POLITECNICO DI TORINO
Repository ISTITUZIONALE

CONFIDERAI: a novel CONFormal Interpretable-by-Design score function for Explainable and Reliable Artificial Intelligence

Original
CONFIDERAI: a novel CONFormal Interpretable-by-Design score function for Explainable and Reliable Artificial Intelligence / Carlevaro, A., Narteni, S., Dabbene, F., Muselli, M., Mongelli, M.. - ELETTRONICO. - (2023).

Availability:
This version is available at: 11583/2982246 since: 2023-09-18T08:59:07Z

Publisher:

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

27 November 2026

CONFIDERAI: a novel CONFormal Interpretable-by-Design score function for Explainable and Reliable Artificial Intelligence

Alberto Carlevaro †, Sara Narteni*†, Fabrizio Dabbene, Marco Muselli and Maurizio Mongelli

arXiv:2309.01778v2 [cs.LG] 6 Sep 2023

Abstract—Everyday life is increasingly influenced by artificial intelligence, and there is no question that machine learning algorithms must be designed to be reliable and trustworthy for everyone. Specifically, computer scientists consider an artificial intelligence system safe and trustworthy if it fulfills five pillars: explainability, robustness, transparency, fairness, and privacy. In addition to these five, we propose a sixth fundamental aspect: conformity, that is, the probabilistic assurance that the system will behave as the machine learner expects. In this paper, we propose a methodology to link conformal prediction with explainable machine learning by defining CONFIDERAI, a new score function for rule-based models that leverages both rules predictive ability and points geometrical position within rules boundaries. We also address the problem of defining regions in the feature space where conformal guarantees are satisfied by exploiting techniques to control the number of non-conformal samples in conformal regions based on support vector data description (SVDD). The overall methodology is tested with promising results on benchmark and real datasets, such as DNS tunneling detection or cardiovascular disease prediction.

Index Terms—Conformal Prediction, Explainable AI, Conformal Critical Regions, Error Control, Rule Based Models, SVDD.

1 Introduction

Trustworthy Artificial Intelligence (AI) is an umbrella-term that gained increasing importance in recent years to establish the requirements of real-world AI systems for their proper design, development and deployment. Among its principles, *explainability* (or transparency) and *technical robustness and safety* (or reliability) have an essential role [1]. Explainability allows each actor involved to understand the reasoning behind any machine learning (ML) decision. There is a plethora of techniques to achieve explainability today, falling under the *eXplainable AI (XAI)* research theme [2]. At a high level, the main categorization of XAI distinguishes post-hoc explanations of black box models and transparent-by-design techniques [3]. The latter category includes rule-based models, where predictions are characterized by easy-to-understand decision rules (often expressed in the *if-then* format).

Nevertheless, even though very helpful thanks to their native interpretability, rule-based models alone are not enough to ensure a correct performance of the model. Therefore, many approaches have been proposed so far to guarantee the safety of ML models [4], [5], also relying on (or devoted to) XAI methodologies [6], [7], [8], [9]. Among them, conformal prediction (CP) stands out with its solid mathematical foundation, that allows to generate prediction sets with predefined probabilistic guarantees for any ML model [10]. However, as discussed in Section 2, we were not able to find in the current literature works specifically focusing on CP for transparent-by-design XAI models. Motivated by this observation, in this paper we investigate this topic, by proposing an innovative score function that enables CP for rule-based models. On the other hand, various studies exploit CP to perform false positives or negatives control. For example, [11] proposes a multi-label conformal prediction approach in which the false positive rate is probabilistically controlled by requiring prediction sets to eliminate non-conformal points. [12] inserts the performance control directly on the expected value of any loss function and [13] introduces the concept of safety score that warns the system when a predefined level of error is reached.

In our work, which is the multidimensional extension of our previous work [14], we will pursue the same objective by defining a set in which the performance on a target class is guaranteed by the score function of the conformal prediction itself. Specifically, we introduce a conformal critical set containing target points with high guarantees provided through the CP.

From this set, a region in the feature space is constructed through a binary classifier that distinguishes between conformal-critical points (i.e., those belonging to the conformal critical set) and non-conformal/non-critical points (not belonging to the set) and, in this case, a false positive

- * *Corresponding author*
- † *These authors contributed equally to the development of the article.*
- *A. Carlevaro, S. Narteni, M. Muselli and M. Mongelli are with CNR-IEIIT, Corso F.M. Perrone 24, Genoa, 16152, Italy.*
- *F. Dabbene is with CNR-IEIIT, Corso Duca degli Abruzzi 24, Turin, 10129, Italy*
  E-mail: name.surname@ieiit.cnr.it
- *A. Carlevaro is also with University of Genoa, Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), 16145 Genoa, Italy*
- *S. Narteni is also with Politecnico di Torino, Department of Control and Computer Engineering (DAUIN), 10129, Turin, Italy*
- *M. Muselli is also with with Rulex Innovations Labs, Via Felice Romani 9, 16122, Genoa, Italy.*

control is applied (i.e., non-conformal/non-critical points predicted as conformal-critical points) as in [15].

1.1 Contribution

Based on the above considerations, in our viewpoint, combining CP framework with XAI is thus essential in the direction of trustworthy AI. However, this topic is little explored in current literature, hence our paper attempts to address such a research gap through the following contributions:

- We design and develop CONFIDERAI, a new score function that allows to build conformal predictors for rule-based models, by leveraging the combination of the global performance properties of decision rules (i.e., their covering and error) and the geometrical position of the points inside rule boundaries.
- We introduce the concept of *conformal critical set*, i.e., the set of target points for which CONFIDERAI indicates high probabilistic guarantees of the underlying ML model. Moreover, by exploiting SVDD-based techniques for the false positives control, we individuate *conformal critical regions* characterized by the largest number of target points and the minimum non-target points, thus ensuring further precision of the decision-making algorithm.

The remaining of the paper is structured as follows: Section 2 reports existing approaches concerning CP and XAI models, with particular focus on rule-based ones; Section 3 introduces the fundamentals of conformal prediction framework and provides the mathematical definition of the *Conformal Critical Set*; Section 4 describes our core contribution, the CONFIDERAI score function, along with simple toy examples to provide the reader with a visual intuition on how the score works; Section 5 describes the fundamental properties of the specific rule-based model adopted in our case studies, i.e., the Logic Learning Machine (LLM); Section 7 reports and discusses the application of the proposed approach on relevant benchmarks and real-world datasets; finally, Section 8 concludes the paper.

2 Related Works

As anticipated, conformal prediction for XAI models is still little investigated in research. Authors in [16] proposed the application of conformal predictors for evaluating the confidence of tree ensemble models, such as random forests. Also, [17] studied CP for random forests in a multi-label classification scenario. Standard inductive conformal prediction (ICP) for classification through decision tree models is investigated in [18], and the same authors [19] present such a framework to perform rule extraction with guarantees, either considering rule extraction *for* opaque models or rule extraction *using* opaque models. Recently, in [20], [21], the same approaches were extended to regression tasks. In all cases, authors define a conformity function based on the margin between true and predicted target probability estimates.

Even more recently, [22] introduced a conformal decision rule learning algorithm, where rule generation and Mondrian conformal prediction [23] processes are combined together, by devising a separate ICP on each rule at hand. Given unseen test samples satisfying a generic rule of any XAI model, conformal decision rules provide a point prediction according to the consequent of the rule and a prediction set based on the results of conformal prediction for that rule. Another study [24] involves the combination of rule-based models and CP frameworks for multi-label classification, by defining a conformity score that, for any candidate point, depends on the predictive quality of the top-performing rule. However, compared to our work, this score does not take into account the different position of instances within rule geometrical shape. To the best of our knowledge, no previous study of these type address score functions and quantile, tailored for rule-based models (Sec. 3-4 here).

