Provably Efficient and Robust Conformal Prediction under a Realistic Threat Model

Alberto Carlevaro[1,2,†] ALBERTO.CARLEVARO@AITEK.IT
Luca Oneto[3,†] LUCA.ONETO@UNIGE.IT
Davide Anguita[3] DAVIDE.ANGUITA@UNIGE.IT
Fabio Roli[3,4] FABIO.ROLI@{UNIGE,UNICA}.IT

[1] *Aitek S.p.A., Funded Research Department, Via della Crocetta 15, 16122 Genova, Italy.*
[2] *CNR-IEIIT, National Council of Research, Corso Duca degli Abruzzi 24, 10129, Turin, Italy.*
[3] *DIBRIS, University of Genoa, Via Opera Pia 11a/13, 16145, Genova, Italy.*
[4] *DIEE, University of Cagliari, Via Marengo, Cagliari, 09123, Italy.*
† *A. Carlevaro and L. Oneto contributed equally to the development of the article.*
(Corresponding author: A. Carlevaro.)

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

Robust conformal prediction is a model-agnostic technique designed to construct predictive sets with guaranteed coverage, assuming data exchangeability, even under adversarial attacks. Two primary strategies have been explored to address vulnerabilities to these attacks. The first strategy employs randomization, which is computationally efficient but fails to provide formal performance guarantees without resulting in overly conservative predictive sets. The second strategy involves formal verification, which restores coverage guarantees but leads to excessively conservative predictive sets and prohibitive computational overhead. Indeed, verification generally becomes NP-hard as it attempts to cope with attacks that are practically impossible, rendering some security claims unfalsifiable. In this paper, we propose a novel, provably efficient robust conformal prediction method by clearly defining a realistic threat model. Specifically, we assume explicit knowledge of the set of potential adversarial attacks, aligning our approach with standard certification procedures designed to certify against specific, identified threats. We demonstrate that attacks targeting the model can effectively be reframed as attacks on the score function, allowing us to recalibrate the score quantile to account for these known attacks and thereby restore desired coverage guarantees. It is worth noting that our approach allows to easily incorporate unknown or emerging (zero-day) attacks upon discovery, thus reestablishing coverage guarantees. By avoiding computationally intensive verification and operating under realistic threat assumptions, our approach achieves both efficiency and provable robustness. Empirical evaluations on real-world classification datasets and comparisons with state-of-the-art methods support the effectiveness and practicality of our proposed solution.

Keywords: Robust Conformal Prediction, Computational Efficiency, Provable Guarantees, Unfalsifiability, Realistic Threat Models.

1. Introduction

Machine Learning (ML) has experienced a significant acceleration in capabilities, driven by breakthroughs in statistical modeling, increased computational power, and abundant data

resources (Jordan and Mitchell, 2015). These advancements have enabled sophisticated applications across numerous domains, including healthcare (Zou et al., 2023), finance (Dixon et al., 2020), and transportation (Wang et al., 2023a), transforming decision-making processes and logistic strategies (Akbari and Do, 2021). As models grow increasingly accurate and scalable, their integration into diverse fields continues to catalyze far-reaching societal (Qian et al., 2024) and technological shifts (Bommasani et al., 2021). However, ML opens challenges that must be carefully addressed to ensure compliance with relevant technical (Paleyes et al., 2022), ethical (Toreini et al., 2020), and regulatory frameworks, such as the Data Act[1] and the proposed Europen AI Act[2].

The first issue (see Section 3.2), stemming from the statistical nature of ML, concerns the quality of predictions which, while correct on average (Vapnik, 1999), may not be accurate at a pointwise level (Amodei et al., 2016). Consequently, robust strategies for handling uncertainty are essential (Gawlikowski et al., 2023). Conformal Prediction (CP) (Vovk et al., 2005) offers a powerful, model-agnostic framework to quantify this uncertainty by constructing prediction sets, i.e., pointwise sets of plausible predictions, with guaranteed coverage under the minimal assumption of data exchangeability (Vovk, 2025). A second issue (see Section 3.3), first highlighted by (Biggio et al., 2013), is the vulnerability of ML models to small, carefully crafted perturbations of the input data, referred to as adversarial samples (Biggio and Roli, 2018). These perturbations are often imperceptible to humans (Biggio and Roli, 2018) or remain physically plausible (Kurakin et al., 2018), thereby exposing critical weaknesses in ML pipelines. Consequently, robust defense strategies are needed to ensure secure and reliable ML-based systems (Biggio and Roli, 2018; Cinà et al., 2023; Vaccari et al., 2022).

