Machine Learning (2024) 113:6645–6661
https://doi.org/10.1007/s10994-024-06571-6

Conformal predictions for probabilistically robust scalable machine learning classification

Alberto Carlevaro[1,4] · **Teodoro Alamo**[3] · **Fabrizio Dabbene**[2] · **Maurizio Mongelli**[1]

Received: 30 December 2023 / Revised: 23 February 2024 / Accepted: 16 May 2024 /
Published online: 9 July 2024

Abstract
Conformal predictions make it possible to define reliable and robust learning algorithms. But they are essentially a method for evaluating whether an algorithm is good enough to be used in practice. To define a reliable learning framework for classification from the very beginning of its design, the concept of scalable classifier was introduced to generalize the concept of classical classifier by linking it to statistical order theory and probabilistic learning theory. In this paper, we analyze the similarities between scalable classifiers and conformal predictions by introducing a new definition of a score function and defining a special set of input variables, the conformal safety set, which can identify patterns in the input space that satisfy the error coverage guarantee, i.e., that the probability of observing the wrong (possibly unsafe) label for points belonging to this set is bounded by a predefined ϵ error level. We demonstrate the practical implications of this framework through an application in cybersecurity for identifying DNS tunneling attacks. Our work contributes to the development of probabilistically robust and reliable machine learning models.

Keywords Conformal predictions · Scalable classifiers · Confidence bounds · Robust AI

Editors: Henrik Boström, Eyke Hüllermeier, Ulf Johansson, Khuong An Nguyen, Aaditya Ramdas.

✉ Alberto Carlevaro
alberto.carlevaro@ieiit.cnr.it

Teodoro Alamo
talamo@us.es

Fabrizio Dabbene
fabrizio.dabbene@cnr.it

Maurizio Mongelli
maurizio.mongelli@ieiit.cnr.it

[1] Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, CNR, Corso Ferdinando Maria Perrone, 24, 16152 Genoa, Italy

[2] Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, CNR, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy

[3] Departamento de Ingenierìa de Sistemas y Automàtica, Universidad de Sevilla, Escuela Superior de Ingenieros, Camino de los Descubrimientos, 41092 Seville, Spain

[4] Funded Research Department, Aitek SpA, Via della Crocetta 15, 16122 Genoa, Italy

1 Introduction

1.1 Context

Conformal predictions (CPs) (Shafer and Vovk 2008) are gaining increasing importance in machine learning (ML) since they validate algorithms in terms of confidence of the prediction. Although it is a fairly recent field of study, there has been an astonishing production of scholarly papers, from the definition of new score functions to different methodologies for constructing conformal sets and, of course, a wide variety of applications. In fact, the ferment of scientific research in this field is so active that even the father of this theory, V. Vovk,[1], continues to actively contribute to the improvement of its knowledge, as in the case of Vovk et al. (2017) where he and his colleagues investigate the concept of validity under nonparametric hypotheses or the innovative introduction of Venn predictors as in Vovk et al. (2022). We refer the reader to the surveys (Angelopoulos and Bates 2023; Fontana et al. 2023; Toccaceli 2022) that largely cover all recent publications and discussions on uncertainty quantification (UQ) through CP for machine learning models.

Under canonical CP theory, the definition of a score function is very peculiar to either the classifier or the application at hand. For example, Forreryd et al. (2018) defines a special conformity measure (corresponding to a score function), based on the residual between the calibration points and the classification hyperplane of a SVM model. Other example, always SVM-based, can be found in Forreryd et al. (2018), Shafer and Vovk (2008) and Balasubramanian et al. (2009), where different definitions of score function (or conformity/non-conformity measure) are given. One of the strengths of our approach, as will become clear later, is the unique definition of such a score function, which, given *any* classifier, allows the conformal prediction framework to be applied in the most natural way. The work in Narteni et al. (2023) defines a score function for rule-based models. The softmax function is used as score function in most image classification problems as in Angelopoulos et al. (2020); Park et al. (2019); Andéol et al. (2024), and many other examples can be provided (see, e.g., the above cited surveys). As a matter of the fact, those definitions come after the setting of the classifier and do not outline a common methodology.

1.2 Contribution

By focusing on binary classification, our goal is to introduce to the CP community a way to link ML classifiers with a natural definition of score function that embeds the conformal guarantee by construction.

We exploit the concept of *scalable classifiers* $f_{\theta}(\boldsymbol{x}, \rho)$ (Sect. 2.1) introduced in Carlevaro et al. (2023) to develop a new class of score functions that rely on the geometry of the problem and that are naturally built from the classifier itself, by inheriting its properties (Sect. 3.1). This allows CP theory to derive the relationship between the input space and the conformity guarantee explicitly. By introducing the new concept of *conformal safety region*, we provide an analytical form of the specific subsets of the input space in which

[1] V. Vovk. is the author of the groundbreaking book *Algorithmic Learning in a Random World* Vovk et al. (2005), which is the foundation of the theory of conformal prediction. His scientific production is still at the top of research in the field, and a constantly updated list of papers on CP by Vovk and his colleagues can be found here http://www.alrw.net/.

marginal coverage guarantees on prediction (Sect. 3.2) can be ensured. Controlling the misclassification rate (either false positives or false negatives) naturally follows from eliciting the following quantities: the confidence level given by the conformal framework, the binary output $y \in \{+1, -1\}$, the confidence error $\varepsilon \in (0, 1)$, as well as the new notion of conformal safety set $\mathcal{S}_\varepsilon$ that satisfies

$$\Pr\{y = -1 \text{ and } \boldsymbol{x} \in \mathcal{S}_\varepsilon\} \leq \varepsilon.$$

In short, the paper defines a methodology in which the optimal shape of a classifier is derived, where the optimality criterion is embedded in the classifier by the conformal guarantee. The proposed methodology thus places itself in the recent and as yet unexplored field of set-value classification (Chzhen et al. 2021), a broad theory that studies predictors that have both good prediction properties and specific performance requirements, two points that underlie the proposed research.

The remainder of the article is organized by providing a brief recall of the concepts of scalable classifiers and conformal prediction and then delving into the details of the definition of scalable score function and conformal safety region. The whole procedure is then validated on an application use case related to cyber-security for identifying DNS tunneling attacks (Sect. 4).

2 Background: scalable classifiers and conformal prediction

The background of the theory we would like to propose in this research refers to a new interpretation of classical classification algorithms, scalable classifiers, and another rather new theory on trustworthy AI, called conformal prediction. Both of these techniques belong to the field of reliable AI, searching for the definition of models, procedures or bounds that can make a learning algorithm probabilistically robust and reliable.