3 Conformal Predictions and Critical Sets

Theory behind CP is substantially based on two, exchangeable, approaches: either $i)$ (non)-conformity measure and p-value as in [10] or $ii)$ score functions and quantile as in [25]. As already pointed out, these two methodologies are actually the same but, in our opinion, the use of score functions and quantile is of more efficiency to understand properly the potential of XAI in CP, since score function links directly the conformal sets with the model. Moreover, we introduce a slightly new concept, the *conformal critical set*, that allows to insert CP in a more safety-based context. As a matter of fact, our aim is to define a subset of the input feature space in which probabilistic guarantees can be provided to the ML model and to exploit explainable techniques to make it a fully trustworthy and easy-to-understand model.

Let $\mathcal{X}$ be a measurable feature space and $\mathcal{Y}$ be the output space. We here consider the case of *binary* classification, with $\mathcal{Y} = \{0, +1\}$. Note that this choice of labels is without loss of generality, since any binary classifier can be converted to these labels. Conformal prediction states that, for any machine learning model $\hat{f}(\boldsymbol{x})_y : \mathcal{X} \longrightarrow \mathcal{Y}$, it is possible to define a *score function* $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$, which depends in some suitable way from the model:

$$s(\boldsymbol{x}, y) \sim \hat{f}(\boldsymbol{x})_y.$$

The score function should be designed so that larger scores encode worse agreement between point $\boldsymbol{x}$ and label y.

In our context, we assume that the label $+1$ denotes the target class S, which is to be interpreted as the presence of *critical situations* in the system. The label 0 refers to the non-target class instead, which denotes the absence of such conditions.

Remark 3.1 (On the meaning of critical points). Note that the meaning of the term "critical" is context-dependent. For example, in a situation in which safety is of paramount

importance, i.e. for instance in the case of collision avoidance, one may be interested in finding regions of the feature space where collision is avoided with high probability. In this case, the terms "critical" and "safe" may be seen as synonyms*. On the other hand, there are cases in which one would like to report only critical cases when the probability of failure is very high, so to avoid false alarms. In this case, the class $+1$ would correspond to the "failure" case.

On the basis of the score function, a *prediction set* at *level of confidence* $1-\varepsilon$, $\varepsilon \in (0,1)$, can be defined for any $\boldsymbol{x} \in \mathcal{X}$:

$$\mathcal{C}_\varepsilon(\boldsymbol{x}) = \{y \mid s(\boldsymbol{x}, y) \leq s_\varepsilon\} \in 2^{\mathcal{Y}}, \tag{1}$$

where s_ε is the $\lceil (n_c+1)(1-\varepsilon) \rceil / n_c$ *quantile* of the score values computed on a *calibration set* $\mathcal{Z}_c \doteq \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_c}$, of size n_c. The prediction set guarantees the *marginal coverage* property

$$1-\varepsilon \leq \Pr\{y \in \mathcal{C}_\varepsilon(\boldsymbol{x})\} \leq 1-\varepsilon + \frac{1}{n_c+1}, \tag{2}$$

where "marginal" means that the probability is averaged over the randomness of the calibration set.
Keeping in mind all the above considerations on how to properly set a conformal prediction, we define the *conformal critical set* (CCS) at *confidence level* $1-\varepsilon$ the subset $\mathcal{S}_\varepsilon \subseteq \mathcal{X}$ as follows:

$$\mathcal{S}_\varepsilon = \left\{ \boldsymbol{x} \mid s(\boldsymbol{x}, +1) \leq s_\varepsilon,\ s(\boldsymbol{x}, 0) > s_\varepsilon \right\}. \tag{3}$$

In words, the CCS is a subset of the input space where the prediction set is composed by only unsafe points $(\boldsymbol{x}, +1)$. This means that the model $\hat{f}$ is likely to make safe predictions for inputs in $\mathcal{S}_\varepsilon$ with a specified level of error ε.

3.1 Explicit bounds on the sample complexity of the calibration set

It should be remarked that the choice of the size of the calibration set is of crucial importance, since its affects the conformal prediction. The point is that the probability in equation (2) can vary by sampling differently the calibration set. However, [26] introduces the concept of (E,δ)-*validity*, i.e., given $\delta \in (0,1)$ and $E \in (\varepsilon, 1)$, it holds that

$$\Pr_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{n_c}} \left\{ \Pr\{y \in \mathcal{C}_\varepsilon(\boldsymbol{x})\} \geq 1-E \right\} \geq 1-\delta. \tag{4}$$

(E,δ)-validity guarantees that, on average, the proportion of errors made by the conformal prediction is bounded by E, with a probability of at least $1-\delta$. In practical terms, this means that if we were to repeat the prediction process multiple times on different calibration sets, the average proportion of errors made by the conformal prediction should not exceed E, and the probability of observing an error larger than E should be at most δ.

We note here that the above concepts allow the number of samples (i.e. the so-called *sample complexity*) in the calibration set to be chosen operationally. Indeed, introducing the notation $\mathbf{B}(n; N, p)$ for the binomial of n trials and probability of success p, it is shown in [26] that (4) is guaranteed if:

$$\mathbf{B}\left(\nu; n_c, E\right) \leq \delta, \tag{5}$$

with

$$\nu \doteq \lfloor (E(n_c+1) - 1 \rfloor.$$

An important observation is that one can then exploit recent results obtained in the field of chance-constraint approximation to derive *explicit bounds* on the number of samples necessary to guarantee the desired probabilistic properties. Indeed, from [27, Theorem 1], it is easy to see that (5) is guaranteed if we choose

$$n_c \geq \frac{1}{E}\left(\nu + \ln\frac{1}{\delta} + \sqrt{2\nu \ln\frac{1}{\delta}}\right) \tag{6}$$

which [28] proves to be satisfied by sampling

$$n_c \geq \frac{7.47}{E} \log\frac{1}{\delta} \tag{7}$$

independent and identically distributed (i.i.d.) samples.

Remark 3.2. In what follows, we identify E with ε. Although this may not match $(\varepsilon,\delta)-$validity, it simplifies the calculations inherent to the desired score function. We use the identification $E = \varepsilon$ for the sake of operativity, thus applying (7) directly. A more accurate setting would investigate how E should be chosen in the interval (ε, 1) under the $(\varepsilon,\delta)-$validity constraint and consequently deriving n_c under the bound in Equation (7). Such a setting is currently under study.

4 Rule-Based Conformity

In the conformal prediction framework, as stated in Section 3, any score function value $s(\boldsymbol{x}, y)$ is higher for any label y that is less likely to be the correct prediction for the considered point $\boldsymbol{x}$. In this work, we aim at designing a new score function suitable for rule-based machine learning models.

4.1 Rule-based models notation

Before going into the design details, let us briefly describe the main characteristics and notation of rule-based models.
Let us consider an input example space for classification $\mathcal{T} = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{N} \in \mathcal{X} \times \mathcal{Y}$, with $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^D$ and $y \in \{0,1\}$.
A rule-based binary classifier $g : \mathcal{X} \longrightarrow \mathcal{Y}$ is expressed by a set of decision rules $\mathcal{R} = \{r_k\}_{k=1}^{M_r}$ in the following form: **if** *premise* **then** *consequence*. The *premise* constitutes the antecedent of the rule and is a logical conjunction ($\wedge$) of conditions c_{i_k}, with $i_k = 1_k, \ldots, N_k$.
Any condition c_{i_k} corresponds to one of the following intervals:

1) $x_{\pi(i)} \geq l_{i_k}$
2) $x_{\pi(i)} \leq u_{i_k}$
3) $l_{i_k} \leq x_{\pi(i)} \leq u_{i_k}$

where l_{i_k}, u_{i_k} are proper numerical thresholds determined by the learning algorithm and $\pi : \mathbb{N} \longrightarrow \mathbb{N}$ denotes

*In this case, the label $+1$ will denote no collision, and 0 will correspond to collision.

the permutation of the indexes of the feature vector $\boldsymbol{x}$ that associates the rule i_{th} condition with the corresponding feature component. Finally, the *consequence* expresses the output class of the decision rule.