In this work, following a recently emerged line of research (Jeary et al., 2024; Lindemann et al., 2024; Gendler et al., 2021; Yan et al., 2024; Ghosh et al., 2023; Feldman et al., 2023; Carlevaro et al., 2024a,b), we address both issues concurrently by designing a CP framework that can handle adversarial samples. In fact, adversarial samples cause a distribution shift that breaks the exchangeability assumption of CP, resulting in leaks in its coverage guarantees (Tibshirani et al., 2019).

Current research efforts can be broadly categorized into two main classes of methods (see Section 2). The first class relies on randomization-based techniques, such as noise injection (Feldman et al., 2023) and conformal smoothing (Yan et al., 2024), aiming to enhance robustness by averaging over perturbations. While these approaches are computationally tractable and broadly applicable (Duchi et al., 2012; Cohen et al., 2019), they typically lack formal guarantees and often produce overly conservative prediction sets in practice (Gendler et al., 2021). The second class comprises verification-based methods (Jeary et al., 2024), which attempt to formally certify coverage by accounting for all possible adversarial perturbations (Wong and Kolter, 2018), including those that are computationally infeasible (Marro and Lombardi, 2023), thus rendering some security claims unfalsifiable (Herley, 2016). Although these methods restore theoretical guarantees (Jeary et al., 2024), they are frequently computationally intractable (Marro and Lombardi, 2023) and may be impractical or even inapplicable for complex architectures (Brix et al., 2023).

1. https://digital-strategy.ec.europa.eu/en/policies/data-act
2. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

In this paper, we propose a novel, *Provably Efficient and Robust Conformal Prediction* (PERCP) framework that tries to fill the gaps of the literature that we just described. Our approach count of two steps. The first step in our proposal is to clearly defining a realistic threat model (Xiong and Lagerström, 2019) (see Section 4.1). Specifically, we assume the attacker's goal is to perform an untargeted evasion attack (Biggio and Roli, 2018) to induce an ML-based classifier[3] into error. Moreover, we assume that the adversary has full knowledge bout the classifier but is able to use just known adversarial attacks (Madry et al., 2018; Goodfellow et al., 2014; Kurakin et al., 2016; Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017), aligning our approach with standard certification procedures (Stuurman and Lachaud, 2022) designed to certify against specific, identified threats. It is worth noting that our approach can seamlessly incorporate unknown or emerging (zero-day) attacks (Ahmad et al., 2023) as they are discovered, by simply adding them to the set of known adversarial attacks. The second step demonstrates that attacks targeting the model can effectively be reframed as attacks on the score function, allowing us to recalibrate the score quantile to account for these known attacks and thereby restore desired coverage guarantees (see Section 4.2). By avoiding computationally intensive verification and operating under realistic threat assumptions, our approach achieves both efficiency and provable robustness. To validate the effectiveness of PERCP, we conduct a series of experiments reported in Section 5. In the first set of experiments, using standard benchmark datasets (CIFAR10 (Krizhevsky, 2009), CIFAR100 (Krizhevsky, 2009) and TinyImageNet (Le and Yang, 2015), we compare PERCP with state-of-the-art approaches to CP under untargeted evasion attack, i.e., two randomization-based techniques (RSCP+ (Gendler et al., 2021) and RSCP+ (PTT) (Yan et al., 2024)) and two verification-based methods (VRCP-I (Jeary et al., 2024) and VRCP-C (Jeary et al., 2024)), comparing the coverage guarantees, computational requirements, and CP sets size. In the second sect of experiments, considering the same benchmarks dataset except for ImageNet (Deng et al., 2009) instead of TinyImageNet, we leverage RobustBench[4] to evaluate performance of PERCP across models with varying degrees of robustness. Results not only confirm the reliability PERCP, but also reveal a non-trivial relationship between model robustness and conformal prediction efficiency: the more robust the model, the smaller the resulting conformal prediction sets.

2. Related Work

In this section, we survey the state of the art on robust CP. We begin by outlining the two principal methodological paradigms, i.e., *randomization-based* (Gendler et al., 2021; Yan et al., 2024) and *formal-verification-based* (Jeary et al., 2024), and then highlight recent advances in probabilistic modeling (Ghosh et al., 2023) and label noise (Feldman et al., 2023). Although these latter lines of work are not directly comparable with the research presented here, they provide valuable context and round out our overview of the field.
Randomized smoothing (Duchi et al., 2012; Cohen et al., 2019; Salman et al., 2019) replaces the standard score function with a Gaussian-smoothed surrogate obtained via Monte Carlo sampling. This idea underlies *Randomly Smoothed Conformal Prediction* (RSCP) (Gendler