2.1 Scalable classifiers

Given an input space $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}^+$, and an output space $\mathcal{Y} = \{-1, +1\}$, scalable classifiers (SCs) were introduced in Carlevaro et al. (2023) as a family of (binary) classifiers parameterized by a scale factor $\rho \in \mathbb{R}$

$$\phi_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) \doteq \begin{cases} +1 & \text{if } f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) < 0, \\ -1 & \text{otherwise.} \end{cases} \tag{1}$$

where the function $f_{\boldsymbol{\theta}} : \mathcal{X} \times \mathbb{R} \longrightarrow \mathbb{R}$ is the so-called *classifier predictor* and the notation with subscript $\boldsymbol{\theta}$ refers to the fact that the classifier also depends on a set of *hyperparameters* $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{n_\theta}]^\top$ to be set in the model (e.g. different choices of kernel, regularization parameters, etc.). To give a meaningful interpretation of this classifier, we refer to the class $+1$ as a "safe" situation we want to target and the other class with -1 as an "unsafe" situation. Some examples might be differentiating between a patient's condition in developing or not developing a certain disease (Lenatti et al. 2022), or understanding what input parameters lead an autonomous car to a collision or non-collision (Carlevaro et al. 2022), among many other applications.

SCs rely on the main assumption that for every $\boldsymbol{x} \in \mathcal{X}$, $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho)$ is continuous and monotonically increasing in ρ, and that $\lim_{\rho \to -\infty} f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) < 0 < \lim_{\rho \to \infty} f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho)$, [Carlevaro et al. (2023)

Assumption 1]. These assumptions imply that, there exists a unique solution $\bar{\rho}(\boldsymbol{x})$ to the equation

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) = 0. \tag{2}$$

The proof of this claim is available in [Carlevaro et al. (2023) Property 2]. In words, a scalable classifier is a classifier that satisfies some crucial properties: *i*) given $\boldsymbol{x}$, there is always a value of ρ, denoted $\bar{\rho}(\boldsymbol{x})$, that establishes the border between the two classes, *ii*) the increase of ρ forces the classifier to predict the -1 class and *iii*) the target $+1$ class of a given feature vector $\boldsymbol{x}$ is maintained for a decrease of ρ. Moreover, [Carlevaro et al. (2023) Property 3] shows how any standard binary classifier can be made scalable by simply including the scaling parameter ρ in an additive way with the classifier predictor. That is, given the function $\hat{f} : \mathcal{X} \longrightarrow \mathbb{R}$ and its corresponding classifier $\hat{\phi}(\boldsymbol{x})$then the function $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) = \hat{f}(\boldsymbol{x}) + \rho$ provides the scalable classifier $\phi_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho)$ Thus, examples of classifiers that can be rendered scalable are support vector machine (SVM), support vector data description (SVDD), logistic regression (LR) but also artificial neural networks. More in detail, given a learning set

$$\mathcal{Z}_\ell \doteq \left\{ \left(\boldsymbol{x}_i, y_i\right) \right\}_{i=1}^{n} \subseteq \mathcal{X} \times \{-1, +1\}$$

containing observed feature points and corresponding labels, $\boldsymbol{z}_i = \left(\boldsymbol{x}_i, y_i\right)$, and assuming that $\varphi : \mathcal{X} \longrightarrow \mathcal{V}$ represents a *feature map* (where $\mathcal{V}$ is an inner product space) that allows to exploit kernels, some examples of scalable classifier predictors are:

- SVM: $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) = \boldsymbol{w}^\top \varphi(\boldsymbol{x}) - b + \rho$,
- SVDD: $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) = \|\varphi(\boldsymbol{x}) - \boldsymbol{w}\|^2 - R^2 + \rho$,
- LR: $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) = \frac{1}{2} - \frac{1}{1 + e^{(\boldsymbol{w}^\top \varphi(\boldsymbol{x}) - b) + \rho}}$,

where the classifier elements $\boldsymbol{w}, b$ and R can be obtained as solution of proper optimization problems. The interested reader can refer to [Carlevaro et al. (2023) Section II c] for a more in depth discussion.

Different values of the parameter ρ correspond to different classifiers that can be considered as the level sets of the classifier predictor with respect to ρ. In particular, since we are interested in predicting the class $+1$ which, we recall, encodes a safety condition, we introduce

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

that is the set of points $\boldsymbol{x} \in \mathcal{X}$ predicted as safe by the classifier with the specific choice of ρ, i.e. the *safety region* of the classifier $f_{\boldsymbol{\theta}}$ for given ρ. It is easy to see that these sets are decreasingly nested with respect to ρ, i.e.

$$\rho_1 > \rho_2 \Longrightarrow \mathcal{S}(\rho_1) \subset \mathcal{S}(\rho_2).$$

2.2 Conformal prediction

Conformal Prediction is a relatively recent framework developed in the late nineties by V. Vovk. We refer the reader to the surveys (Angelopoulos and Bates 2023; Shafer and Vovk 2008; Fontana et al. 2023) for a gentle introduction to this methodology. CP is mainly an

a-posteriori verification of the designed classifier, and in practice returns a measure of its "conformity" to the calibration data. We consider the particular implementation of CP discussed in Angelopoulos and Bates (2023), relative to the so-called "inductive" CP: in this setting, starting from a given predictor and a *calibration* set, CP allows to construct a new predictor with given probabilistic guarantees.

To this end, the first key step is the definition of a *score function* $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$. Given a point $\boldsymbol{x} \in \mathcal{X}$ and a *candidate* label $\hat{y} \in \{-1, 1\}$, the score function returns a score $s(\boldsymbol{x}, \hat{y})$. Larger scores encode worse agreement between point $\boldsymbol{x}$ and the candidate label $\hat{y}$. Then, assume to have available a second set of n_c observations, usually referred to as *calibration* set, defined as follows

$$\mathcal{Z}_c \doteq \left\{(\boldsymbol{x}_i, y_i)\right\}_{i=1}^{n_c} = \mathcal{X}_c \times \mathcal{Y}_c \subseteq \mathcal{X} \times \mathcal{Y}, \tag{4}$$

that are pairs of points $\boldsymbol{x}$ with their corresponding true label y.

We assume that the observations $\boldsymbol{x}_i \in \mathcal{X}_c$ come from the same distribution Pr of the observations in the test set $\mathcal{Z}_{ts} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_{ts}} = \mathcal{X}_{ts} \times \mathcal{Y}_{ts} \subseteq \mathcal{X} \times \mathcal{Y}$. Additionally, CP requires that the data are *exchangeable*, which is a weaker assumption than that of i.i.d.. Exchangeability means that the joint distribution of the data $z_1, z_2, \ldots, z_n$ is unchanged under permutations:

$$(z_1, z_2, \ldots, z_n) \sim (z_{\sigma(1)}, z_{\sigma(2)}, \ldots, z_{\sigma(n)}), \text{ for all permutations } \sigma.$$

Then, given a user-chosen confidence rate $(1 - \varepsilon) \in (0, 1)$, a *conformal set* $C_\varepsilon(\boldsymbol{x})$ is defined as the set of candidate labels whose score function is lower than the $(\lceil (n_c + 1)(1 - \varepsilon) \rceil / n_c)$ -quantile, denoted as s_ε, computed on the $s_1, \ldots, s_{n_c}$ calibration scores. That is, to every point $\boldsymbol{x}$, CP associates a set of "plausible labels"

$$C_\varepsilon(\boldsymbol{x}) = \{\ \hat{y} \in \{-1, 1\}\ \ :\ \ s(\boldsymbol{x}, \hat{y}) \leq s_\varepsilon\ \}.$$

The usefulness of the conformal set is that, according with Vovk et al. (1999), $C_\varepsilon(\boldsymbol{x})$ possesses the so-called *marginal conformal coverage guarantee* property, that is, given any (unseen before) observation $(\tilde{\boldsymbol{x}}, \tilde{y})$, the following holds

$$\Pr\left\{\tilde{y} \in C_\varepsilon(\tilde{\boldsymbol{x}})\right\} \geq 1 - \varepsilon. \tag{5}$$

In other words, the true label $\tilde{y}$ belongs with high probability – at least $(1 - \varepsilon)$ – to the conformal set.