Another useful concept in rule-based learning is the notion of *rule relevance*, assigning to each rule a value in the [0,1] range which resembles its predictive ability. Specifically, it is computed by combining the covering $C(r_k)$ and error $E(r_k)$ metrics (commonly known as True Positive Rate and False Positive Rate of the rule, respectively), defined as follows:

$$C(r_k) = \frac{TP(r_k)}{TP(r_k) + FN(r_k)} \tag{8}$$

$$E(r_k) = \frac{FP(r_k)}{TN(r_k) + FP(r_k)} \tag{9}$$

Denoting with $\hat{y}_j$ the class label predicted by the rule r_k for point $(\boldsymbol{x}_j, y_j)$, $TP(r_k)$ and $FP(r_k)$ are defined as the number of instances that correctly and wrongly satisfy rule r_k, being $\hat{y}_j = y_j$ and $\hat{y}_j \neq y_j$ respectively; conversely, $TN(r_k)$ and $FN(r_k)$ represent the number of samples $(\boldsymbol{x}_j, y_j)$ which do not meet at least one condition in rule r_k, with $\hat{y}_j \neq y_j$ and $\hat{y}_j = y_j$, respectively.
Then, *rule relevance* $R(r_k)$ of rule r_k can be found as:

$$R(r_k) = C(r_k) \cdot (1 - E(r_k)) \tag{10}$$

4.2 CONFIDERAI score function

Given a rule r_k generated by a rule-based model after training, and predicting an output class y, it implies a hyper-rectangle as a decision boundary in the feature space (being defined by the premise of the rule). The closer a point covered by r_k is to this boundary, the higher is its probability of being wrongly covered by the rule. Conversely, points lying inside the rule hyper-rectangle, but farther from the boundary are most probably well conforming to the rule output. So we have to take into account a score that penalizes more the points closer to the classification boundary. For this reason, we introduce the quantity $\gamma = \gamma(\boldsymbol{x}, r_k)$ defined as:

$$\gamma = \sum_{i=1}^{N_k} \left(\frac{1}{d_i^-(\boldsymbol{x}, c_{i_k})} + \frac{1}{d_i^+(\boldsymbol{x}, c_{i_k})} \right), \tag{11}$$

where

$$d_i^-(\boldsymbol{x}, c_{i_k}) = |x_{\pi(i)} - l_{i_k}| \quad \text{and} \quad d_i^+(\boldsymbol{x}, c_{i_k}) = |x_{\pi(i)} - u_{i_k}|$$

and $\varphi(d) : \mathbb{R} \to \mathbb{R}$ being a monotonically decreasing (scalar) function.
In the sequel, we will let $\varphi(d) = \frac{1}{d}$, but other choices are possible. For example, one could set $\varphi(d) = \exp(-\alpha d)$. In this way, a variation on α leads to a variation on the velocity of descent, allowing to control it properly.

In order to compute both d_i^- and d_i^+ when either l_{i_k} or u_{i_k} are missing, i.e., when condition c_{i_k} assumes, respectively, the second or the first form described in Section 4.1, the minimum and maximum value of feature $\boldsymbol{x}_{\pi(i)}$ across the dataset is considered.

Example 4.1. To give the reader a complete understanding of the γ factor, consider the following two-dimensional example as in Figure 1. In this case, the green point should be more conformal than the red point, since its position is farther from the boundaries of the rule. Then, the γ factor associated with $\boldsymbol{x}_1$, γ_1, should be smaller than the one associated with $\boldsymbol{x}_2$, γ_2:

$$\begin{aligned}\gamma_1 &= \sum_{i=1}^{2} \left(\frac{1}{d_i^-(\boldsymbol{x}_1, c_{i_k})} + \frac{1}{d_i^+(\boldsymbol{x}_1, c_{i_k})} \right) \\ &= \frac{1}{8} + \frac{1}{14} + \frac{1}{4} + \frac{1}{6} = \frac{103}{168} = 0.6131\end{aligned}$$

$$\begin{aligned}\gamma_2 &= \sum_{i=1}^{2} \left(\frac{1}{d_i^-(\boldsymbol{x}_2, c_{i_k})} + \frac{1}{d_i^+(\boldsymbol{x}_2, c_{i_k})} \right) \\ &= \frac{1}{18} + \frac{1}{4} + \frac{1}{2} + \frac{1}{8} = \frac{67}{72} = 0.9306\end{aligned}$$

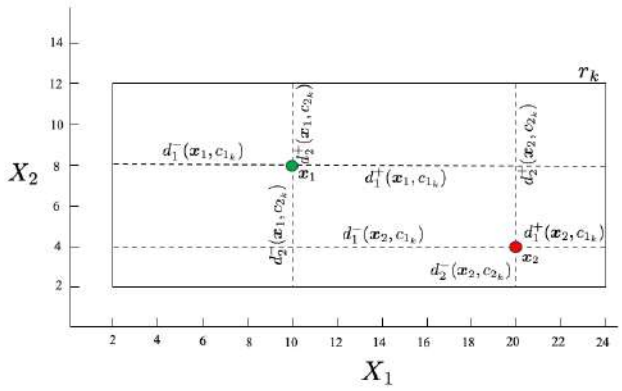


Figure 1. Example of γ factor construction for a two-dimensional feature vector $\boldsymbol{x} = (x_1, x_2)$.

Finally, we apply the sigmoid function to γ so that its values vary in the $[0, 1]$ range, thus obtaining the following term:

$$\tau(\boldsymbol{x}, r_k) = \frac{1}{1 + e^{-\gamma}}, \tag{12}$$

which is used in combination with rule relevance to define a *score* for point $\boldsymbol{x}$ and class label y:

$$s(\boldsymbol{x}, y) \doteq \sum_{r_k \in \mathcal{R}_{\boldsymbol{x}}^y} \tau(\boldsymbol{x}, r_k)(1 - R(r_k)), \tag{13}$$

where the sum is on the set $\mathcal{R}_{\boldsymbol{x}}^y$ of rules predicting label y and verified by the input point $\boldsymbol{x}$.

Remark 4.1. The presence of the sum term brings the assumption that multiple rules can *overlap*. However, the proposed score function does not lose generality and remains valid even for models resulting in non-overlapping rules: in this case, ruleset $\mathcal{R}_{\boldsymbol{x}}^y$ will have cardinality fixed to one.

In this way, the introduced score takes into account both the geometrical position of points with respect to rule boundaries and, by depending on rule relevance, the predictive ability of the rules. The latter contribution is

