgeon, reviewed 15 predictions provided by a classical (weakly supervised) predictive model h (which also gave the probability score associated with the diagnostic advice), with moderate accuracy (approximately 70%), and compared them with the set-valued predictions provided by a corresponding DCTWD classifier (defined on the basis of a TWD $\mathcal{W}_h$ classifier grounding on h), together with the related probability bound as described in Section 3.2. See Table 10 for a brief summary of the annotated cases. The spine surgeon evaluated the usefulness of the advice and, then, discussed about the rationale for the potential adoption of these approaches in clinical decision support.

Table 10: Summary of the information regarding the medical cases reviewed by the domain expert. For each case, we report both the single-valued prediction and the set-valued prediction provided by the DCTWD (in parentheses, the probability scores of the two methods), the target labels, and the perceived usefulness of the two types of predictions, measured in an ordinal scale ranging from 1 (very low) to 5 (very high). We also report, for each case, the pelvic tilt (PT) and sacral slope (SS) angles.

Case ID	PT	SS	Target	h (prob.)	DCTWD (prob.)	Usefulness (h)	Usefulness (DCTWD)
83	29	8	L	L (0.68)	L, LL (0.86)	4	4
8	23	24	G	L (0.64)	L, G (0.86)	5	5
87	43	24	L	L (0.76)	L (0.86)	4	5
100	11	22	TL, L	L (0.32)	N, TL, L (0.71)	3	5
115	21	35	TL, LL	T (0.62)	T, L, LL (0.86)	5	3
71	27	12	G, L	G (0.29)	T, L, LL, G (0.94)	5	2
3	39	4	L	L (0.33)	L, G (0.71)	3	5
24	16	39	LL	L (0.71)	L, LL (0.94)	3	4
101	13	33	TL	T (0.64)	T, TL (0.86)	2	4
4	32	26	L	T (0.74)	T, L (0.94)	2	4
124	57	16	L	L (0.75)	L, LL (0.94)	5	3
109	16	28	N	N (0.67)	N, T (0.86)	4	5
104	22	24	N, LL	N (0.32)	N, L, LL (0.71)	5	4
2	20	19	L	N (0.33)	N, T, L (0.86)	2	5
61	15	44	N	N (0.89)	N, L (0.86)	5	4

(a) A sagittal x-ray for the case no. 100. This depicts a thoracolumbar kyphosis that is so light that the alignment is almost normal.

(b) A sagittal x-ray that depicts a manifest, easy-to-detect thoracolumbar kyphosis.

From the quantitative point of view, the output of the DCTWD classifier was found to be more useful (or informative), with a mean score of 4.13 (SD=1.21) (vs 3.80, SD=0.92), but not significantly so (Mann Whitney test, $U = 97.5$. p-value = 0.55).

On a more qualitative level, the traditional approach was deemed preferable whenever the classifier would be able to provide the decision makers with highly-

confident advice. In the case at hand, which we recall was a 7-class diagnostic task, a probability score for a single class was considered high if it was at least 3 times higher than those from a uniform probability distribution (i.e., 1/6). Nonetheless, also the conformal approach proposed in this paper was deemed valuable in these cases, for its capability to point out the most plausible alternatives, that the decision maker could further consider to definitely rule them out in favor of the diagnostic class singled-out by the traditional approach.

Conversely, when the traditional approach gives predictions associated with low confidence scores, the set-valued output (provided by the DCTWD) was found to be more useful. This was mainly the case because the set-valued prediction provides an enumeration of the main alternatives, and thus indirectly suggests what further evidence or elements should be considered by the decision maker to rule out some options and keep those that are more compatible with the case at hand.

In the 15 cases considered in this evaluation, the set-valued predictions provided by the DCTWD of alternatives were deemed to be always close to the set of natural candidates for the correct diagnosis that an expert surgeon would have considered if unaided (only for case 115 the DCTWD did not include the label TL in the set-valued predictions, although it included both T and L).

Generalizing, we can assert that whenever the traditional approach provides confidence scores close to the uniform probability distribution, and the probability bounds of the DCTWD are sufficiently high, then presenting both these pieces of advice would be the best option.