3. Due to space constraints we focus on classification but our approach can be readily generalized to other supervised problems like regression.
4. https://robustbench.github.io/.

et al., 2021), which inflates the conformal quantile according to the distribution of smoothed scores. Although RSCP empirically improves robustness, it often yields overly conservative prediction sets and was subsequently shown to lack formal coverage guarantees (Yan et al., 2024). To address this shortcoming, Yan et al. (2024) introduced RSCP+, which leverages Hoeffding's inequality to restore finite-sample validity. Despite its theoretical rigour, RSCP+ can still produce trivial—i.e., excessively large and thus uninformative—prediction sets. To mitigate this drawback, Yan et al. (2024) proposed a *post-training transformation* (PTT) that recalibrates the cumulative distribution function of the smoothed scores through a sigmoid transformation fitted on a hold-out validation set. This adjustment relaxes the conformal quantile, typically leading to smaller, more informative prediction sets. However, the efficiency gains are not guaranteed and depend critically on the size and representativeness of the hold-out data, as later observed by Jeary et al. (2024).

Formal-verification approaches seek to endow conformal prediction with *provable* robustness by certifying coverage for *all* perturbations within a prescribed threat set. The flagship method, *Verifiably Robust Conformal Prediction* (VRCP) (Jeary et al., 2024), appears in two flavours. *VRCP-I* performs neural-network verification at inference time, computing instance-wise *lower* bounds on the conformal score (best case). *VRCP-C* instead carries out verification during calibration, deriving *upper* bounds on the calibration scores (worst case) so that standard scores can be used at test time. Both variants draw on state-of-the-art verification machinery—interval-bound propagation (Zhang et al., 2020; Xu et al., 2021) and convex relaxations (Wong and Kolter, 2018). Despite their formal elegance, these techniques often struggle to scale to deep or highly non-linear networks (Brix et al., 2023; Marro and Lombardi, 2023). Furthermore, their guarantees hinge on threat models that can be restrictive or unrealistic, raising concerns about the falsifiability of the resulting claims (Herley, 2016). Then, a recent work from Zargarbashi et al. (2024) proposes a robust yet efficient conformal framework based on bounding the worst case change in conformity scores but marginal coverage is still an open problem being CP sets either over-covered or under-covered.

As a final remark, we highlight several recent studies that explore alternative robustness paradigms. Probabilistic Conformal Prediction (Ghosh et al., 2023) frames robustness through the lens of distributional uncertainty, whereas approaches such as those of Feldman et al. (2023) focus on resilience to label noise. Although these methods do not explicitly address adversarial robustness, they share the overarching aim of improving the reliability of conformal prediction under various forms of distribution shift.

3. Background Concepts

Before we present the theoretical contributions of this work, we introduce the fundamentals of multiclass classification, conformal prediction, and adversarial machine learning. This section also establishes the relevant notation and key concepts that will be used throughout the paper.

3.1. Multiclass Classification

Given an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and a finite output space $\mathcal{Y} = \{1, \dots, m\}$, a *classifier* (Shalev-Shwartz and Ben-David, 2014; Bishop and Bishop, 2023) is a function $f \in \mathcal{F} \subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$

that assigns a label $\hat{y} \in \mathcal{Y}$ to sample $\boldsymbol{x} \in \mathcal{X}$. Usually, this process is guided by a model's internal, such as estimated probabilities or logits, such that

$$\hat{y} \doteq f(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \boldsymbol{f}(\boldsymbol{x})_y, \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{x})_y \in \mathbb{R}$ is the y-th entry of the model's internal, $\boldsymbol{f} : \mathcal{X} \longrightarrow \mathbb{R}^m$. When probability estimates are needed, they can be obtained via the softmax function

$$\mathbb{f}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{f}(\boldsymbol{x})). \tag{2}$$

3.2. Conformal Prediction

For a classification task, we follow the inductive Conformal Prediction (CP) (Shafer and Vovk, 2008), which, given a learning data set of labeled exchangeable data (i.e., where the joint probability distribution is invariant under permutations) $\mathcal{D} \doteq \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$, partitions it into a proper training set $\mathcal{D}_\mathcal{T} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_t, y_t)\}$, used to fit ($\mathcal{D}_{\text{train}}$), validate ($\mathcal{D}_{\text{hold}}$) and test ($\mathcal{D}_{\text{test}}$) f, and a calibration set $\mathcal{D}_\mathcal{C} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_c, y_c)\}$, used to estimate the uncertainty of the predictor. $\mathcal{D}_\mathcal{C}$ is independent from $\mathcal{D}_\mathcal{T}$, namely $\mathcal{D} = \mathcal{D}_\mathcal{T} \cup \mathcal{D}_\mathcal{C}$ and $\mathcal{D}_\mathcal{T} \cap \mathcal{D}_\mathcal{C} = \oslash$, meaning it consists of unseen data that was not used in the fitting process of the model, ensuring unbiased estimation of uncertainty. Such uncertainty is then encoded by a negative-oriented (i.e., higher, more uncertain) *score function* $s : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which, given a *confidence level* $1 - \alpha \in (0, 1)$, is used to compute the $(1 - \alpha)$-empirical quantile of the score values in the calibration set