3 Notion of score function for scalable classifiers and conformal safety sets

In this section we introduce two concepts: *i*) a definition of score function for scalable classifiers (see Definition 1) and *ii*) the notion of *conformal safety region* (see Definition 2).

3.1 Natural definition of score function for scalable classifiers

In this paragraph, we show how scalable classifiers allow for a natural definition of the score function, based on their own classifier predictor.

Definition 1 [Score Function for Scalable Classifier] Given a scalable classifier $\phi_\theta(\boldsymbol{x}, \rho)$ with classifier predictor $f_\theta(\boldsymbol{x}, \rho)$, given a point $\boldsymbol{x}$ and an associated candidate label $\hat{y}$, the score function associated to the scalable classifier is defined as

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

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

We notice that, since f_θ is a SC predictor, the existence and uniqueness of such $\bar{\rho}(\boldsymbol{x})$ is guaranteed (Sect. 2.1) and consequently s is well defined.

In practice, the score function evaluates how much it is necessary to vary the original classification boundary $f_\theta(\boldsymbol{x}, 0)$ such that the point $\boldsymbol{x}$ falls on the classification boundary of the new classifier $f_\theta(\boldsymbol{x}, \bar{\rho}(\boldsymbol{x}))$, starting from class $\hat{y}$. Alternatively, it is possible to think of the score function as a measure of the "difficulty" of making the classifier predict a certain class: very large values for $\bar{\rho}(\boldsymbol{x})$ imply that it is difficult to render $f_\theta(\boldsymbol{x}, \bar{\rho}(\boldsymbol{x}))$ positive, or equivalently that the class -1 is not conformal (thus, when $\hat{y} = -1$, the score function is $_{\bar{\rho}(\boldsymbol{x}) = -\hat{y}\ \bar{\rho}(\boldsymbol{x})}$). Very negative values of $\bar{\rho}(\boldsymbol{x})$ imply that it is difficult to render the output equal to $+1$, thus the score function is in this case $-_{\bar{\rho}(\boldsymbol{x}) = -\hat{y}}\ \bar{\rho}(\boldsymbol{x})$.

Example 1 Scalable SVDD is the most straightforward example of correctly understanding such a definition for score function. In this case the score function takes this form

$$s(\boldsymbol{x}, \hat{y}) = -\hat{y}\left(R^2 - \|\boldsymbol{w} - \boldsymbol{\varphi}(\boldsymbol{x})\|^2\right).$$

This represents exactly the quantity that needs to be removed ($\hat{y} = +1$ for the point inside the sphere, $\|\boldsymbol{w} - \boldsymbol{\varphi}(\boldsymbol{x})\|^2 - R^2 < 0$) or added ($\hat{y} = -1$ for the point outside the sphere, $R^2 - \|\boldsymbol{w} - \boldsymbol{\varphi}(\boldsymbol{x})\|^2 > 0$) to the radius such that $\boldsymbol{x}$ falls on the boundary of the classifier.

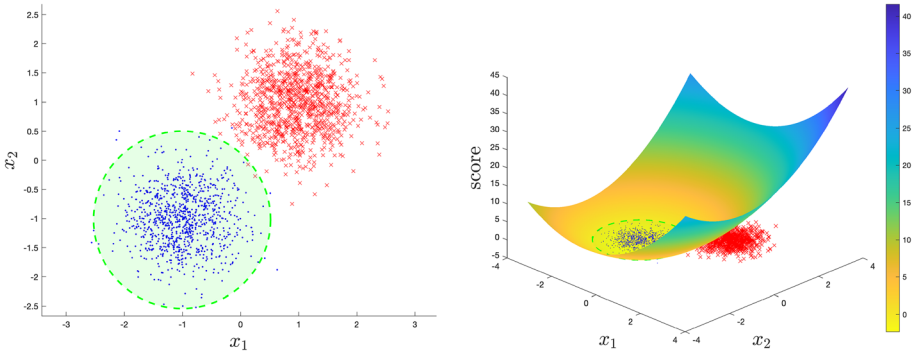
For example, consider two classes of points, "safe" ($+1$, in blue in the following figure) and "unsafe" (-1, in red in the following figure) sampled from two two-dimensional Gaussian distributions with respectively means and covariance matrices

$$\mu_{\mathrm{S}} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \sigma_{\mathrm{S}} = \frac{1}{2}\mathrm{I}\ ;\ \mu_{\mathrm{U}} = \begin{bmatrix} +1 \\ +1 \end{bmatrix}, \sigma_{\mathrm{U}} = \frac{1}{2}\mathrm{I}$$

where I is the identity matrix. We trained a linear SVDD classifier (Fig. 1a) and plotted the respectively score function (Fig. 1b). Exactly the behavior described above can be observed: the score function associates values according to the geometry provided by the classifier. In this case, points belonging to the boundary of the circumference have score function values of 0 (dashed green line) and negative or positive depending on whether the point is inside or outside the circumference. It is worth noting that the classifier can be interpreted as a level set of the score function, and this interpretation is crucial as will become clear in the following.

3.2 Conformal safety regions

Classical CPs define subsets of the output space that satisfy the probabilistic marginal coverage constraint, but it is equally important to understand the relationship between the input space and the conformal sets. In other words, it would be meaningful to define



(a) Linear SVDD classifier (b) Linear SVDD score function

Fig. 1 Relationship between the SVDD classifier and the corresponding score function: the absolute value of the score function assigns to a sample its distance to the circumference boundary. The color bar on the right helps to understand the behavior of the score function: darker colors indicate regions with less conformity with the target class, warmer the opposite. The zero value of the score function is obtained exactly on the boundary

regions in the input space classified on the basis of the conformal set of their samples to identify for which inputs the classifier is most reliable in making a certain prediction. For example, one should be interested in finding the region of classification uncertainty ($C_\varepsilon(\boldsymbol{x}) = \{-1, +1\}$) or the region in which the conformal classifier predicts a specific label ($C_\varepsilon(\boldsymbol{x}) = \{+1\}$ or $C_\varepsilon(\boldsymbol{x}) = \{-1\}$) or in which it has no guess at all ($C_\varepsilon(\boldsymbol{x}) = \varnothing$).