The medical expert also provided some comments on two noteworthy cases that we report in what follows, to highlight the kind of reasoning that cautious learning can facilitate in diagnostic tasks:

- Case 100 was described as an odd one (see Figure 6a). The involved surgeon said that it was probably a normal subject (because the pelvic tilt and the SVA were normal and the combined normality of both parameters leaves little room for a pathological case to be confirmed), who nevertheless exhibited a value of lumbar lordosis that was too low. He agreed upon the fact that other, equally expert, colleagues could have defined the unusual shape of the spine exhibited by case 100 (presenting small pelvic incidence, relatively small lumbar lordosis and small thoracic kyphosis) as unharmonious and weird, irrespective of its occurrence in asymptomatic subjects. Interestingly, the DCTWD was capable to capture this "oddness" and it provided a set-valued prediction that encompassed both normality, lumbar lordosis and thoracic kyphosis as plausible labels.

- Case 8 was deemed extremely interesting. In [20], subjects like case 8, who present values of lumbar lordosis lower than the normative values, and thoracic kyphosis above the normative values, are considered clear instances of global kyphosis. Nonetheless, insufficient lumbar lordosis, combined with decreased thoracic kyphosis, indicates cases of manifest lumbar kyphosis. This puts patients like case 8 in an area of uncertainty, and no current spine deformity classification can associate these patients

with a clear-cut category, without the risk of misdiagnosis. Interestingly, the DCTWD method recognizes and reflects this intrinsic uncertainty, by not imposing any specific diagnosis over the others.

The described qualitative evaluation, and the brief discussion of the two cases mentioned above, are just exemplifications. Nonetheless, they allow us to hint at how computational tools, like those integrating some form of machine learning, can support human reasoning, and how decision makers and these tools should interact in naturalistic settings and real-world scenarios.

This latter aspect also relates to *how* plausible classes should be presented, that is *how many* and whether in terms of confidence or probability. Likewise, the usefulness of set-valued predictions was appreciated in almost all the decision settings, as long as the interval did not encompass more than 3 or 4 alternative candidates, irrespective of the number of potential disjoint options.

In our short, but indicative, use case, we showed how human decision makers can collect observations in a medical scenario; combine this information with knowledge on spinal bio-mechanics, developed in either direct or indirect clinical experience (e.g., historical trial and errors, case reports, clinical comparisons); and formulate hypotheses on the basis of what a computational decision support gives them. In regard to set-valued output, we saw how this type of support can reflect compatible patterns of spine deformation and compensation, and hence be a useful aid to choose appropriate treatments even if a single option is not highlighted. In fact, the predictions provided by the DCTWD classifier were found to be useful even for the cases for which traditional systems could suggest a single diagnosis with high accuracy, because they acted as triggers for double check and review of less-than-obvious options.

In light of our study, we then make the point that decision support in real-world settings should always leverage some form of cautious prediction; either in conjunction with more traditional approaches, or in isolation. Their usefulness especially emerges in those cases where real life comes in shades of grey, and even well-trained and long-experienced experts cannot classify specific cases with total certainty. Cautious learning approaches can better reflect this intrinsic uncertainty, compared with traditional approaches, and thus they could provide more useful and interpretable decision support [14, 23] for decision makers in critical settings, or for under-specified tasks.

6. Conclusion

In this article, we studied the relationship between TWD and CP, two popular cautious learning approaches. To this aim, we introduced the *three-way non-conformity measure*, as well as the *three-way conformal predictor* (TWCP), and discussed two classes of conformal TWD classifiers (i.e., the DCTWD and PCTWD classifiers) by which a conformal predictor can be transformed into a TWD classifier. Through this relationship, the validity of TWD-based ML models is proven for the first time (to our knowledge): this allows to establish reliable learning-theoretic guarantees and error bounds for TWD classifiers.

Furthermore, the definition of optimal cost-sensitive cautious classification algorithms is addressed, along with a characterization of the conditions under which CP and TWD would provide identical results.

From an empirical point of view, we illustrated how the proposed constructions can be used to obtain TWD classifiers that were shown to outperform state-of-the-art TWD, and CP, methods.

Finally, we highlighted the positive potential of the proposed approaches – and cautious learning methods more in general – in the development of reliable decision support, through an illustrative use case, involving a subject-matter expert in a complex medical classification problem.