$$Q_{1-\alpha}(\boldsymbol{s}) \doteq \min \left[q : q \in \mathbb{R}, \frac{1}{c} \sum_{i=1}^{c} [s_i \leq q] \geq 1 - \alpha \right], \tag{3}$$

where $\boldsymbol{s} = [s(f, \boldsymbol{x}, y) : (\boldsymbol{x}, y) \in \mathcal{D}_\mathcal{C}]$ and where we use the Iverson bracket notation[5]. Then, for a given test sample $(\boldsymbol{x}_{n+1}, y_{n+1})$, the CP procedure constructs a conformal set

$$\mathcal{C}(\boldsymbol{x}_{n+1}) \doteq \{y : y \in \mathcal{Y}, s(f, \boldsymbol{x}_{n+1}, y) \leq Q_{1-\alpha}(\boldsymbol{s})\} \subseteq \mathcal{Y} \tag{4}$$

that satisfies the marginal coverage guarantee at *level of confidence* $1 - \alpha$, i.e.,

$$\mathbb{P}\{y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{n+1})\} \geq 1 - \alpha. \tag{5}$$

The CP procedure works with any classifier predictor and any choice of score function, but the statistical efficiency (e.g., the size of conformal sets , $|\mathcal{C}(\boldsymbol{x}_{n+1})|$) of CP is significantly influenced by the choice of the score. Following the theoretical description of conformity scores for classification of MAPIE[6] (Cordier et al., 2023), an open-source Python library for uncertainty quantification in machine learning using CP, we recall these examples of score function:

- **Least Ambiguous set-value Classifier (LAC)**, $s(f, \boldsymbol{x}, y) = 1 - \mathbb{f}(\boldsymbol{x})_y$, (Sadinle et al., 2019). It prioritizes selecting the most confident predictions by assigning lower scores to labels with higher (probabilistically estimated) model confidence: the lower the score the more confident the model in its prediction;

5. $[P] = 1$ if P is true, else 0.
6. https://mapie.readthedocs.io/en/latest/theoretical_description_classification.html

- **Top-K**, $s(f, \boldsymbol{x}, y) = j$ where $y = \pi_j$, and $\mathbb{f}(\boldsymbol{x})_{\pi_1} > \cdots > \mathbb{f}(\boldsymbol{x})_{\pi_j} > \cdots > \mathbb{f}(\boldsymbol{x})_{\pi_m}$, (Angelopoulos et al., 2020), where $\boldsymbol{\pi}$ is the permutation of $\mathcal{Y}$ that sorts in descending order $\mathbb{f}(\boldsymbol{x})_{\pi_i}$. The conformity score is simply the rank of the true label: as the confidence of the model on the label decreases, its rank increases (y moves lower in the ranking);
- **Adaptive Prediction Set (APS)**, $s(f, \boldsymbol{x}, y) = \sum_{i=1}^{j} \mathbb{f}(\boldsymbol{x})_{\pi_i}$, where $j = \pi_y$, (Romano et al., 2020; Angelopoulos et al., 2020). The conformity score is calculated by summing the ranked scores of the labels, starting from the highest and continuing until the true label is included, ensuring prediction sets that are, by construction, non-empty.

3.3. Adversarial Attacks

Adversarial ML (Biggio and Roli, 2018; Cinà et al., 2023) is a broad field of research which deals with the study and mitigation of adversarial threats that exploit vulnerabilities in ML models, aiming to degrade their performance or cause erroneous predictions. In this work, we will deals with the evasion attacks, namely small carefully crafted modifications to $\boldsymbol{x}$ that try to induce f into error. Formally the adversary has to search in a set of possible perturbation of $\boldsymbol{x}$, i.e., $\mathcal{P} : \mathbb{R}^d \to \{\subset \mathbb{R}^d\}$ such that $f(\tilde{\boldsymbol{x}}) \neq f(\boldsymbol{x})$ (untargeted evasion attack) where $\tilde{\boldsymbol{x}} \in \mathcal{P}(\boldsymbol{x})$. Sometimes the goal can be more challenging, and the adversary wants to force f predicting a particular wrong label $\tilde{y} \in \mathcal{Y}$ such that $\tilde{y} = f(\tilde{\boldsymbol{x}}) \neq f(\boldsymbol{x})$ (targeted evasion attack). $\mathcal{P}(\boldsymbol{x})$, in general, is not always easy to model as allowed perturbations may be not easy to define (Wong and Kolter, 2023). A common simplification is to define $\mathcal{P}(\boldsymbol{x})$ as an L_p ball of radius $\varepsilon \in \mathbb{R}$, i.e.,