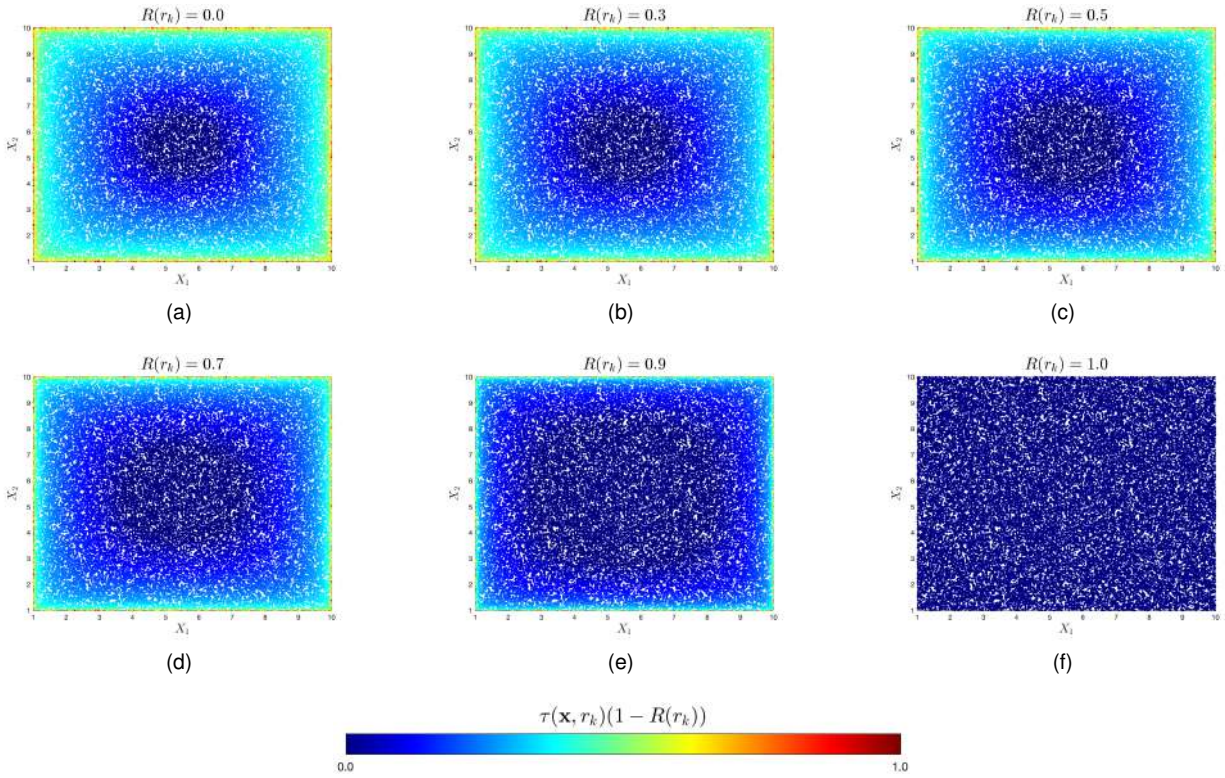


Figure 2. Toy example showing rule relevance contribution to the score function

expressed through the term $(1 - R(r_k))$ (and not directly through $R(r_k)$) in order to keep the score low when classification has better performance, that is when rule relevance is higher. To better show this behavior, an illustrative example is shown in the next Section 4.3.

4.3 Toy Examples in 2D

To point out the contribution of rule relevance on the score values, we designed a simple yet explicative example.
Let us consider a bidimensional feature space formed by features X_1 and X_2, and suppose that a rule r_k is learned on such a space, being characterized by the following premise:

$$1 \leq X_1 \leq 10 \ \wedge \ 1 \leq X_2 \leq 10$$

Assuming these thresholds fixed, the geometrical boundaries of the rule remain unchanged and Figure 2 shows the effect of increasing relevance values of r_k (from 0 in Fig. 2a to 1 in Fig. 2f). By looking at the figure, we can observe that when $R(r_k) \leq 0.5$, the score values mainly depend on the geometrical contribution defined by Eq. 11 and 12: indeed, points that are closer to rule boundaries are well distinguishable to the others. Conversely, as relevance grows ($R(r_k) = 0.7$), its contribution gets more significant, by lowering the score value even for points that lie close to the boundaries. This is even more pronounced in the extreme case of $R(r_k) = 1$, where the predictive ability of the rule would be so high that it overwhelms the geometrical contribution.
In practice, this design choice handles the possible case when multiple rules have the same geometrical shape (in terms of aspect ratio of their boundary), but different relevance value. As shown in Fig. 3, two points (red cross) located at the same distance to the respective rule boundary are scored with a higher value when the rule has a low relevance (left rectangle), and, viceversa, a lower value when the relevance is high (right rectangle).

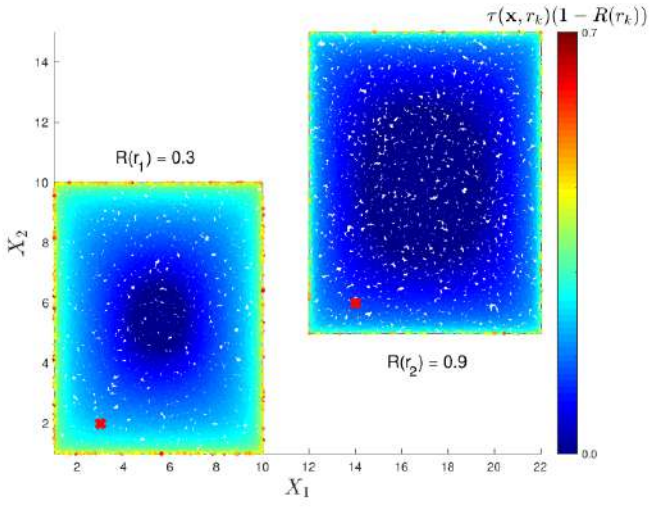


Figure 3. Toy example showing two rules $r_k, k = \{1, 2\}$ with relevance $R(r_1) = 0.3$ and $R(r_2) = 0.9$, respectively, whose boundaries share the same aspect ratio. The red cross point in r_1 has a higher score than the one in r_2.

5 Logic Learning Machine

The rule-based model adopted in our work is the Logic Learning Machine (LLM), designed and developed by Rulex[†] as a more efficient variant of Switching Neural Networks [29]. Rule generation through LLM takes place in three steps: first, the process starts by discretizing the features and binarizing them via the inverse-only-one coding. The resulting binary strings are then concatenated into a single large string representing the considered samples. Subsequently, shadow clustering is used to build logical structures, called implicants, in the Boolean lattice, which are finally transformed into sets of conditions and combined into a collection of intelligible rules [30], [31]. It is worth underlying that the LLM design process is thus based on an *aggregate-and-separate* approach [32] able to generate a set of rules that can *overlap*. As a result, an input sample $\boldsymbol{x}$ may verify multiple rules predicting the same class label and it even may cover rules predicting different output classes.

Let us denote with $\mathcal{R}_{\boldsymbol{x}}$ the set of all rules satisfied by $\boldsymbol{x}$. LLM class assignment is then performed based on relevance values.

Specifically, given a generic point $\boldsymbol{x}$, and the set $\mathcal{R}_{\boldsymbol{x}}^{y}$ of rules predicting label y and verified by the point, a class label $\hat{y}$ is assigned to $\boldsymbol{x}$ by solving the following problem [33]:

$$\hat{y} = arg\max_{y} \left(\sum_{r \in \mathcal{R}_{\boldsymbol{x}}^{y}} R(r) \right) \tag{14}$$

The topical issue in the conformal framework relies on the fact that the predictions $\hat{y}$ in (14) are not exploited directly as they do not provide any guarantee alone (on the reliability of label assignments). They rather drive the search of guaranteed subspaces of data, on the basis of the set of predictions in $\mathcal{C}_{\varepsilon}(\boldsymbol{x})$. This is the argument of the following section.

6 From conformal critical sets to conformal critical regions

As per Equation 3, a conformal critical set at a fixed ε can be identified. Subsequently, test points belonging to this set can be labelled as *conformal-critical*, providing a new way to look at the dataset. Indeed, we can train a new classifier to individuate the largest region of only *conformal-critical* points as possible, i.e., a *Conformal Critical Region* (CCR), that is a good approximation of the CCS defined in (3).

The conformal-critical points enclosed in these regions thus constitute the set of points with label $+1$ for which the classification is statistically validated. The identification of their boundaries proves very important in real applications, since going outside of them identifies a zone in the feature space where the correct classification of $+1$ points is no more guaranteed, hence other solutions should be sought, such as another training configuration, another model, etc. In light of the Trustworthy AI principle of technical robustness and safety, this result is crucial.

[†]https://www.rulex.ai

6.1 Examples

Some examples may help understand. In case of a dynamical system (e.g., robots moving in a given environment, vehicle approaching specific maneuvers), the critical region may represent the subsets of states in which the application is considered safe (e.g., collision avoidance states, as to the vehicle platooning considered in the performance evaluation section). Finding those subsets may be a hard task in many practical situations and only data driven solutions are applicable (e.g., platooning is again a good example in that respect). On the other hand, those solutions need some guarantee and this is where the conformal framework comes into play. It consists of finding the regions with zero (or at least controllable) false negatives of (predicted) safety states, namely, the zones in which the dynamical system may move without experiencing any danger (such as collision with other agents or obstacles). The conformal critical set gives such an assurance. The safety engineer knows that the AI application has been properly designed as soon as the trajectories lye within that set. Characterizing the boundaries of the set is even more important in order to trigger appropriate alarms before any danger may take place.