In conclusion, we believe that our theoretical analysis and the promising results from the empirical study represent a first step, as well as a foundation, for further investigations aimed at characterizing the theoretical aspects of TWD-based ML, and of cautious-learning approaches more in general. For this reason, we believe that the following open problems should be further investigated:

- In Section 3.2, a characterization of the conditions for the equivalence between a TWD classifier and the corresponding PCTWD classifier was proved, under the assumption of a uniform-error loss function. It would be interesting to generalize this characterization to general-loss functions;

- In this paper, we focused on the most basic notion of *validity* (i.e. conservative validity). It would thus be interesting to study also the *probabilistic validity* of TWD classifiers, or their validity in non-i.i.d. settings [2];

- The three-way non-conformity measure was introduced to define CP algorithms based on TWD classifiers. Though this approach allowed a natural comparison among the two studied approaches, it is not optimal in terms of efficiency. It would thus be interesting to study appropriate generalizations of other, efficient [31], CP approaches to the TWD setting;

- The proven validity bounds are instance-wise and can be applied in both online and inductive settings (using a validation set). Nonetheless, it could be interesting to study validation-independent finite-sample bounds. This would require, in turn, to generalize the framework of *PAC learning* theory to TWD-based ML and, more in general, to cautious learning [12];

- Finally, through a simple but indicative case study, in Section 5, we discussed the usefulness of the proposed approaches to develop more reliable and supportive decision support tools. We deem that further assessing the perceived usefulness of TWD, CP and other cautious learning approaches as support tools for human decision makers could be of great interest towards the development of truly reliable Decision Support Systems.

References

[1] Adamskiy, D., Nouretdinov, I., & Gammerman, A. (2011). Conformal predictors in semisupervised case. *Statistical Learning and Data Science*, (p. 43).

[2] Balasubramanian, V., Ho, S.-S., & Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

[3] Bhagat, S., Vozar, V., Lutchman, L., Crawford, R., & Rai, A. (2013). Morbidity and mortality in adult spinal deformity surgery: Norwich spinal unit experience. *European Spine Journal*, *22*, 42–46.

[4] Bosc, N., Atkinson, F., Felix, E., Gaulton, A., Hersey, A., & Leach, A. R. (2019). Large scale comparison of qsar and conformal prediction methods and their applications in drug discovery. *Journal of cheminformatics*, *11*, 4.

[5] Campagner, A., Cabitza, F., & Ciucci, D. (2020). The three-way-in and three-way-out framework to treat and exploit ambiguity in data. *International Journal of Approximate Reasoning*, *119*, 292–312.

[6] Campagner, A., Cabitza, F., & Ciucci, D. (2020). Three–way decision for handling uncertainty in machine learning: a narrative review. In *Proceedings of International Joint Conference on Rough Sets 2020* (pp. 137–152). Springer volume 12179 of *LNCS*.

[7] Del Coz, J. J., Díez, J., & Bahamonde, A. (2009). Learning nondeterministic classifiers. *Journal of Machine Learning Research*, *10*.

[8] Dubois, D., & Prade, H. (2016). Practical methods for constructing possibility distributions. *International Journal of Intelligent Systems*, *31*, 215–239.

[9] Dubois, D., Prade, H., & Sandri, S. (1993). On possibility/probability transformations. In *Fuzzy logic* (pp. 103–112). Springer.

[10] Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (pp. 973–978). Lawrence Erlbaum Associates Ltd volume 17.

[11] Ferri, C., & Hernández-Orallo, J. (2004). Cautious classifiers. *ROCAI*, *4*, 27–36.

[12] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in neural information processing systems* (pp. 4878–4887).

[13] Gu, P., Liu, J., & Zhou, X. (2021). Approaches to three-way decisions based on the evaluation of probabilistic linguistic terms sets. *Symmetry*, *13*, 764.

[14] Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, *3*, 119–131.

[15] Hüllermeier, E. (2014). Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, *55*, 1519–1534.

[16] Hüllermeier, E., Destercke, S., & Couso, I. (2019). Learning from imprecise data: adjustments of optimistic and pessimistic variants. In *International Conference on Scalable Uncertainty Management* (pp. 266–279). Springer.

[17] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*, 457–506.

[18] Johansson, U., Boström, H., & Löfström, T. (2013). Conformal prediction using decision trees. In *2013 IEEE 13th international conference on data mining* (pp. 330–339). IEEE.

[19] Klein, G. (2008). Naturalistic decision making. *Human factors*, *50*, 456–460.

[20] Lamartina, C., & Berjano, P. (2014). Classification of sagittal imbalance based on spinal alignment and compensatory mechanisms. *European Spine Journal*, *23*, 1177–1189.

[21] Laxhammar, R., & Falkman, G. (2010). Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In *Proceedings of the first international workshop on novel data stream pattern mining techniques* (pp. 47–55).

[22] Li, Y., Zhang, L., Xu, Y., Yao, Y., Lau, R. Y. K., & Wu, Y. (2017). Enhancing binary classification by modeling uncertain boundary in three-way decisions. *IEEE Transactions on Knowledge and Data Engineering*, *29*, 1438–1451.