In particular, since the goal is to find the input values that bring the classification to a "safe" situation (i.e., in our notation, $y = +1$) with a certain level of confidence, we introduce the concept of *conformal safety region*.

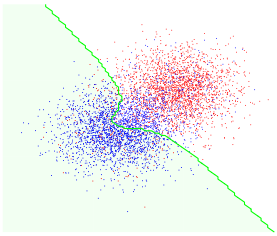
Definition 2 [Conformal Safety Region] Consider a calibration set $\mathcal{Z}_c = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_c}$ from the same data distribution of the test set Z_{ts}. Given a level of error $\varepsilon \in (0, 1)$, a score function $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$, and its corresponding $(\lceil (n_c + 1)(1 - \varepsilon) \rceil / n_c)$-quantile s_ε computed on the calibration set, the conformal safety region (CSR) of level ε is defined as follows

$$\Sigma_\varepsilon = \{\, \boldsymbol{x} \in \mathcal{X} \;:\; s(\boldsymbol{x}, +1) \leq s_\varepsilon,\; s(\boldsymbol{x}, -1) > s_\varepsilon \,\}. \tag{6}$$

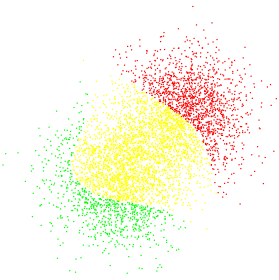
In words, a conformal safety region (CSR) is the subset of the input space where the conformal set is composed by only safe labels, $C_\varepsilon(\boldsymbol{x}) = \{+1\}$, which can be inferred directly from the definition. Note that the above definition is independent on the choice of the score function s. What we will prove in the next is that using the score function defined for SCs (Definition 1) it is possible to give an analytical form to Σ_ε.

Example 2 Consider the same configuration as in Example 1 but with covariance matrices $\sigma_S = \sigma_U = I$ and with a probability to sample an outlier for each class $p_O = 0.1$. Consider the LR classifier and its corresponding score function

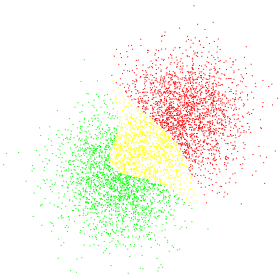
$$s(\boldsymbol{x}, \hat{y}) = -\hat{y}(b - \boldsymbol{w}^\top \varphi(\boldsymbol{x})),$$



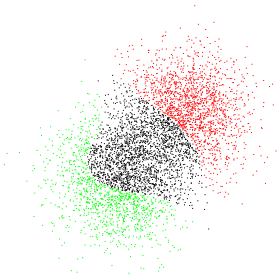
(a) LR classifier



(b) $\varepsilon = 0.05$



(c) $\varepsilon = 0.1$



(d) $\varepsilon = 0.5$

Fig. 2 Scatter-plots of the conformal set varying ε for cubic LR. Green and red points correspond to singleton conformal set ($C_\varepsilon(\boldsymbol{x}) = \{+1\}$ and $C_\varepsilon(\boldsymbol{x}) = \{-1\}$ respectively) yellow points to double predictions ($C_\varepsilon(\boldsymbol{x}) = \{+1, -1\}$) and black points to empty prediction ($C_\varepsilon(\boldsymbol{x}) = \varnothing$)

which is the same of the SVM since the solution of the equation $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho) = 0$ is for both $\bar{\rho}(\boldsymbol{x}) = b - \boldsymbol{w}^\top \varphi(\boldsymbol{x})$.

We trained on a training set composed by 3000 samples the LR classifier with cubic polynomial kernel (Fig. 2a) and then we computed the score values on a calibration set of 5000 samples. We computed the quantiles varying ε (0.05, 0.1 and 0.5) and we plotted (on a test set of 10000 samples) the scatter of the points according to the conformal set. Green points belong to the CSR Σ_ε and it is easily understandable that the smaller ε is, the smaller Σ_ε. This behavior is in line with CP theory, since small values of ε mean that the conformal prediction must be very precise, and this is achievable only if the classifier itself is "very confident" of assigning the true label to a sample. Also, it should be noted that the smaller ε is, the larger the region of uncertainty for the conformal prediction ($C_\varepsilon(\boldsymbol{x}) = \{-1, +1\}$, in yellow in Figs. 2a, 2b). Again, since for small ε high levels of marginal coverage must be satisfied, conformal prediction tends to give both labels to a point when it is uncertain. Contrarily, for high values of ε (Fig. 2c) the conformal sets for uncertain points tend to be empty (in black) because the score is too high and no output meets the specifications to belong to C_ϵ. Finally, it is worth noting that the regions into which the points scatter have a well-defined shape: as introduced in Example 1 and as will become clear in the next section, these regions correspond to level sets of the score function.

3.3 Analytical form of conformal safety regions for scalable classifiers

The definition of score we gave for SCs in Definition 1 identifies a particular value of the scalable parameter, which is the one corresponding to the quantile s_ε, that we can define formally as

$$\rho_\varepsilon \doteq |s_\varepsilon|. \tag{7}$$

To this value, we can associate a level set $\mathcal{S}(\rho_\varepsilon)$ defined as in (3), i.e. the ρ_ε-safe set

$$\mathcal{S}_\varepsilon = \{\, \boldsymbol{x} \in \mathcal{X} \;:\; f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho_\varepsilon) < 0 \,\}. \tag{8}$$

We can prove that non-trivial relationships link $\mathcal{S}_\varepsilon$ to the CSR Σ_ε. But before, let us split Σ_ε in two contribution:

$$\Sigma_\varepsilon = \Sigma_\varepsilon^a \cup \Sigma_\varepsilon^b, \tag{9}$$

where

$$\Sigma_\varepsilon^a = \{\, \boldsymbol{x} \in \mathcal{X} \;:\; s(\boldsymbol{x}, +1) < s_\varepsilon, s(\boldsymbol{x}, -1) > s_\varepsilon \,\}, \tag{10}$$

and

$$\Sigma_\varepsilon^b = \{\, \boldsymbol{x} \in \mathcal{X} \;:\; s(\boldsymbol{x}, +1) = s_\varepsilon, s(\boldsymbol{x}, -1) > s_\varepsilon \,\}. \tag{11}$$

The relationship between $\mathcal{S}_\varepsilon$ and Σ_ε is explored in the following results that provide as final and major contribution the fact that $\mathcal{S}_\varepsilon \subseteq \Sigma_\varepsilon$.