Other examples may be provided with respect to the control of false positive rate. A first example deals with cybersecurity. In many cases, the alarms inherent to ongoing attacks lead to severe digital (sometimes even physical) countermeasures, such that security analysts want to be sure that an attack is really in play before unleashing all the necessary restrictions. In that respect, the conformal critical set profiles the conditions of the system (a network or a critical infrastructure) surely associated with the presence of attack (i.e., zero false positive). A second example deals with disease diagnosis by AI, as being made on the basis of clinical data over a population of patients (in comparison with healthy individuals). In this case, positive answer by the AI means the disease is predicted; false positive means the prediction was not realistic (e.g., after additional exams, the inauspicious diagnosis becomes invalidated). Circumventing the cases in which patients are surely affected by the disease is of great interest, also in respect to differentiating them with respect to the other cases where the disease prediction does not lie in the conformal borders. In the latter situation, additional exams should be even more urgent to settle the matter (disease or not disease). We believe the reader may think to many other circumstances of analogous practical interest.

6.2 Approximating the conformal critical sets

The construction of conformal critical regions is model-agnostic, i.e. it is possible to use any binary classifier that individuates conformal-critical points, obtaining a region $\tilde{S}_{\varepsilon}$ that approximates the CCS S_{ε}. However, it should be pointed out that our target is to construct closed and well defined sets. In this perspective, a good model is the Support Vector Data Description [34], a variation of the well-known SVM, since it is able to define closed envelopes enclosing target points (i.e. conformal-critical) controllable by a radius and a center. In this case, a

gaussian-kernel based SVDD has been trained to separate and characterize the conformal-critical points. Moreover, a technique to minimize the number of misclassified points inside the conformal critical region was adopted as in [15] by performing successive iterations of SVDD inside the classification boundary (SafeSVDD, or SSVDD in short). In this case, since conformal critical regions must guarantee the highest level of confidence as possible, the number of false positives (i.e. non-conformal/non-critical points wrongly classified as conformal-critical points) has been minimized. We also remark that this kind of control is equivalent to ensuring that the minimum number of non-target points remains within the CCR.

Remark 6.1. What we obtain with the CCR $\tilde{S}_\varepsilon$ is only a (good) approximation of the conformal critical set, i.e. a region where conformity is expected but not guaranteed.

In the following we present a pseudo-code to describe the algorithm for the construction of the CCR. But first, let us indicate with n_+ the number of critical points (i.e. labelled with $+1$) in the calibration set, and with n_0 the non-critical ones. We have then $n_c = n_+ + n_0$. Moreover we indicate with the acronym FPR the false positive rate. [‡]

Algorithm 1 `Conformal Critical Region`
Input Critical class calibration set $(\boldsymbol{x}_1, +1)$, $(\boldsymbol{x}_2, +1), \ldots, (\boldsymbol{x}_{n_+}, +1)$, conformal critical set $\mathcal{S}_\varepsilon$, threshold on FPR η.
Output CCR $\tilde{\mathcal{S}}_\varepsilon$.

1: Classify

$$\boldsymbol{x}_i \longrightarrow \begin{cases} +1 & \text{if} \quad \boldsymbol{x}_i \in \mathcal{S}_\varepsilon \\ -1 & \text{otherwise} \end{cases}$$

and create a new dataset

$$\mathcal{X} \times \tilde{\mathcal{Y}} = \{(\boldsymbol{x}_i, \tilde{y}_i) \mid \tilde{y}_i \in \{-1, +1\}\}.$$

2: Train a binary classifier on $\mathcal{X} \times \tilde{\mathcal{Y}}$ (e.g. SVDD).
3: While FPR $< \eta$, control the misclassification error (e.g. SSVDD as in [15]).
4: If FPR $< \eta$ then **output** $\tilde{S}_\varepsilon$.

We remark again that the procedure to obtain a CCR is model-agnostic, i.e. in principle it is possible to use any (binary) classifier. We decided to use SVDD (and its controllable version SSVDD as in [15]) because it already contains in its definition the property of defining closed regions: since our intention is to contour critical points, we evaluated it to be the most suitable choice for an initial evaluation of our method.
As final remarks, it is worthy to underline that CCRs represent an *applicative* tool to assess conformity in real contexts, where the aim is to individuate an envelope with high probabilistic guarantees on a specific class of interest. Indeed, the introduced CCS and the related new way to label the dataset to find CCRs can be defined in a flexible way, according to the desired kind of guarantees for the problem at hand. In this paper, our focus was on finding a proper envelope for target ($+1$) points and the CCS of Eq. 3 was accordingly defined to contain points that led to singleton $+1$ labels; therefore, the boundaries of the obtained CCRs in our experiments outline the zones of the features space that lead to a critical situation with very low presence of false alarms. However, by moving the focus on the non-target (0) points, a CCS could have been defined so to enclose only singleton 0 labels, and the same SVDD-based approach (or any other classifier) could have been used to find CCRs that delimited these points, this time defining an area where a non-critical situation could have been guaranteed with high probability, with a reduced number of missed alarms.

[‡]We are interested in FPs since our target (critical) class is labelled with $+1$. Since our intention is to find a region that contains the highest number of conformal-critical points ($+1$) we want to minimize the number of non-critical points (0) incorrectly identified as conformal-critical, i.e. false positives.

7 Experimental Results

In this Section, we present the results of the experiments devoted to test CONFIDERAI score functions, both in terms of canonical metrics in conformal prediction evaluation (i.e., accuracy and efficiency, see Sec. 7.2) and of our newly introduced conformal critical set (Sec. 7.3).

7.1 Datasets description

To evaluate the goodness of CONFIDERAI, we tested the method on 10 datasets, which we briefly describe:

- **P2P** and **SSH**: two datasets concerning peer-to-peer (P2P) and secure shell (SSH) applications of a Domain Name Server (DNS) tunneling detection system [35]; the aim is to detect the presence or absence of DNS attacks by monitoring network traffic and collecting statistical information.
- **BSS**: the Body Signals of Smoking dataset [§] collects personal and biological measurements from subjects, with the aim of predicting if these quantities can represent biomarkers of *smoking* or *non-smoking* habits.
- **CHD**: the Cardiovascular Heart Disease dataset[¶] contains patients records with personal, clinical and behavioral features to predict the presence or the absence of a cardiovascular disease.
- **Vehicle Platooning**: the dataset consists of simulations of a vehicle platooning system [36] with a binary output of *collision* or *not-collision* under physical features like the number of cars per platoon or the initial distance between cars.
- **RUL**: the Turbofan Engine Degradation Simulation dataset[||] deals with damage propagation modeling for aircraft engines. The goal is to understand which conditions are inherent to imminent faults of the engine by estimating its Remaining Useful Life.

[§]Reference link: https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking?select=smoking.csv
[¶]Reference link: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.
[||]Reference link: https://www.kaggle.com/datasets/behrad3d/nasa-cmaps.

- **EEG**: the Eye State Classification EEG dataset[**] reports the state of patients' eyes (open or closed) based on continuous electroencephalogram (EEG) measurements.
- **MQTTset** [37]: based on Message Queue Telemetry Transportation communication protocol, this dataset collects measurements from different Internet of Things devices to simulate a smart environment; cyber-attacked data are also included to detect *malicious* and *legitimate* traffic.
- **Magic**: the Magic Gamma Telescope dataset[††] reports Monte Carlo simulations of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope to distinguish between gamma and hadron radiation.
- **Fire Alarm**: this dataset[‡‡] contains data to develop an AI-based smoke detection device.

7.2 Accuracy and Efficiency

For the evaluation, we explored the bounds introduced in Section 3.1, considering both accuracy and efficiency, by setting $\varepsilon = 0.01$, $\varepsilon = 0.05$, $\varepsilon = 0.1$ and $\varepsilon = 0.2$. Accuracy was measured by the average error, over the test set, of the conformal prediction sets considering points of both classes (*AvgErr*), only class $y = 0$ points (*AvgErr0*) and only class $y = 1$ points (*AvgErr1*). We remind that an error occurs whenever the true label is not contained in the prediction set. Efficiency was quantified through the rate of test points prediction sets with singleton predictions (*Single*), no predictions (*Empty*) and two predictions (*Double*). The obtained results are reported in Table 1.

