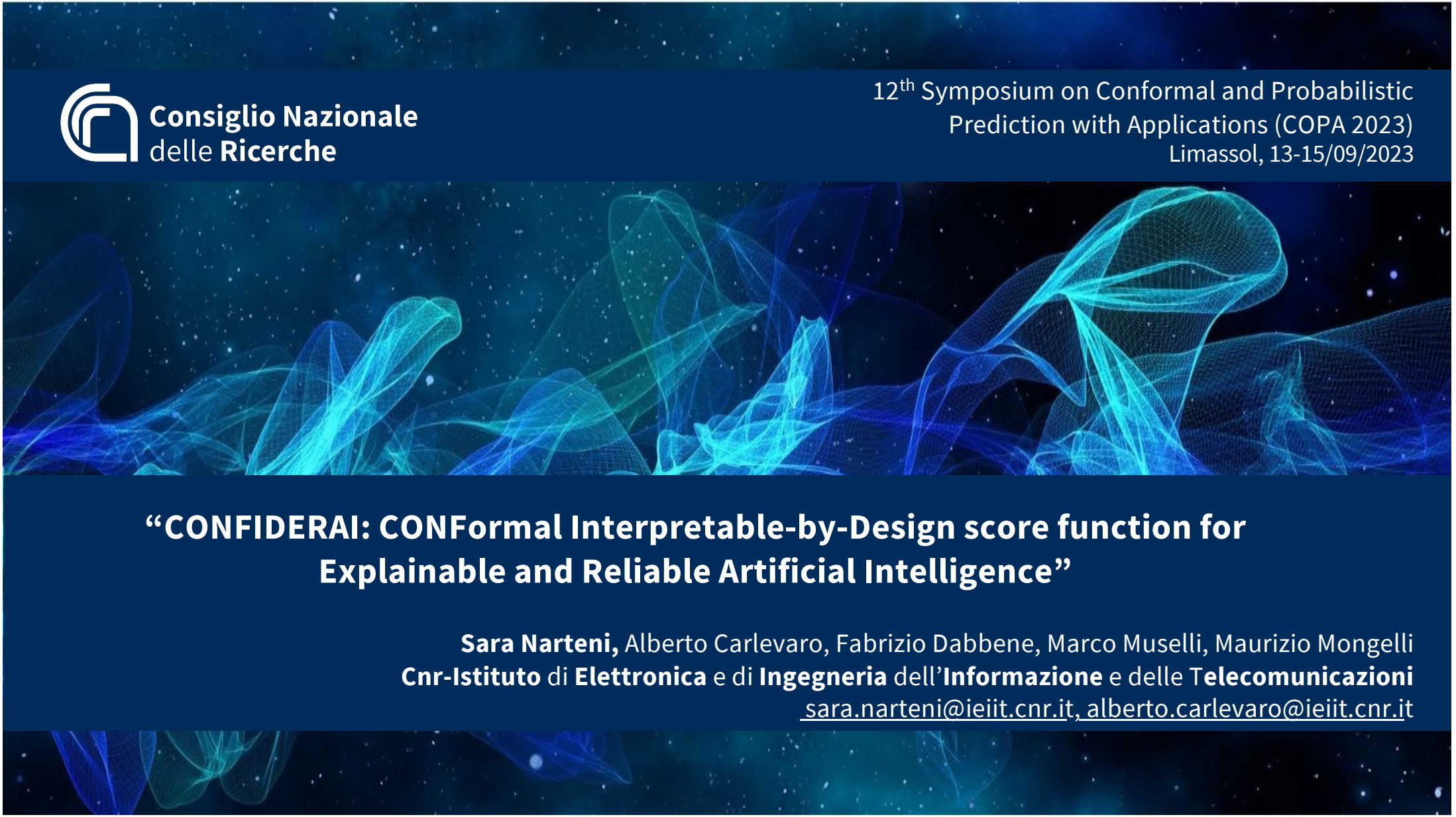


Consiglio Nazionale delle Ricerche

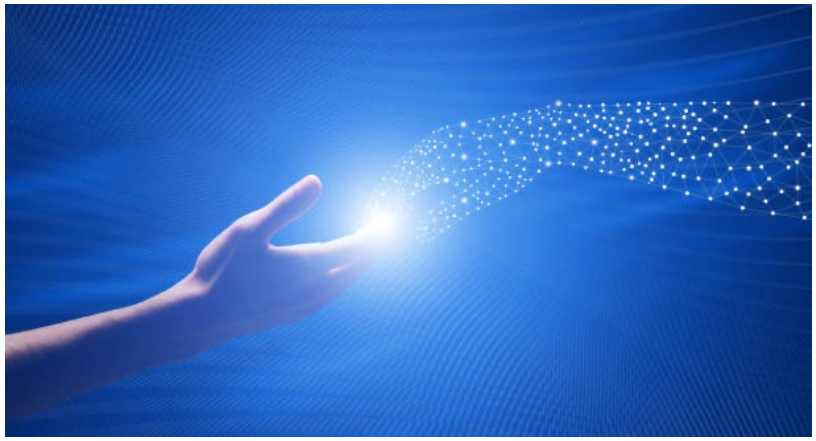
12th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2023)
Limassol, 13-15/09/2023

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

Sara Narteni, Alberto Carlevaro, Fabrizio Dabbene, Marco Muselli, Maurizio Mongelli
Cnr-Istituto di **Elettronica** e di **Ingegneria** dell'**Informazione** e delle T**elecomunicazioni**
sara.narteni@ieiit.cnr.it, alberto.carlevaro@ieiit.cnr.it

OUTLINE

- Motivation and background
- CONFIDERAI
  - Conformal Safety Set
  - Novel score function for rule-based models
- Preliminary applications
- Conclusions and future steps



MOTIVATION AND BACKGROUND

- Increased diffusion of **trustworthy AI**
  1. Need of ***transparency***: eXplainable AI (XAI)
  2. Need of ***safety guarantees***

- Strong conformal prediction frameworks
  1. (non)-conformity measure + p-values [1]
  2. score functions + quantile [2]

- Little investigation on CP for XAI models (with the 1st technique)

Our work: combining XAI and CP by designing a new score function for rule-based models

Picture credit: European Commission, Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*, Publications Office, 2019, https://data.europa.eu/doi/10.2759/346720

[1] Glenn Shafer, Vladimir Vovk, a tutorial on conformal prediction, arXiv, 2007, https://arxiv.org/abs/0706.3188