Proposition 1

$$\mathcal{S}_\varepsilon = \Sigma_\varepsilon^a \subseteq \Sigma_\varepsilon.$$

Proof

$$\begin{aligned}
\boldsymbol{x} \in \mathcal{S}_\varepsilon &\iff f_{\boldsymbol{\theta}}(\boldsymbol{x}, |s_\varepsilon|) < 0, \\
&\iff f_{\boldsymbol{\theta}}(\boldsymbol{x}, |s_\varepsilon|) < f_{\boldsymbol{\theta}}(\boldsymbol{x}, \bar{\rho}(\boldsymbol{x})), \\
&\iff |s_\varepsilon| < \bar{\rho}(\boldsymbol{x}), \\
&\iff -s_\varepsilon < \bar{\rho}(\boldsymbol{x}) \text{ and } s_\varepsilon < \bar{\rho}(\boldsymbol{x}), \\
&\iff -s_\varepsilon < -s(\boldsymbol{x}, +1) \text{ and } s_\varepsilon < s(\boldsymbol{x}, -1), \\
&\iff s(\boldsymbol{x}, +1) < s_\varepsilon \text{ and } s(\boldsymbol{x}, -1) > s_\varepsilon, \\
&\iff \boldsymbol{x} \in \Sigma_\varepsilon^a \subseteq \Sigma_\varepsilon.
\end{aligned}$$

□

Corollary 1

$$\mathcal{S}_\varepsilon = \Sigma_\varepsilon \text{only if} \Sigma_\varepsilon^b = \varnothing.$$

Proof Trivial, from

$$\Sigma_\varepsilon = \Sigma_\varepsilon^a \cup \Sigma_\varepsilon^b = \mathcal{S}_\varepsilon \cup \Sigma_\varepsilon^b.$$

□

Proposition 2

$$\Sigma_\varepsilon^b \neq \varnothing \Longrightarrow s_\varepsilon > 0.$$

Proof

$$\boldsymbol{x} \in \Sigma_\varepsilon^b \iff s(\boldsymbol{x}, +1) = s_\varepsilon \text{ and } s(\boldsymbol{x}, -1) < s_\varepsilon, \tag{12}$$

$$\iff -\bar{\rho}(\boldsymbol{x}) = s_\varepsilon \text{ and } \bar{\rho}(\boldsymbol{x}) < s_\varepsilon, \tag{13}$$

$$\iff -s_\varepsilon < s_\varepsilon, \tag{14}$$

$$\iff s_\varepsilon > 0. \tag{15}$$

□

We can then summarize all these information in a single theorem that defines the "analytical form" of the CSR, i.e. that it is possible to express Σ_ε in terms of a single scalar parameter.

Theorem 3 [*Analytical Representation of the Conformal Safety Region via Scalable Classifiers*] *Consider the classifier* (1) *and suppose that* [Carlevaro et al. (2023) *Assumption* 1] *holds and that* $\Pr\{\boldsymbol{x} \in \mathcal{X}\} = 1$. *Consider then a calibration set* $\mathcal{Z}_c = \left\{(\boldsymbol{x}_i, y_i)\right\}_{i=1}^{n_c}$ (n_c *exchangeable samples*), *a level of error* $\varepsilon \in (0,1)$, *a score function* $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ *as in Definition* 1 *with* $\lceil (n_c+1)(1-\varepsilon) \rceil / n_c$*-quantile* s_ε *computed on the calibration set. Define the conformal scaling of level* ε *as follows*

$$\rho_\varepsilon = |s_\varepsilon|, \tag{16}$$

and define the corresponding ρ_ε*-safe set*

$$\mathcal{S}_\varepsilon = \{\, \boldsymbol{x} \in \mathcal{X} \,:\, f_{\boldsymbol{\theta}}(\boldsymbol{x}, \rho_\varepsilon) < 0 \,\}. \tag{17}$$

Then, given the conformal safety region of level ε, Σ_ε, *we have*

i) $\mathcal{S}_\varepsilon \subseteq \Sigma_\varepsilon$.
ii) $\mathcal{S}_\varepsilon = \Sigma_\varepsilon$ *if* $s_\varepsilon \leq 0$.

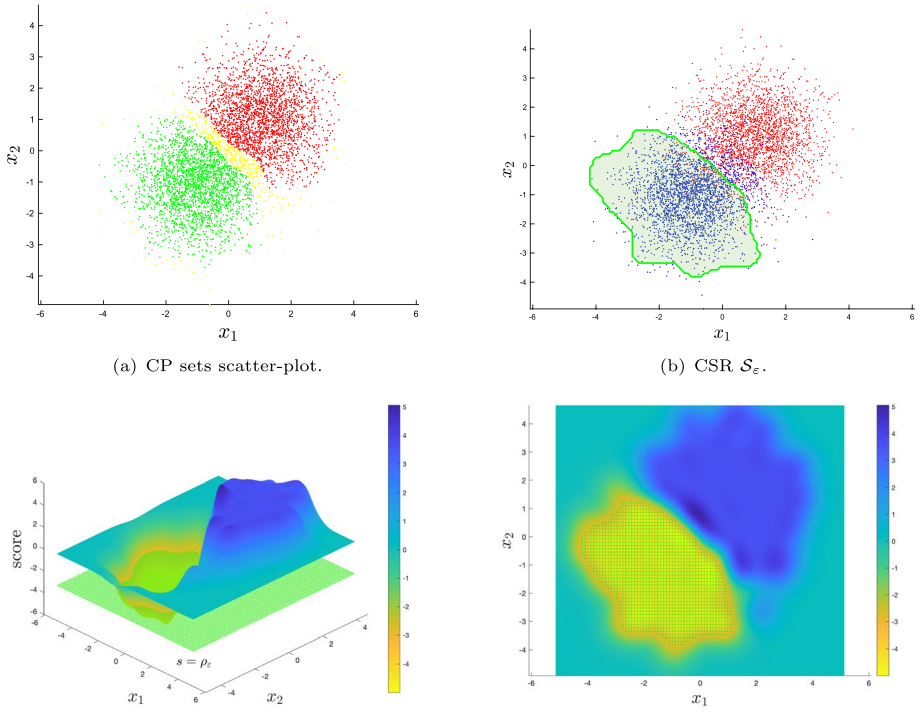
that is, $\mathcal{S}_\varepsilon$ is a CSR.

Proof Proof follows directly from Propositions (1) and (2) and Corollary (1). □

In its classical definition, conformal prediction is a local property, that is, the conformal coverage guarantee is valid only punctually. However, conformal labels map each point in a subset of the input space, depending on the size of the respective conformal set. Theorem 3 then provides a new classifier that maps the samples contained in $\mathcal{S}_\varepsilon$ to the target class $+1$. Once ρ_ε has been computed, it is then possible to write

$$\mathcal{S}_\varepsilon = \phi_{\boldsymbol{\theta}}(\cdot, \rho_\varepsilon)^{-1}(y = +1),$$

identifying a unique relationship between the target class of the classification and the CSR.



(a) CP sets scatter-plot. (b) CSR $\mathcal{S}_\varepsilon$.

(c) Gaussian SVM score function. (d) "score" = 0 plane projection.