$$\mathcal{P}(\boldsymbol{x}) = \mathcal{B}_{p,\varepsilon}(\boldsymbol{x}) \doteq \left\{ \boldsymbol{x}' : \boldsymbol{x}' \in \mathbb{R}^d, \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_p \leq \varepsilon \right\}. \tag{6}$$

In this setting, it is possible to formalize the untargeted and targeted evasion attacks, if exist, as follows

$$\tilde{\boldsymbol{x}} = \arg \min_{\boldsymbol{x}' \in \mathcal{B}_{p,\varepsilon}(\boldsymbol{x})} [f(\boldsymbol{x}') \neq f(\boldsymbol{x})], \quad \tilde{\boldsymbol{x}}_{\tilde{y}} = \arg \min_{\boldsymbol{x}' \in \mathcal{B}_{p,\varepsilon}(\boldsymbol{x})} [\tilde{y} = f(\boldsymbol{x}') \neq f(\boldsymbol{x})], \tag{7}$$

In general, these attacks are computationally challenging, in particular NP-hard problems (Marro and Lombardi, 2023). For this reason, different heuristics have been proposed (Biggio and Roli, 2018). The most effective ones rely on the fact that $\boldsymbol{f}$ is, by construction, differentiable with respect to $\boldsymbol{x}$ as the majority of the state of the art models are trained with gradient-based methods (Bishop and Bishop, 2023). Then, Problems (7) are reformulated as

$$\tilde{\boldsymbol{x}} = \arg \min_{\boldsymbol{x}' \in \mathcal{B}_{p,\varepsilon}(\boldsymbol{x})} \ell(f(\boldsymbol{x}'), y), \tag{8}$$

where ℓ is chosen, differentiable, based on the attack type. For example, to get $\tilde{\boldsymbol{x}}$ for the untargeted and targeted attacks of Problems (7), one can set $\ell(f(\boldsymbol{x}'), y) = \boldsymbol{f}(\boldsymbol{x}')_y$ and $\ell(f(\boldsymbol{x}'), y) = -\boldsymbol{f}(\boldsymbol{x}')_{\tilde{y}}$ respectively[7]. Nevertheless, many other ℓ have been proposed in the literature like (Carlini and Wagner, 2017) which leverages a margin-based loss function

7. From now on, we will focus on untargeted attacks for space constraints even if it can be easy generalized to targeted attacks

or (Kullback, 1951) which uses the cross-entropy loss or the Kullback–Leibler divergence. Then, Problem (8) can be addressed with all the tools for gradient based optimization available in the literature (Bishop and Bishop, 2023) plus a simple projection on the L_p ball (Biggio and Roli, 2018). As a consequence, any solution to Problems (7) will lead us, in general, to a sub-optimal solutions. This means that, based on the f, ℓ, and $\boldsymbol{x}$, the optimization algorithm $\mathtt{O}_{\boldsymbol{\rho}}$ (characterized by a set of parameters $\boldsymbol{\rho}$, used to solve Problems (8), namely the realistic/practical attack) will generate a set of reachable perturbation $\hat{\mathcal{B}}_{p,\varepsilon}(\boldsymbol{x}) \subseteq \mathcal{B}_{p,\varepsilon}(\boldsymbol{x})$, that is the one we can actually explore. Testing multiple ℓ and $\mathtt{O}_{\rho}$, (i.e., again, multiple realistic/practical attacks) and taking the most effective one, it may lead to increase $\hat{\mathcal{B}}_{p,\varepsilon}(\boldsymbol{x})$ toward $\mathcal{B}_{p,\varepsilon}(\boldsymbol{x})$ but, in practice, without never reaching the entire ball, i.e., $\hat{\mathcal{B}}_{p,\varepsilon}(\boldsymbol{x}) \subset \mathcal{B}_{p,\varepsilon}(\boldsymbol{x})$. This stands in contrast to verification algorithms, which consider all perturbations within the L_p-ball, including those that are computationally unreachable via optimization, resulting in an approach overly pessimistic. As a consequence, we define a realistic/practical attack as an operator

$$\mathtt{A}_{\boldsymbol{\theta}} : \mathcal{F} \times \mathcal{X} \to \mathbb{R}^d, \tag{9}$$

that, given a set of parameters $\boldsymbol{\theta}$ (i.e., ℓ and $\mathtt{O}_{\boldsymbol{\rho}}$), tries to perform the evasion attack solving Problem (8). In the following, it is reported a list of the most used realistic/practical untargeted evasion attacks that can be represented by (9) and that we will actually use in the paper as benchmark examples to test the robustness of our CP procedure:

- **Projected Gradient Descent (PGD)** (Madry et al., 2018). The loss for PGD is $\ell(f(\boldsymbol{x}), y) = \mathbb{f}(\boldsymbol{x})_y$. Then, starting from $\boldsymbol{x}$, a gradient step of size γ is performed, project onto the $\mathcal{B}_{p,\varepsilon}$, and this procedure is repeated for T iterations. As a consequence, the set of hyperparameters for PGD is given by $\boldsymbol{\theta} = \{T, \gamma, \varepsilon, p\}$.
- **Fast Gradient Sign Method (FGSM)** (Goodfellow et al., 2014). FGSM perturbs the input by moving it in the direction of the gradient of $\ell(f(\boldsymbol{x}), y) = \mathbb{f}(\boldsymbol{x})_y$, scaled by a fixed perturbation budget ε. The direction is determined by the sign of the gradient (for $p = \infty$) or normalized by the gradient's norm (for $p < \infty$). The set of parameters for FGSM is $\boldsymbol{\theta} = \{\varepsilon, p\}$.
- **Basic Iterative Method (BIM)** (Kurakin et al., 2016). BIM iteratively minimizes $\ell(f(\boldsymbol{x}), y) = \mathbb{f}(\boldsymbol{x})_y$ by applying FGSM multiple times, with each perturbed sample being projected back into the L_p-ball after each update. This process continues for T iterations, where each step γ moves the input in the direction of the gradient, scaled by ε. The parameters are $\boldsymbol{\theta} = \{T, \gamma, \varepsilon, p\}$.
- **DeepFool** (Moosavi-Dezfooli et al., 2016). DeepFool minimizes $\ell(f(\boldsymbol{x}), y) = \mathbb{f}(\boldsymbol{x})_y$ by iteratively computing the smallest perturbation needed to cross the decision boundary. At each step, it adjusts $\boldsymbol{x}$ in the direction that reduces the margin between the target class and the most competitive alternative within the L_p-ball, ensuring minimal perturbation. This process continues for T iterations with a step-size of γ, with parameters $\boldsymbol{\theta} = \{T, \gamma, \varepsilon, p\}$.
- **Carlini-Wagner (CW)** (Carlini and Wagner, 2017). For untargeted attacks, the CW algorithm searches for an adversarial example $\tilde{\boldsymbol{x}} \in \mathcal{B}_{2,\varepsilon}(\boldsymbol{x})$ that minimizes the model's confidence in the correct class y. Starting from $\boldsymbol{x}$, a gradient step of size γ is performed to reduce $\mathbb{f}(\tilde{\boldsymbol{x}})_y$ and this procedure is repeated T times. The set of

hyperparameters for CW is given by $\boldsymbol{\theta} = \{\gamma, T, c, \varepsilon\}$, where c balances the trade-off between perturbation size and misclassification.

For a complete review of the attacks please refer to (Biggio and Roli, 2018; Pitropakis et al., 2019).

4. Provably Efficient and Robust Conformal Prediction (PERCP)

In this section, we introduce the main contribution of our work: a methodology to construct *Provably Efficient and Robust Conformal Prediction* (PERCP) sets $\tilde{\mathcal{C}} \subseteq \mathcal{Y}$, which satisfies the following *robust marginal coverage* guarantee

$$\mathbb{P}\left\{y_{n+1} \in \tilde{\mathcal{C}}(\tilde{\boldsymbol{x}}_{n+1})\right\} \geq 1 - \alpha, \tag{10}$$

under a realistic threat model. We first define this threat model and then present the PERCP methodology.

4.1. Threat Model

In this section, we introduce the threat model addressed in this work, namely a systematic description of how an adversary could attack the CP sets. Specifically, threat model describes the adversary's goals, knowledge, and capabilities.

4.1.1. Attacker's Goal

The adversary's goal is to perform an untargeted evasion attack, namely design a small carefully crafted modifications to the input that try to induce the classifier into error (see Section 3.3). As a consequence, the exchangeability assumption in CP is violated, leading to a breach of the marginal coverage guarantee of classical CP (see Section 3.2). In other words, the attacker goal is to implicitly attack the CP sets by performing an actual evasion attack on the classifier.