[2] Anastasios N. Angelopoulos, Stephen Bates, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, arXiv, 2021, https://arxiv.org/abs/2107.07511

Consiglio Nazionale delle Ricerche

FROM CONFORMAL PREDICTION TO CONFORMAL SAFETY SETS

Machine Learning model $\hat{f}(\boldsymbol{x})_y : \mathcal{X} \longrightarrow \mathcal{Y}$ $\sim$ **Score function** $s(\boldsymbol{x}, y) : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$

Size of ***calibration set***

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

$$\frac{\lceil (n_c + 1)(1 - \varepsilon) \rceil}{n_c} \text{ quantile}$$

$$\mathcal{S}_\varepsilon = \left\{ \boldsymbol{x} \mid s(\boldsymbol{x}, y) \leq s_\varepsilon, \forall y \in \mathcal{Y} \right\}$$

Conformal Safety Set
(all labels)

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

New dataset labelling

Binary Classifier

Conformal Safety Region

FROM CONFORMAL PREDICTION TO CONFORMAL SAFETY SETS

Machine Learning model $\hat{f}(\boldsymbol{x})_y : \mathcal{X} \longrightarrow \mathcal{Y}$ $\sim$ **Score function** $s(\boldsymbol{x}, y) : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$

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

Size of ***calibration set***

$$\frac{\lceil (n_c + 1)(1 - \varepsilon) \rceil}{n_c} \text{ quantile}$$

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

Ongoing extension:
focus on a **target class** (i.e., +1)