Fig. 3 CSR computed with a Gaussian SVM at $\varepsilon = 0.05$. Scattered CSR Σ_ε, **a**, coincides with the analytical CSR $\mathcal{S}_\varepsilon$, **b** that coincides with the level set $z = \rho_\varepsilon$ of the score function, (**c**). **d** is the planar representation on $x_1 - x_2$ plane of the score function

Example 3 In the same configuration as in Example 2, we trained a Gaussian SVM and calculated the score values on the calibration set. Figure 3 shows exactly what Theorem 3 claims: CSRs are level sets of the score function that correspond to a specific quantile and thus to a specific confidence level. Specifically, in this example it is shown the CSR at level of confidence $1 - 0.05 = 0.95$ that results in a quantile equal to 2.8113 and corresponding conformal scaling $\rho_{0.05} = -2.8113$. The hyperplane "score" $= -2.8113$ exactly cuts the score function at the level set corresponding to the CSR.

Remark 1 [On the Usefulness of Conformal Safety Regions] The introduction of the concept of CSR brings inevitably to understand how this instrument can be useful in practice. First of all it allows to identify reliable prediction regions and quantify uncertainty: in decision making problems where a certain amount of confidence in the prediction is required (like for example in medical applications) CSRs can suggest the best set of input features that guides the predictions reliably, minimizing the presence of misclassification samples. Moreover CSRs provide an interpretable way to understand the model's behavior in different regions of the input space. This can be useful for the model explanation and for possible improvements and corrections to the model. Finally, CSRs are very "regulatory compliant": in applications with regulatory requirements, CSRs ensure compliance by providing a clear understanding of where model's predictions are reliable.

In addition, CSRs can provide strong information about the prediction of points belonging to them. Indeed, it can be proved that the number of false positives is limited by the ε error.

Theorem 4 *Consider the classifier* (1) *and the corresponding CSR developed as in Theorem* 3 *with a level of error* $\varepsilon \in (0, 1)$. *Then, it can be stated that*

$$\Pr\left\{y = -1 \text{ and } \boldsymbol{x} \in \mathcal{S}_\varepsilon\right\} \leq \varepsilon. \tag{18}$$

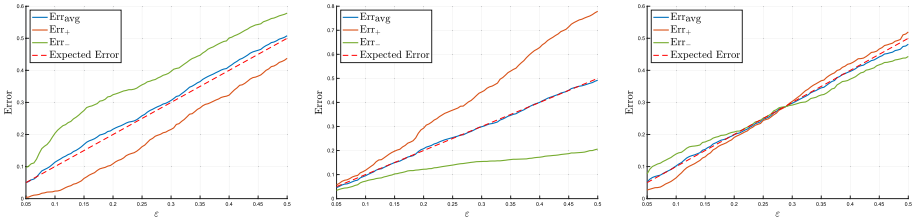
Proof Since $\mathcal{S}_\varepsilon \subseteq \Sigma_\varepsilon$:

$$\begin{aligned}
\Pr\{y = -1 \text{ and } \boldsymbol{x} \in \mathcal{S}_\varepsilon\} &\leq \Pr\{y = -1 \text{ and } \boldsymbol{x} \in \Sigma_\varepsilon\} \\
&= \Pr\{y = -1 \text{ and } \boldsymbol{x} \in \{\boldsymbol{x} : C_\varepsilon(\boldsymbol{x}) = \{+1\}\}\} \\
&\leq \Pr\{y = -1 \text{ and } \boldsymbol{x} \in \{\boldsymbol{x} : -1 \notin C_\varepsilon(\boldsymbol{x})\}\} \\
&= \Pr\{y = -1 \text{ and } y \notin C_\varepsilon(\boldsymbol{x})\} \\
&\leq \Pr\{y \notin C_\varepsilon(\boldsymbol{x})\} \leq \varepsilon,
\end{aligned}$$

where the last inequality holds for the marginal coverage property of CP (5). □

The significance of this statement cannot be overstated, as it implies that thanks to CSRs, it becomes feasible to identify regions in feature space where the conformal coverage of the target class is assured. Consequently, these regions identify feature points with a high degree of certainty, thereby enhancing the reliability, trustworthiness, and robustness of (any) classification algorithm, especially with regard to safety considerations. Specifically, the final output of the proposed method is a region, $\mathcal{S}_\varepsilon$, in which with high probability the chance of finding the unwanted label is small (and thus as small as desired). This means that the scalable classifier together with the conforming prediction can handle the natural uncertainty arising both from the data (to the extent that the data are representative of the information they provide, i.e. *aleatoric uncertainty*) and the model (to the extent that it is accurate in modeling, i.e. *epistemic uncertainty*), providing "safety" sets that have a volume proportional to ε, i.e., to the confidence of the prediction (Hüllermeier and Waegeman 2021).This is very much in line with recent and ongoing literature in the field of geometric uncertainty quantification, as in Sale et al. (2023) where the authors propose the idea of "credal sets" (Abellán et al. 2006) that, as our CSR does, guarantee the correctness of the prediction bounding the input set in polytopes. In this regard, the idea of quantifying uncertainty through functions that give a measure of distance (such as the score function proposed here) is something that is sparking the UQ community, enabling future comparisons with other methods such as the "second order UQ" discussed in Sale et al. (2023).

Remark 2 [On the link with Probably Approximate Correct theory] Probably approximate correct (PAC) learning is a theory developed in the 1980 s by Valiant (2013) for quantifying uncertainty in learning processes, with a focus on the case of undersampled data. PAC learning has been used to define sets of predictions that can satisfy probabilistic guarantees with nonparametric probabilistic assumptions (see, for example, Park et al. (2022)) with similarities with our (and in general with CP theory) approach. Specifically, PAC learning is a broad theory where it is possible to insert the research presented in this paper on uncertainty quantification of machine learning classifiers with conformal prediction. For example, the confidence bounds on which conformal prediction theory is based (and so is



(a) Error plot SVM. (b) Error plot SVDD. (c) Error plot LR.

Fig. 4 Trend of the average error as ϵ varies in [0.05, 0.5] for different classifiers. The errors vary in [0, 0.6] for SVM, [0, 0.8] for SVDD and [0, 0.6] for LR

our research) are inherited from PAC learning theory. As shown in [Vovk (2012) Prop 2a], the concept of (ε, δ)-validity (i.e. the marginal coverage guarantee of equation (5) together with the randomness of the calibration set) is a PAC style guarantee on the (inductive) conformal prediction. As reported in our previous work Carlevaro et al. (2023), there are nontrivial relationships between the number of samples on the calibration set and probabilistic guarantees on prediction. All these relationships can be read into the PAC learning formalism, and future work will focus on this topic.

In the next section we report some numerical examples of Theorem 4, see Fig. 6.