[23] Liu, D. (2021). The effectiveness of three-way classification with interpretable perspective. *Information Sciences*, *567*, 237–255.

[24] Liu, D., Li, T., & Liang, D. (2014). Incorporating logistic regression to decision-theoretic rough sets for classifications. *International Journal of Approximate Reasoning*, *55*, 197–210.

[25] Liu, D., Yang, X., & Li, T. (2020). Three-way decisions: beyond rough sets and granular computing. *International Journal of Machine Learning and Cybernetics*, *11*, 989–1002.

[26] Liu, J., Li, H., Zhou, X., Huang, B., & Wang, T. (2019). An optimization-based formulation for three-way decisions. *Information Sciences*, *495*, 185–214.

[27] Liu, Z.-G., Pan, Q., Dezert, J., & Mercier, G. (2014). Credal classification rule for uncertain data based on belief functions. *Pattern Recognition*, *47*, 2532–2541.

[28] Makili, L. E., Sánchez, J. A. V., & Dormido-Canto, S. (2012). Active learning using conformal predictors: application to image classification. *Fusion Science and Technology*, *62*, 347–355.

[29] Min, F., Liu, F.-L., Wen, L.-Y. et al. (2019). Tri-partition cost-sensitive active learning through knn. *Soft Computing*, *23*, 1557–1572.

[30] Nouretdinov, I., Gammerman, J., Fontana, M., & Rehal, D. (2020). Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection. *Neurocomputing*, *397*, 279–291.

[31] Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, *114*, 223–234.

[32] Savchenko, A. V. (2021). Fast inference in convolutional neural networks based on sequential three-way decisions. *Information Sciences*, *560*, 370–385.

[33] Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, *9*, 371–421.

[34] Vovk, V. (2002). On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.* (pp. 187–196). IEEE.

[35] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

[36] Vovk, V., Nouretdinov, I., Fedorova, V., Petej, I., & Gammerman, A. (2017). Criteria of efficiency for set-valued classification. *Annals of Mathematics and Artificial Intelligence*, *81*, 21–46.

[37] Wechsler, H. et al. (2015). Cyberspace security using adversarial learning and conformal prediction. *Intelligent Information Management*, *7*, 195.

[38] Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information sciences*, *507*, 772–794.

[39] Xu, Y., Tang, J., & Wang, X. (2020). Three sequential multi-class three-way decision models. *Information Sciences*, *537*, 62–90.

[40] Yang, B., & Li, J. (2020). Complex network analysis of three-way decision researches. *International Journal of Machine Learning and Cybernetics*, (pp. 1–15).

[41] Yang, G., Destercke, S., & Masson, M.-H. (2017). Cautious classification with nested dichotomies and imprecise probabilities. *Soft Computing*, *21*, 7447–7462.

[42] Yao, Y. (2009). Three-way decision: an interpretation of rules in rough set theory. In *International Conference on Rough Sets and Knowledge Technology* (pp. 642–649). Springer.

[43] Yao, Y. (2012). An outline of a theory of three-way decisions. In *International Conference on Rough Sets and Current Trends in Computing* (pp. 1–17). Springer.

[44] Yao, Y. (2019). Tri-level thinking: models of three-way decision. *International Journal of Machine Learning and Cybernetics*, (pp. 1–13).

[45] Yao, Y. (2021). The geometry of three-way decision. *Applied Intelligence*, (pp. 1–28).

[46] Yao, Y. (2021). Set-theoretic models of three-way decision. *Granular Computing*, *6*, 133–148.

[47] Yue, X., Chen, Y., Yuan, B., & Lv, Y. (2021). Three-way image classification with evidential deep convolutional neural networks. *Cognitive Computation*, (pp. 1–13).

[48] Zhan, X., Wang, Z., Yang, M., Luo, Z., Wang, Y., & Li, G. (2020). An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction. *Measurement*, (p. 107588).

[49] Zhang, Y., Miao, D., Wang, J., & Zhang, Z. (2019). A cost-sensitive three-way combination technique for ensemble learning in sentiment classification. *International Journal of Approximate Reasoning*, *105*, 85 – 97.

[50] Zhou, J., Pedrycz, W., Gao, C., Lai, Z., & Yue, X. (2021). Principles for constructing three-way approximations of fuzzy sets: A comparative evaluation based on unsupervised learning. *Fuzzy Sets and Systems*, *413*, 74–98.