4.1.2. Attacker's Knowledge and Capability

We assume that the adversary has full knowledge, i.e., white box attacks, about the classifier, i.e., architecture and associated parameters. However, the adversary is able to use just known attacks $\mathcal{A}$, namely attacks, and relative parameters, that are available in the literature. In other words, given a classifier f the attacker can select (one or the best performing) $\mathtt{A}_{\boldsymbol{\theta}} \in \mathcal{A}$.

Note that, this assumption does not cover unknown or emerging (zero-day) attacks, such threats can be swiftly incorporated (patched) into our framework upon discovery, incorporating it into $\mathcal{A}$.

Note also that our assumption avoids unfalsifiable security claims. Indeed, most works in the literature focus on providing guarantees against every possible $\mathcal{P}(\boldsymbol{x})$, e.g., $\mathcal{B}_{p,\varepsilon}(\boldsymbol{x})$ (Madry et al., 2018; Wong and Kolter, 2018; Cohen et al., 2019), including those that cannot be implemented in practice (Gilmer et al., 2018) (because they are computationally infeasible) making such guarantees unfalsifiable.

Note finally, that, theoretically $\mathcal{A}$ may be very large (or even infinite dimensional) as, e.g., some parameters of the attacks are real numbers. Nevertheless, it is well known that practically this is not the case as some attacks configurations are ineffective or computationally unfeasible (Biggio and Roli, 2018; Marro and Lombardi, 2023; Muthalagu et al., 2025).

4.2. Statistical Guarantees

In this section we will present our PERCP for the threat model described in Section 4.1.

With this goal in mind, we first observe that

(i) most of the untargeted evasion attacks focus on minimizing, implicitly or explicitly, $\mathbb{f}(\boldsymbol{x})_y$ (see Section 3.3). This means that the attack, will always worsen $\mathbb{f}(\boldsymbol{x})_y$ until, when attacks is successful as, e.g., ε is large enough, actually change the classification;

(ii) most of the score functions are monotonic decreasing (i.e., negative-oriented) in $\mathbb{f}(\boldsymbol{x})_y$ (see Section 3.2).

It thus becomes clear that the attacker's goal to perform an actual evasion attack on the classifier implicitly targets also the CP set. In fact, when we perform an attack $\tilde{\boldsymbol{x}} = \mathtt{A}_{\boldsymbol{\theta}}(f, \boldsymbol{x})$, with any of the attacks of Section 3.3, we will have that

$$\mathbb{f}(\tilde{\boldsymbol{x}})_y \leq \mathbb{f}(\boldsymbol{x})_y. \tag{11}$$

As a consequence, for the LAC (see Section 3.2), the score function increases

$$\mathbb{f}(\tilde{\boldsymbol{x}})_y \leq \mathbb{f}(\boldsymbol{x})_y \Rightarrow 1 - \mathbb{f}(\tilde{\boldsymbol{x}})_y \geq 1 - \mathbb{f}(\boldsymbol{x})_y \Rightarrow s(f, \tilde{\boldsymbol{x}}, y) \geq s(f, \boldsymbol{x}, y). \tag{12}$$

While this property is immediate for LAC, it reflects a more general principle that negative-oriented score functions are expected to satisfy: the score should increase under adversarial attacks. Intuitively, this captures the idea that adversarial perturbations degrade the model's prediction, so the agreement between the true label y and the perturbed input $\tilde{\boldsymbol{x}}$ cannot be better than that with the original, unperturbed input $\boldsymbol{x}$.

Beyond LAC, where this statement is clear, also the Top-k and APS score functions (see Section 3.2) satisfy this property. In the case of the Top-k score, after the attack, the rank of $\mathbb{f}(\boldsymbol{x})_y$ in the sorted list of class scores can only stay the same or worsen, meaning that the true label y moves lower in the ranking (i.e., the rank increases). For the APS score, if the true label y remains within the top-k after the attack, then the score remains the sum of the top-k class probabilities, which, following the same behavior as Top-k score, does not decrease $s(f, \tilde{\boldsymbol{x}}, y) = \sum_{j=1}^{k} \mathbb{f}(\tilde{\boldsymbol{x}})_{\pi_j} \geq \sum_{j=1}^{k} \mathbb{f}(\boldsymbol{x})_{\pi_j} = s(f, \boldsymbol{x}, y)$. If y drops out of the top-k, the resulting score no longer includes the (decreased) value $\mathbb{f}(\tilde{\boldsymbol{x}})_y$, and instead includes class probabilities associated with incorrect labels, which are higher. Therefore, the score increases further, and the property still holds.