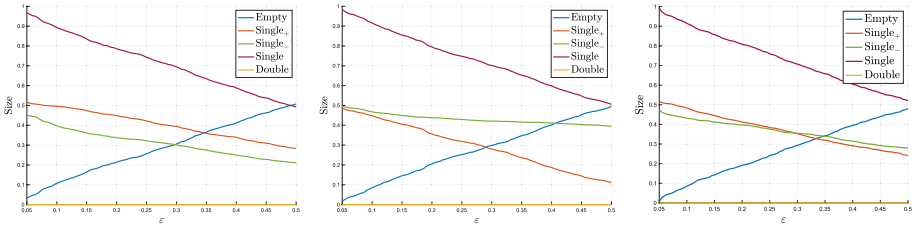
4 A real world application: detection of SSH-DNS tunnelling

The dataset chosen for the example application deals with covert channel detection in cybersecurity (Aiello et al. 2015). The aim is detecting the presence of secure shell domain name server (SSH-DNS) intruders by an aggregation-based monitoring that avoids packet inspection, in the presence of silent intruders and quick statistical fingerprints generation. By modulating the quantity of anomalous packets in the server, we are able to modulate the difficulty of the inherent supervised learning solution via canonical classification schemes (Carlevaro and Mongelli (2021); Vaccari et al. (2022)).

Let q and a be the packet sizes of a *query* and the corresponding *answer*, respectively (what answer is related to a specific query can be understood from the packet identifier) and Dt the time-interval intercurring between them. The information vector is composed of the statistics (mean, variance, skewness and kurtosis) of q, a and Dt for a total number of 12 input features:

$$\boldsymbol{x} = [m_A, m_Q, m_{Dt}, v_A, v_Q, v_{Dt}, s_A, s_Q, s_{Dt}, k_A, k_Q, k_{Dt}],$$

and an overall size of 10000 examples. High-order statistics give a quantitative indication of the asymmetry (skewness) and heaviness of tails (kurtosis) of a probability distribution, helping to improve the detection inference. The output space $\mathcal{Y} = \{-1, +1\}$ is generated by associating each sample $\boldsymbol{x}$ with the label -1 when "no tunnel" is detected and $+1$ when "tunnel" is detected. In this sense, the idea of safety should be interpreted as an indication that the system has detected the presence of a "tunnel" or abnormal behavior, i.e., the system believes that there is a potential security threat or intrusion. This could trigger various security countermeasures, such as blocking incoming traffic or applying filters to the connection.

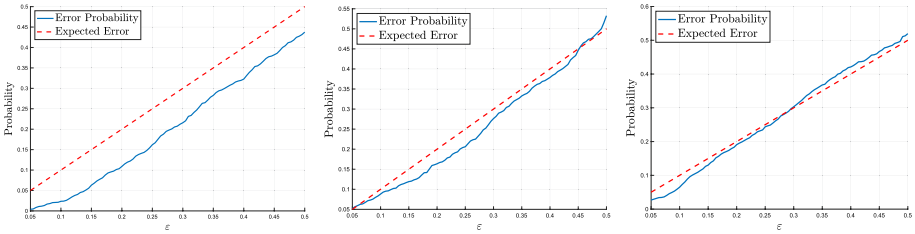


(a) Size plot SVM. (b) Size plot SVDD. (c) Size plot LR.

Fig. 5 Trend of the average size of conformal sets as ε varies in [0.05, 0.5] for different classifiers. The size varies from 0 (empty) to 1 (full)

Conformal predictions assess the goodness of an algorithm by two basic metrics of evaluation: accuracy and efficiency. Accuracy is measured by the average error, over the test set, of the conformal prediction sets considering points of both classes (*err*), only class $y = -1$ points (err $_-$) and only class $y = +1$ points (err $_+$). We remind that an error occurs whenever the true label is not contained in the prediction set. Efficiency is quantified through the rate of test points prediction sets with no predictions (*empty*), two predictions (*double*) and singleton predictions (*single*), these ones also divided by class (*single*$_-$ and *single*$_+$). The obtained results (as the classifier varies) are reported in Figs. 4 and 5 for accuracy and efficiency, respectively.

The overall metrics computed on the benchmark dataset outline the expected behavior of the conformal prediction, with slight differences between the example classifiers. For all values of ε, the average error is indeed bounded by ε in all cases. Also, *err* increases linearly with ε. *This means that the classification is reliable*. As for the size of the conformal set, the overall results point out that for small values of ε the model produces more double-sized regions, since in this way it would be "almost certain" that the true label is contained in the conformal set. Then, the size reduces by increasing ε, allowing for the presence of more empty prediction sets. The number of singleton conformal set remains always sufficiently high (it increases as double conformal sets decrease and it decreases as empty conformal set increase) meaning that *the classification is efficient*. Regarding the use of the example classifiers, it is interesting to note that LR is the most stable with respect to ε and the error conditional on classes: the error rate for both classes is nearly linear with ε, suggesting that the prediction is reliable even conditional on the single class or, better, that the classifier is able to clearly separate the classes while maintaining the expected confidence. The same behavior is also observed for SVM, although the errors per class deviate more from the average error. The error for class "tunnel" is always lower than that for class "no tunnel", suggesting that the classifier is more likely to minimize the number of false positives, losing in accuracy for true negatives. The opposite behavior is observed for SVDD, which instead tries to classify negative instances better, resulting in a lower expected classification error for class "no tunnel". The most interesting aspect, however, is that the algorithm is less conformal when conditioned on the error of the single class, increasing the spread with respect to the average error as ε increases. Conformal prediction together with scalable classifiers define then a totally new framework to deal with uncertainty quantification in classification based scenarios. The results shown in this



(a) Error coverage SVM (b) Error coverage SVDD (c) Error coverage LR

Fig. 6 Error coverage plot as ε varies in [0.05, 0.5] for the example classifiers. The probability varies in [0, 0.5] for SVM, [0.05, 0.55] for SVDD and [0, 0.6] for LR

application drastically overcome the ones obtained on the same dataset in Carlevaro and Mongelli (2021). The previous approach relied on an iterative procedure to control the number of misclassified points that could only be used with a specific algorithm (SVDD) and without a-priori confidence bounds, but only on the basis of an smart trial-and-error algorithm. The point that the reader should observe is precisely this: the presented theory allows dealing with the uncertainty naturally brought by machine learning approaches in a simple and probabilistically grounded way, allowing setting confidence in prediction by design. Finally, Fig. 6 shows the behavior of the coverage error of the CSR with respect to the example classifiers. As stated in Theorem 4, the probability that the wrong label -1 is predicted for the points belonging to $\mathcal{S}_\varepsilon$ is underlinear with respect the expected error ε.

5 Conclusions

Scalable classifiers allow for the development of new techniques to assess safety and robustness in classification problems. With this research we explored the similarities between scalable classifiers and conformal prediction. Through the definition of a score function that naturally derives from the scalable classifier, it is possible to define the concept of conformal safety region, a region that possesses a crucial property known as error coverage, which implies that the probability of observing the wrong label for data points within this region is guaranteed to be no more than a predefined confidence error of ε. Moreover, ongoing studies on the conformal coverage (that is, the probability of observing the true safe label in the CSR is no less than $1-\varepsilon$) suggest that a mathematical proof for this property is conceivable. The idea is to exploit the results on class-conditional conformal prediction as in Vovk (2012). In addition, future work will include the possibility of extending the formulation of scalable classifiers, and thus the conformal safety region, to the multi-class and multi-label context.