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

New dataset labelling

Binary Classifier

Conformal Safety Region

NOTATION FOR RULE-BASED MODELS

$$\mathcal{T} = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{N}, \text{ with } \boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \text{ and } y \in \{0, 1\}$$

$$\mathcal{R} = \{r_k\}_{k=1}^{M} \longrightarrow \textbf{if } \textit{premise } \textbf{then } \textit{consequence}$$

The *premise* constitutes the antecedent of the rule and is a logical conjunction ($\wedge$) of conditions c_{i_k}, with $i_k = 1_k, \ldots, N_k$

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

$$\left.\begin{aligned} C(r_k) &= \frac{TP(r_k)}{TP(r_k) + FN(r_k)} \\ E(r_k) &= \frac{FP(r_k)}{TN(r_k) + FP(r_k)} \end{aligned}\right\} \quad R(r_k) = C(r_k) \cdot (1 - E(r_k)) \quad \textbf{Rule relevance}$$

NOVEL SCORE FUNCTION FOR RULE-BASED MODELS

$\boldsymbol{x} = (x_1, x_2)$: 2D input point
y: output label
r_k: k_{th} rule of the model

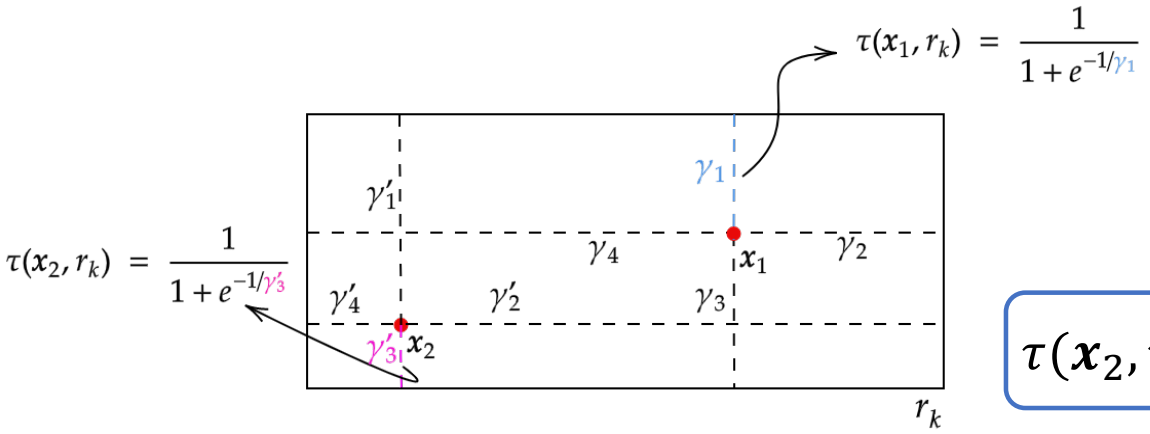
$\gamma = \gamma(\boldsymbol{x}, r_k)$: minimum of Euclidean distances between $\boldsymbol{x}$ and rule boundary sides