The overall metrics computed on the benchmark datasets outline the expected behavior of the conformal prediction. For all values of ε, the average error is indeed bounded by ϵ in all cases, except for the MQTTset dataset at $\varepsilon = 0.2$ whose average error is lower than expected, probably due to the complexity of the dataset. Also, *AvgErr* increases linearly with ε. As for the size of the conformal set, results in their overall 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 the presence of more empty or singleton prediction sets. Exception to this general trend is observed for P2P and Fire Alarm cases, where the rates of singleton prediction sets are considerably high even at low ε. These results also denote that the underlying LLM model has a reliable performance on the two mentioned datasets, while the relatively low efficiency on the other datasets indicates that the original model should be improved.

As an example, we chose SSH dataset to show the average errors and prediction regions size obtained by varying $\varepsilon \in [0.05, 0.5]$. Figure 4 reports the trends of these metrics, pointing out the aforementioned behaviors at the increase of ε. In most of the cases, the average error on class 0 , i.e. the *legitimate* samples, is lower than the average error on class 1, i.e. the *attack* points. Concerning the size, we can notice that double-size prediction regions are always dominant with respect to singleton and empty regions: their percentage decreases with ε, while that of singleton regions increases symmetrically and empty regions are rarely observed. Also, at $\varepsilon = 0.4$ *Single* and *Double* metrics assume a constant trend.

7.3 Conformal Critical Regions

Besides evaluating the error and the size of the obtained prediction regions, we derived the conformal critical regions by following our definition of conformal critical set (Eq. 3) and the SVDD-based approach as explained in Section 6. The performance for the CCRs is evaluated by considering the number of conformal-critical points inside the regions, i.e. the empirical probability $\Pr\{S \mid \boldsymbol{x} \in \tilde{S}_\varepsilon\}$. Specifically, we denote this probability with $\Pr_{\text{SVDD}}$ for the classical SVDD, and $\Pr_{\text{SSVDD}}$ for the optimized Safe-SVDD version.

The results are shown in Figure 5 for SSH dataset at $\varepsilon = 0.05$ and in Table 1 for all the other datasets. By analyzing the two empirical probabilities obtained throughout all datasets and ε values, we can observe that SafeSVDD always overcomes classical SVDD (as expected) and achieves more than the 80% of correct detections for the majority of the experiments. In particular, the value of $\Pr_{\text{SSVDD}}$ for the SSH case indicates that within the region the prediction of DNS tunneling attacks is performed correctly in over the 92% of cases with the SafeSVDD method.

8 Conclusion

This paper introduced CONFIDERAI, a new score function for rule-based models directly designed on top of the properties of these models. Indeed, starting from the decision rules generated by the model, conformity is derived as a function of the placement of the samples with respect to the geometry of the model, also taking into account rule relevance, a measure that reflects the predictive quality of a rule. Extensive experimentation by considering the Logic Learning Machine model on several datasets has shown a behavior in line with conformal prediction framework, both in terms of accuracy and efficiency of the prediction sets.

In addition, by leveraging on the results of CONFIDERAI, we moved a step beyond the probabilistic guarantees provided by the conformal predictions, in the direction of a more safety-preserving solution. Thus, we first defined the notion of conformal critical set that provides guarantee to efficiently predict the target class points (i.e., the critical ones) in high probability (thanks to the CP); then, we exploited further classifiers to reduce the false positives (i.e., non-target points) associated to such a set, leading to the individuation of conformal critical regions, which proved still a good approximation of the conformal critical set. We also highlight that the idea of finding CCSs and CCRs

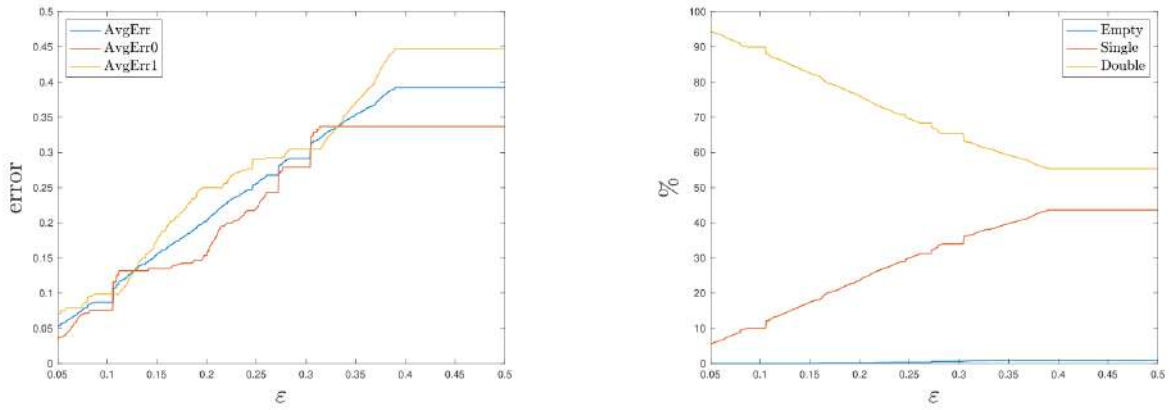
[**]Reference link: https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State.

[††]Reference link: https://www.kaggle.com/datasets/abhinand05/magic-gamma-telescope-dataset.

[‡‡]Reference link: https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset.

Table 1
Evaluation metrics for CONFIDERAI on Logic Learning Machine model tested on 10 benchmark datasets.