The exploration of conformal and error coverages introduces a novel and meaningful concept that holds great promise for applications in the field of reliable and trustworthy artificial intelligence. It has the potential to enhance the assessment of safety and robustness, contributing to the advancement of AI systems that can be trusted and relied upon in critical applications.

Acknowledgements The authors would like to thank Anastasios Angelopulos of University of California, Berkeley and Sara Narteni of CNR-IEIIT for their thoughtful suggestions about Conformal Prediction.

Author contributions Alberto Carlevaro: Methodology, Validation, Investigation, Software, Data curation, Writing. Teodoro Alamo: Methodology, Validation, Investigation, Writing, Supervision. Fabrizio Dabbene: Methodology, Validation, Investigation, Writing, Supervision. Maurizio Mongelli: Methodology, Validation, Investigation, Data curation, Writing, Supervision.

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement. This work was supported in part by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement ID: 101070028. The work was also supported by Future Artificial Intelligence Research (FAIR) project, Italian Recovery and Resilience Plan (PNRR), Spoke 3 - Resilient AI. Moreover T. Alamo acknowledges support from grant PID2022-142946NA-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF, A way of making Europe.

Data availability The datasets generated and analysed during the current study are not publicly available due to potentially sensitive data from the CNR-IEIIT's internal network but are available from the corresponding author on reasonable request.

Data availability The authors prefer not to make the codes available yet. It can be requested from the corresponding author upon reasonable request.

Declarations

Conflict of interest There are no either Conflict of interest or Conflict of interest to declare.

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication Not Applicable.

References

Abellán, J., Klir, G. J., & Moral, S. (2006). Disaggregated total uncertainty measure for Credal sets. *International Journal of General Systems, 35*(1), 29–44.

Andéol, L., Fel, T., De Grancey, F., & Mossina, L. (2024). Conformal prediction for trustworthy detection of railway signals. *AI and Ethics, 4*, 1–5.

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and trends®in machine learning, 16*(4), 494–591. https://doi.org/10.1561/2200000101

Angelopoulos, A.N., Bates, S., Jordan, M. & Malik, J. (2020) Uncertainty sets for image classifiers using conformal prediction. In: International conference on learning representations

Aiello, M., Mongelli, M., & Papaleo, G. (2015). DNS tunneling detection through statistical fingerprints of protocol messages and machine learning. *International Journal of Communication Systems, 28*(14), 1987–2002.

Balasubramanian, V.N., Gouripeddi, R., Panchanathan, S., Vermillion, J., Bhaskaran, A. & Siegel, R. (2009) Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure. In: 2009 36th annual computers in cardiology conference (CinC), (pp. 5–8) . IEEE

Carlevaro, A., Alamo, T., Dabbene, F. & Mongelli, M. (2023) Probabilistic safety regions via finite families of scalable classifiers arXiv:2309.04627 [stat.ML]

Carlevaro, A., Lenatti, M., Paglialonga, A., & Mongelli, M. (2022). Counterfactual building and evaluation via explainable support vector data description. *IEEE Access, 10*, 60849–60861. https://doi.org/10.1109/ACCESS.2022.3180026

Carlevaro, A., & Mongelli, M. (2021). A new SVDD approach to reliable and eXplainable AI. *IEEE Intell Syst*. https://doi.org/10.1109/MIS.2021.3123669

Chzhen, E., Denis, C., Hebiri, M. & Lorieul, T. (2021) Set-valued classification–overview via a unified framework. arXiv:2102.12318 [stat.ML]

Fontana, M., Zeni, G., & Vantini, S. (2023). Conformal prediction: A unified review of theory and new challenges. *Bernoulli, 29*(1), 1–23. https://doi.org/10.3150/21-BEJ1447

Forreryd, A., Norinder, U., Lindberg, T., & Lindstedt, M. (2018). Predicting skin sensitizers with confidence–Using conformal prediction to determine applicability domain of gard. *Toxicology in Vitro, 48*, 179–187. https://doi.org/10.1016/j.tiv.2018.01.021

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*, 457–506. https://doi.org/10.1007/s10994-021-05946-3

Lenatti, M., Carlevaro, A., Guergachi, A., Keshavjee, K., Mongelli, M., & Paglialonga, A. (2022). A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations. *PLOS ONE, 17*(11), 1–24. https://doi.org/10.1371/journal.pone.0272825

Narteni, S., Carlevaro, A., Dabbene, F., Muselli, M. & Mongelli, M. (2023) Confiderai: Conformal interpretable-by-design score function for explainable and reliable artificial intelligence. In: Conformal and probabilistic prediction with applications, (pp. 485–487).

Park, S., Bastani, O., Matni, N. & Lee, I. (2019) Pac confidence sets for deep neural networks via calibrated prediction. In: International conference on learning representations.

Park, S., Dobriban, E., Lee, I., & Bastani, O. (2022). PAC prediction sets for meta-learning. *Advances in Neural Information Processing Systems, 35*, 37920–37931.

Sale, Y., Bengs, V., Caprio, M. & Hüllermeier, E. (2023) Second-order uncertainty quantification: A distance-based approach arXiv:2312.00995 [cs.LG]

Sale, Y., Caprio, M. & Hüllermeier, E. (2023) Is the volume of a credal set a good measure for epistemic uncertainty? arXiv:2306.09586 [cs.LG]

Shafer, G., & Vovk, V. (2008). *A tutorial on conformal prediction, 9*, 371–421. http://jmlr.org/papers/v9/shafer08a.html.

Toccaceli, P. (2022). Introduction to conformal predictors. *Pattern Recognition, 124*, 108507. https://doi.org/10.1016/j.patcog.2021.108507

Vaccari, I., Carlevaro, A., Narteni, S., Cambiaso, E., & Mongelli, M. (2022). eXplainable and reliable against adversarial machine learning in data analytics. *IEEE Access, 10*, 83949–83970. https://doi.org/10.1109/ACCESS.2022.3197299

Valiant, L. (2013). *Probably approximately correct: Nature's algorithms for learning and prospering in a complex world*. USA: Basic Books Inc.

Vovk, V. (2012) Conditional validity of inductive conformal predictors. In: Hoi, S.C.H., Buntine, W. (eds.) Proceedings of the Asian conference on machine learning. Proceedings of machine learning research, vol. 25, pp. 475–490. PMLR, Singapore Management University, Singapore . https://proceedings.mlr.press/v25/vovk12.html

Vovk, V., Gammerman, A. & Saunders, C. (1999) Machine-learning applications of algorithmic randomness. In: Proceedings of the sixteenth international conference on machine learning. ICML '99, pp. 444–453. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Berlin, Heidelberg: Springer.

Vovk, V., Gammerman, A., & Shafer, G. (2022). *Probabilistic classification: Venn predictors* (pp. 157–179). Cham: Springer. https://doi.org/10.1007/978-3-031-06649-8_6

Vovk, V., Shen, J., Manokhin, V. & Xie, M. (2017) Nonparametric predictive distributions based on conformal prediction. In: Conformal and probabilistic prediction and applications, pp. 82–102