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

rule relevance

rule predicting y and verified by $\boldsymbol{x}$

position of point within rule boundary

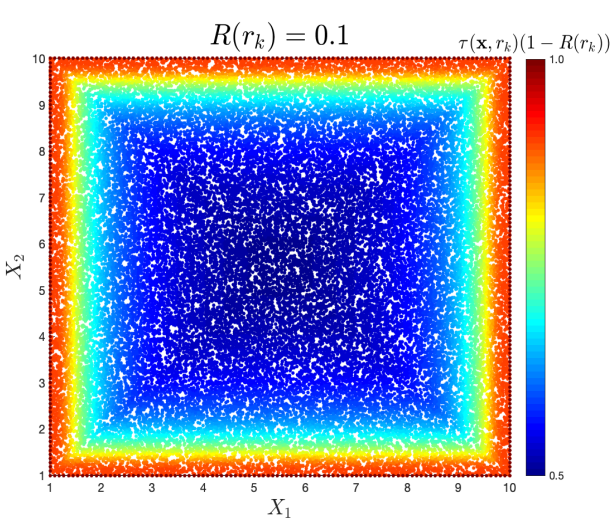


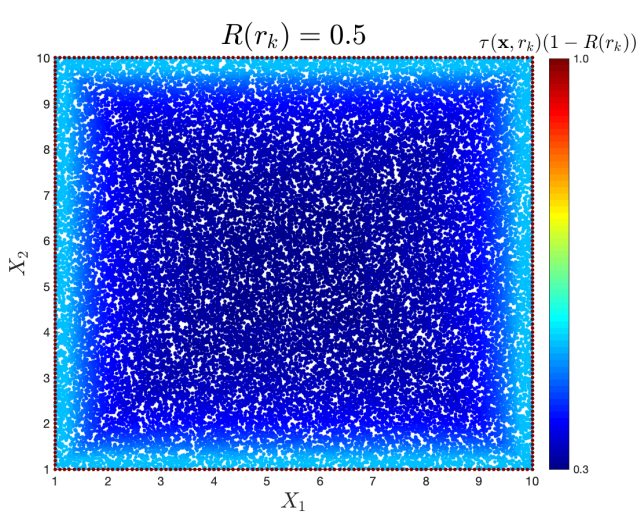
$$\tau(\boldsymbol{x}, r_k) = \frac{1}{1 + e^{-1/\gamma}}$$

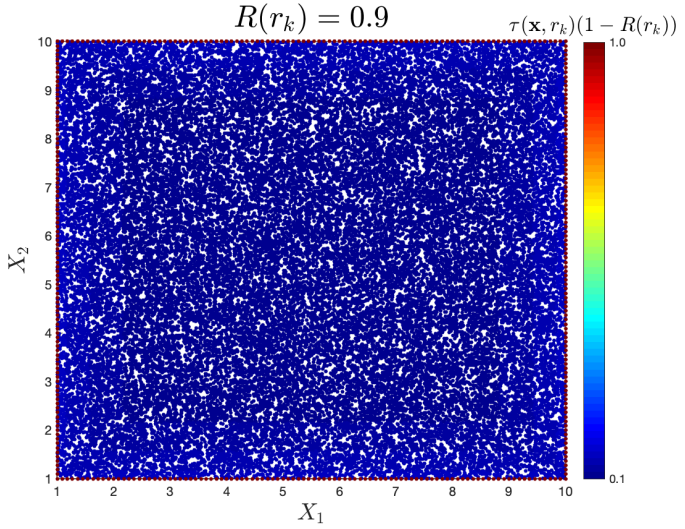
$$\tau(\boldsymbol{x}_2, r_k) > \tau(\boldsymbol{x}_1, r_k)$$