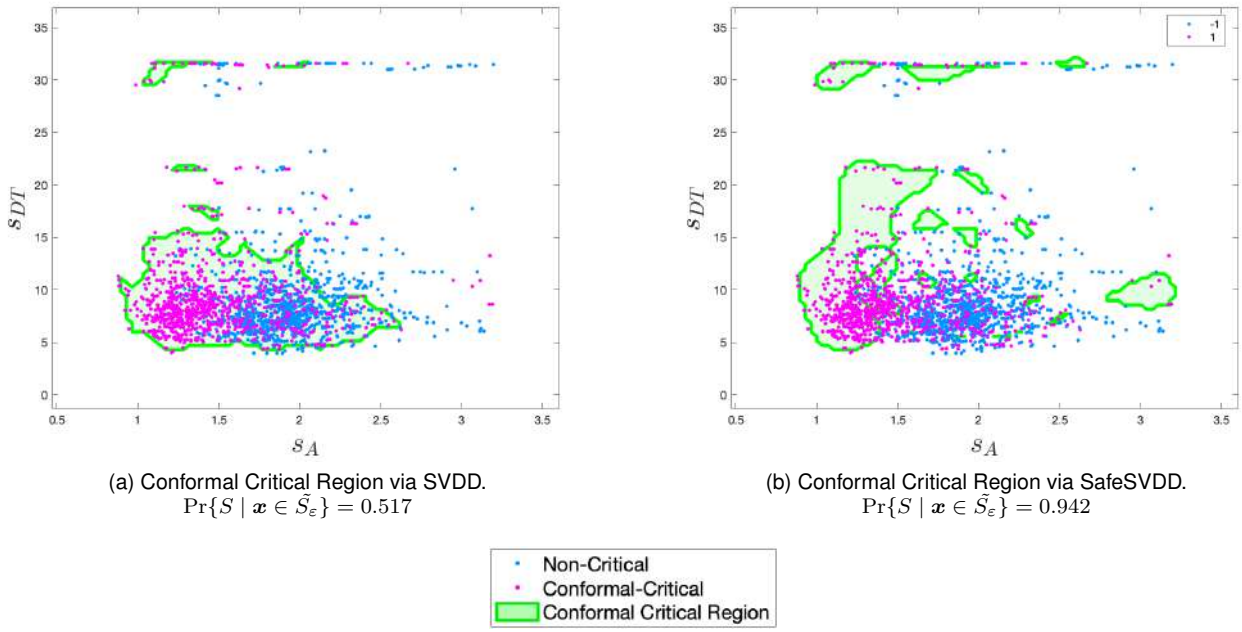
		Error			**Size**			**CCR**	
		avgErr	avgErr0	avgErr1	Empty	Single	Double	$\Pr_{SVDD}$	$\Pr_{SSVDD}$
P2P	$\varepsilon = 0.01$	0.009	0.017	0.000	0.009	0.938	0.054	0.505	0.995
	$\varepsilon = 0.05$	0.053	0.106	0.000	0.053	0.893	0.054	0.512	0.962
	$\varepsilon = 0.1$	0.096	0.192	0.000	0.075	0.910	0.015	0.496	1.000
	$\varepsilon = 0.2$	0.190	0.379	0.000	0.168	0.818	0.014	0.478	0.962
SSH	$\varepsilon = 0.01$	0.012	0.007	0.016	0.000	0.012	0.988	0.484	1.000
	$\varepsilon = 0.05$	0.053	0.035	0.070	0.000	0.056	0.944	0.517	0.942
	$\varepsilon = 0.1$	0.087	0.076	0.099	0.000	0.101	0.899	0.527	0.892
	$\varepsilon = 0.2$	0.202	0.154	0.250	0.002	0.236	0.763	0.489	1.000
BSS	$\varepsilon = 0.01$	0.009	0.012	0.003	0.000	0.017	0.983	0.514	0.986
	$\varepsilon = 0.05$	0.050	0.059	0.033	0.003	0.094	0.903	0.529	0.935
	$\varepsilon = 0.1$	0.102	0.110	0.087	0.014	0.191	0.795	0.547	0.949
	$\varepsilon = 0.2$	0.196	0.220	0.153	0.038	0.374	0.588	0.545	0.909
CHD	$\varepsilon = 0.01$	0.012	0.003	0.019	0.000	0.032	0.968	0.246	1.000
	$\varepsilon = 0.05$	0.051	0.011	0.087	0.002	0.089	0.909	0.356	1.000
	$\varepsilon = 0.1$	0.104	0.025	0.176	0.007	0.250	0.743	0.268	1.000
	$\varepsilon = 0.2$	0.208	0.064	0.338	0.020	0.386	0.594	0.316	0.998
Vehicle Platooning	$\varepsilon = 0.01$	0.012	0.020	0.003	0.000	0.012	0.988	0.452	0.997
	$\varepsilon = 0.05$	0.050	0.064	0.034	0.000	0.054	0.946	0.445	0.996
	$\varepsilon = 0.1$	0.100	0.138	0.052	0.000	0.105	0.895	0.454	0.994
	$\varepsilon = 0.2$	0.206	0.230	0.175	0.004	0.240	0.756	0.425	0.954
RUL	$\varepsilon = 0.01$	0.010	0.012	0.004	0.000	0.011	0.989	0.321	0.605
	$\varepsilon = 0.05$	0.044	0.050	0.032	0.001	0.067	0.931	0.285	0.606
	$\varepsilon = 0.1$	0.086	0.110	0.032	0.001	0.112	0.887	0.304	0.850
	$\varepsilon = 0.2$	0.182	0.203	0.134	0.027	0.243	0.730	0.361	0.875
EEG	$\varepsilon = 0.01$	0.011	0.008	0.014	0.000	0.019	0.981	0.404	0.996
	$\varepsilon = 0.05$	0.049	0.053	0.045	0.000	0.065	0.935	0.416	0.989
	$\varepsilon = 0.1$	0.095	0.103	0.085	0.005	0.123	0.872	0.385	0.957
	$\varepsilon = 0.2$	0.176	0.202	0.148	0.008	0.210	0.782	0.465	0.987
MQTTset	$\varepsilon = 0.01$	0.011	0.000	0.021	0.000	0.011	0.989	0.496	0.750
	$\varepsilon = 0.05$	0.050	0.000	0.099	0.000	0.050	0.950	0.439	0.689
	$\varepsilon = 0.1$	0.106	0.000	0.209	0.000	0.106	0.894	0.470	0.821
	$\varepsilon = 0.2$	0.135	0.000	0.266	0.000	0.135	0.865	0.571	0.920
Magic	$\varepsilon = 0.01$	0.013	0.012	0.014	0.000	0.016	0.984	0.329	1.000
	$\varepsilon = 0.05$	0.057	0.035	0.093	0.001	0.072	0.927	0.325	0.532
	$\varepsilon = 0.1$	0.095	0.064	0.147	0.004	0.123	0.873	0.338	1.000
	$\varepsilon = 0.2$	0.210	0.173	0.274	0.007	0.244	0.749	0.268	0.690
Fire Alarm	$\varepsilon = 0.01$	0.006	0.013	0.000	0.006	0.967	0.026	0.506	0.868
	$\varepsilon = 0.05$	0.044	0.067	0.022	0.034	0.954	0.013	0.526	0.813
	$\varepsilon = 0.1$	0.076	0.131	0.022	0.065	0.923	0.013	0.505	0.787
	$\varepsilon = 0.2$	0.188	0.359	0.022	0.178	0.810	0.013	0.540	0.834



(a) Average error on both and single classes

(b) Percentage of empty, single-label and two-labels prediction regions

Figure 4. Trend of the performance metrics obtained on the SSH dataset by varying $\varepsilon \in [0.05,\ 0.5]$



(a) Conformal Critical Region via SVDD.
$\Pr\{S \mid \boldsymbol{x} \in \tilde{S}_\varepsilon\} = 0.517$

(b) Conformal Critical Region via SafeSVDD.
$\Pr\{S \mid \boldsymbol{x} \in \tilde{S}_\varepsilon\} = 0.942$

Figure 5. Conformal safety regions with the (optimized) classical SVDD (5a) and the region obtained reducing the number of non-conformal points with Safe-SVDD (5b).

is actually independent on the proposed score function for rule-based models, but it rather can be adopted to whatever CP framework and underlying algorithm.
The current work is a starting point for the development of a fully conformal rule-based methodology for trustworthy AI. Future works will involve a more in-depth experimentation of other rule-based models and their assessment on real world applications. Moreover it will be of crucial importance to prove the probabilistic bounds given in Section 3, in order to determine a well specified framework for probabilistic conformal prediction.

ACKNOWLEDGMENTS

The authors would like to thank Teodoro Alamo of University of Seville for thoughtful discussions about conformal predictions and probabilistic safety sets.
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.

REFERENCES

[1] High-Level Expert Group on AI, "Ethics guidelines for trustworthy ai," report, European Commission, Brussels, Apr. 2019.