Consiglio Nazionale delle Ricerche

ILLUSTRATIVE EXAMPLE





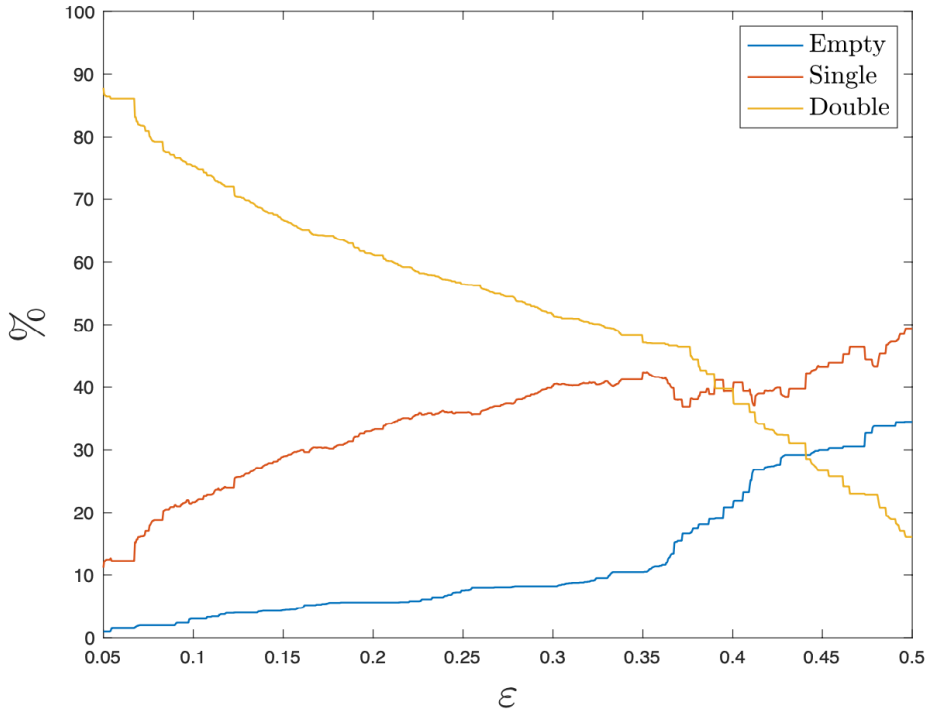


rule relevance increase
constant $\tau(\boldsymbol{x}, r_k)$

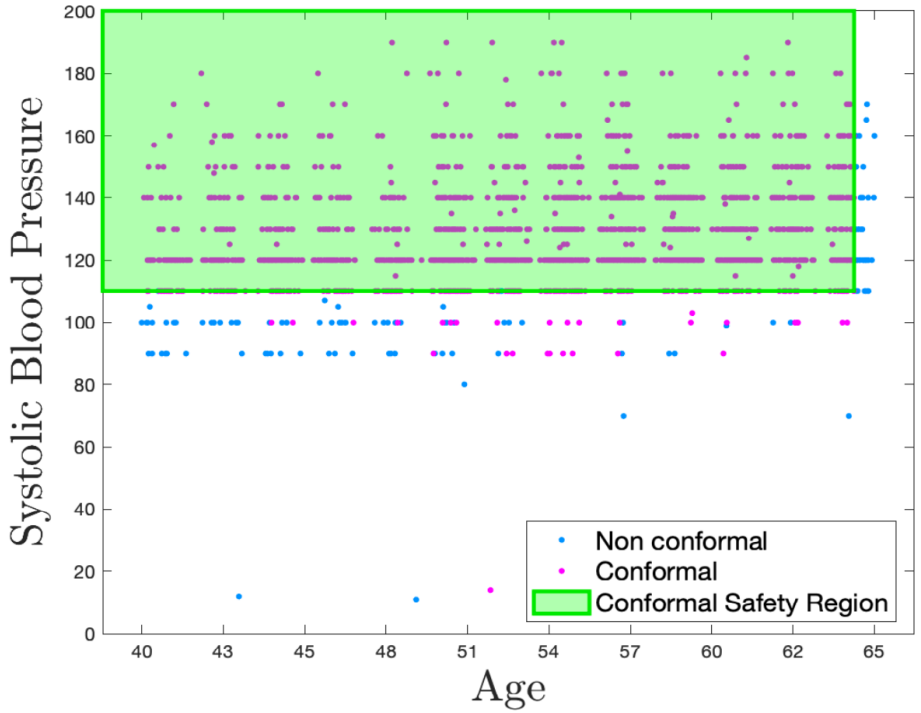
Consiglio Nazionale delle Ricerche

PRELIMINARY RESULTS

Efficiency:
average size of the prediction region



Conformal Safety Region



Dataset: https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

Consiglio Nazionale delle Ricerche

CONCLUSIONS

- New **score function** to perform CP on rule-based (transparent-by-design) XAI models in 2D, accounting for:
  - Performance ability
  - Geometry

- Step beyond pure CP:
  - definition of **conformal safety sets (CSS) and regions**

- Future (ongoing) works:
  - > 2D extension
  - Refinement of CSS definition towards safety preservation

Consiglio Nazionale delle Ricerche

12th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2023)
Limassol, 13-15/09/2023





THANK YOU!
Q&A Time