[2] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, p. 101805, 2023.
[3] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, pp. 1–66, 2022.
[4] A. V. S. Neto, J. B. Camargo, J. R. Almeida, and P. S. Cugnasca, "Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work," *IEEE Access*, vol. 10, pp. 130733–130770, 2022.
[5] S. Dey and S.-W. Lee, "Multilayered review of safety approaches for machine learning-based systems in the days of ai," *Journal of Systems and Software*, vol. 176, p. 110941, 2021.
[6] J. Marques-Silva and A. Ignatiev, "Delivering trustworthy ai through formal xai," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 12342–12350, 2022.
[7] A. Hurault and J. Marques-Silva, "Certified logic-based explainable ai," 2023.
[8] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, "Learning certifiably optimal rule lists," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 35–44, 2017.
[9] S. Narteni, V. Orani, I. Vaccari, E. Cambiaso, and M. Mongelli, "Sensitivity of logic learning machine for reliability in safety-critical systems," *IEEE Intelligent Systems*, vol. 37, no. 5, pp. 66–74, 2022.
[10] G. Shafer and V. Vovk, "A tutorial on conformal prediction," 2007.
[11] A. Fisch, T. Schuster, T. Jaakkola, and D. Barzilay, "Conformal prediction sets with limited false positives," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 6514–6532, PMLR, 17–23 Jul 2022.
[12] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," *arXiv preprint arXiv:2208.02814*, 2022.
[13] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, and M. Pavone, "Sample-efficient safety assurances using conformal prediction," in *International Workshop on the Algorithmic Foundations of Robotics*, pp. 149–169, Springer, 2022.
[14] S. Narteni, A. Carlevaro, F. Dabbene, M. Muselli, and M. Mongelli, "Confiderai: Conformal interpretable-by-design score function for explainable and reliable artificial intelligence," in *Conformal and Probabilistic Prediction with Applications*, pp. 485–487, PMLR, 2023.
[15] A. Carlevaro and M. Mongelli, "A new svdd approach to reliable and explainable ai," *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 55–68, 2022.
[16] S. Bhattacharyya, "Confidence in predictions from random tree ensembles," in *2011 IEEE 11th International Conference on Data Mining*, pp. 71–80, 2011.
[17] H. Wang, X. Liu, B. Lv, F. Yang, and Y. Hong, "Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional chinese medicine," *PloS one*, vol. 9, no. 6, p. e99565, 2014.
[18] U. Johansson, H. Boström, and T. Löfström, "Conformal prediction using decision trees," in *2013 IEEE 13th International Conference on Data Mining*, pp. 330–339, 2013.
[19] U. Johansson, R. König, H. Linusson, T. Löfström, and H. Boström, "Rule extraction with guaranteed fidelity," in *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pp. 281–290, Springer, 2014.
[20] U. Johansson, H. Linusson, T. Löfström, and H. Boström, "Interpretable regression trees using conformal prediction," *Expert systems with applications*, vol. 97, pp. 394–404, 2018.
[21] U. Johansson, C. Sönströd, T. Löfström, and H. Boström, "Rule extraction with guarantees from regression models," *Pattern Recognition*, vol. 126, p. 108554, 2022.
[22] H. Abdelqader, E. Smirnov, M. Pont, and M. Geijselaers, "Interpretable and reliable rule classification based on conformal prediction," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I*, pp. 385–401, Springer, 2023.
[23] V. Vovk, D. Lindsay, I. Nouretdinov, and A. Gammerman, "Mondrian confidence machine," *Technical Report*, 2003.
[24] E. Hüllermeier, J. Fürnkranz, and E. Loza Mencia, "Conformal rule-based multi-label classification," in *KI 2020: Advances in Artificial Intelligence: 43rd German Conference on AI, Bamberg, Germany, September 21–25, 2020, Proceedings 43*, pp. 290–296, Springer, 2020.
[25] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," 2021.
[26] V. Vovk, "Conditional validity of inductive conformal predictors," 2012.
[27] M. Mammarella, V. Mirasierra, M. Lorenzen, T. Alamo, and F. Dabbene, "Chance-constrained sets approximation: A probabilistic scaling approach," *Automatica*, vol. 137, p. 110108, 2022.
[28] V. Mirasierra, M. Mammarella, F. Dabbene, and T. Alamo, "Prediction error quantification through probabilistic scaling," *IEEE Control Systems Letters*, vol. 6, pp. 1118–1123, 2022.
[29] M. Muselli, "Switching neural networks: A new connectionist model for classification," 01 2005.
[30] S. Parodi, C. Manneschi, D. Verda, E. Ferrari, and M. Muselli, "Logic learning machine and standard supervised methods for hodgkins lymphoma prognosis using gene expression data and clinical variables," *Health Informatics Journal*, vol. 24, 06 2016.
[31] D. Cangelosi, F. Blengio, R. Versteeg, A. Eggert, A. Garaventa, C. Gambini, M. Conte, A. Eva, M. Muselli, and L. Varesio, "Logic learning machine creates explicit and stable rules stratifying neuroblastoma patients," *BMC bioinformatics*, vol. 14, no. 7, pp. 1–20, 2013.
[32] M. Mongelli, E. Ferrari, M. Muselli, and A. Fermi, "Performance validation of vehicle platooning through intelligible analytics," *IET Cyber-Physical Systems: Theory & Applications*, vol. 4, no. 2, pp. 120–127, 2019.
[33] E. Ferrari, D. Verda, N. Pinna, and M. Muselli, "A novel rule-based modeling and control approach for the optimization of complex water distribution networks," in *Advances in System-Integrated Intelligence: Proceedings of the 6th International Conference on System-Integrated Intelligence (SysInt 2022), September 7-9, 2022, Genova, Italy*, pp. 33–42, Springer, 2022.
[34] D. Tax and R. Duin, "Support vector domain description," *Pattern Recognition Letters 20*, pp. 1191–1199, 1999.
[35] M. Aiello, M. Mongelli, and G. Papaleo, "Dns tunneling detection through statistical fingerprints of protocol messages and machine learning," *International Journal of Communication Systems*, vol. 28, no. 14, pp. 1987–2002, 2015.
[36] M. Mongelli, M. Muselli, E. Ferrari, and A. Fermi, "Performance validation of vehicle platooning via intelligible analytics," *IET Cyber-Physical Systems: Theory & Applications. 4. 10.1049/iet-cps.2018.5055*, 2018.
[37] I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso, "Mqttset, a new dataset for machine learning techniques on mqtt," *Sensors*, vol. 20, no. 22, 2020.



Alberto Carlevaro received the Master Degree in Applied Mathematics from the University of Genoa, Italy, in 2020 with a physics-mathematics thesis. He is a PhD student in the Department of Electrical, Electronic and Telecommunications Engineering and Naval Architecture (DITEN) of the University of Genoa, in collaboration with CNR-IEIIT and S.M.E. Aitek, and now visiting research scholar at EECS department of UC Berkeley. His current fields of research are Machine Learning, Statistical Learning, Explainable AI and Physics Informed Machine Learning.



Sara Narteni got her M.Sc. in Bioengineering at the University of Genoa on March 2020, with a thesis entitled “Pleural line ultrasound videos analysis for computer aided diagnosis in acute pulmonary failure”. She is currently a PhD student in the Italian National PhD program on Artificial Intelligence for Industry 4.0 based at Politecnico di Torino, working at CNR-IEIIT Institute in Genoa. Her research interests include trustworthy artificial intelligence, with specific focus on Explainable Artificial Intelligence methods and applications to different fields, such as healthcare, industry and automotive.



Fabrizio Dabbene (Senior Member, IEEE) received the Laurea and Ph.D. degrees from the Politecnico di Torino, Italy, in 1995 and 1999, respectively. He is currently the Director of Research with the Institute IEIIT, National Research Council of Italy (CNR), Milan, Italy, where he is coordinates the Information and Systems Engineering Group. He has held visiting and research positions with The University of Iowa, Penn State University, and the Russian Academy of Sciences, Institute of Control Science, Moscow, Russia. He has authored or coauthored more than 100 research papers and two books. Dr. Dabbene was an Elected Member of the Board of Governors, from 2014 to 2016. He has served as the vice president for publications, from 2015 to 2016. He is currently chairing the IEEE-CSS Italy Chapter. He has also served as an Associate Editor for Automatica, from 2008 to 2014, and IEEE Transactions on Automatic Control, from 2008 to 2012. He is also a Senior Editor of the IEEE Control Systems Society Letters..



Marco Muselli graduated from the University of Genoa in 1985, with a Master’s degree in electronic engineering. In 1988, he joined the Italian National Research Council (CNR) where he currently works as a Senior Researcher at the Institute of Electronics, Information Engineering and Telecommunications (IEIIT). His research interests include artificial intelligence, bioinformatics, optimization, mathematical statistics, and probability theory. His work is mainly focused on developing new efficient machine learning methods, capable of generating intelligible rules, and on the theoretical analysis of their convergence properties. In 2014, Marco Muselli founded Rulex Inc., an international company operating in the sector of data-driven decisioning, with the aim of developing general and specific techniques for analyzing real data, providing decision support and knowledge extraction. Currently, as the CEO of Rulex, Muselli drives company strategies and activities, with a keen focus on research and development. He is the author or co-author of more than 150 scientific publications, most of which have been published in international journals.



Maurizio Mongelli obtained his PhD. Degree in Electronics and Computer Engineering from the University of Genoa in 2004. He worked for Selex and the Italian Telecommunications Consortium (CNIT) from 2001 until 2010. He is now a researcher at CNR-IEIIT, where he deals with machine learning applied to health and cyber-physical systems. He is co-author of over 100 international scientific papers, 2 patents and is participating in the SAE G-34/EUROCAE WG-114 AI in Aviation Committee.