We can now formalize these concepts in the following property, which holds for any combination of evasion attacks and score function reported in Sections 3.3 and 3.2.

Property 1 (Evasion Attacks Imply Higher Scores):
Let us consider a classifier f and a sample $\boldsymbol{x} \in \mathcal{X}$ and associate true label $y \in \mathcal{Y}$. For any evasion attack $\tilde{\boldsymbol{x}} = \mathtt{A}_{\boldsymbol{\theta}}(f, \boldsymbol{x})$ that, implicitly or explicitly, minimizes $\mathbb{f}(\boldsymbol{x})_y$ and for any monotonic increasing in $\mathbb{f}(\boldsymbol{x})_y$ score function $s(f, \boldsymbol{x}, y)$ we have that

$$\mathbb{f}(\tilde{\boldsymbol{x}})_y \leq \mathbb{f}(\boldsymbol{x})_y \Rightarrow s(f, \tilde{\boldsymbol{x}}, y) \geq s(f, \boldsymbol{x}, y). \tag{13}$$

Proof By definition of monotonic decreasing in $\mathbb{f}(\boldsymbol{x})_y$ score function we have that $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ such that $\mathbb{f}(\boldsymbol{x}_1)_y \leq \mathbb{f}(\boldsymbol{x}_2)_y$ we have that $s(f, \boldsymbol{x}_1, y) \geq s(f, \boldsymbol{x}_2, y)$. By definition, given $\tilde{\boldsymbol{x}} = \mathtt{A}_{\boldsymbol{\theta}}(f, \boldsymbol{x})$ we have that $\mathbb{f}(\tilde{\boldsymbol{x}})_y \leq \mathbb{f}(\boldsymbol{x})_y$. From these two observations Eq. (13) comes straightforward. ■

Under the threat model of Section 4.1 and Property 1 we can now report the main result of this paper, our PERCP.

Theorem 1 *Let us consider a classifier f and a perturbed sample $\tilde{\boldsymbol{x}}_{n+1}$ and associated, unknown, original sample $\boldsymbol{x}_{n+1} \in \mathcal{X}$ and true label $y_{n+1} \in \mathcal{Y}$ under the threat model of Section 4.1, namely*

$$\tilde{\boldsymbol{x}}_{n+1} = \mathtt{A}_{\boldsymbol{\theta}}(f, \boldsymbol{x}_{n+1}), \quad \mathtt{A}_{\boldsymbol{\theta}} \in \mathcal{A}. \tag{14}$$

Let us define the following quantities

$$\mathcal{P}(\boldsymbol{x}_{n+1}) = \{\mathtt{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_{n+1}) : \mathtt{A}_{\boldsymbol{\theta}} \in \mathcal{A}\}, \tag{15}$$

$$\boldsymbol{x}_{\mathcal{A}}(\boldsymbol{x}_{n+1}) = \arg \min_{\tilde{\boldsymbol{x}} \in \mathcal{P}(\boldsymbol{x}_{n+1})} \mathbb{f}(\tilde{\boldsymbol{x}})_y, \tag{16}$$

$$\boldsymbol{s}_{\mathcal{A}} = [s(f, \boldsymbol{x}_{\mathcal{A}}(\boldsymbol{x}), y) : (\boldsymbol{x}, y) \in \mathcal{D}_{\mathcal{C}}]. \tag{17}$$

Then for any set of possible attacks $\mathcal{A}$ and score functions that satisfy Property 1 it is possible to prove that

$$\mathbb{P}\left\{y_{n+1} \in \tilde{\mathcal{C}}(\tilde{\boldsymbol{x}}_{n+1})\right\} \geq 1 - \alpha, \quad \forall \mathtt{A}_{\boldsymbol{\theta}} \in \mathcal{A}, \tag{18}$$

where

$$\tilde{\mathcal{C}}(\tilde{\boldsymbol{x}}_{n+1}) = \{y : y \in \mathcal{Y}, s(f, \tilde{\boldsymbol{x}}_{n+1}, y) \leq Q_{1-\alpha}(\boldsymbol{s}_{\mathcal{A}})\}. \tag{19}$$

Proof As first step, note that $\boldsymbol{x}_{n+1}$ and y_{n+1} are unknown and the defender knows just $\tilde{\boldsymbol{x}}_{n+1}$ and that it is generated by applying one or more attacks from the set $\mathcal{A}$.
To defend against the worst-case scenario, we must consider that the adversary could choose the most harmful among all potential perturbations, including the clean input. The worst-case adversarial sample is then defined as the one that minimizes the classifier's confidence in the true label over all perturbations in $\mathcal{P}(\boldsymbol{x}_{n+1})$, as in Eq. (16). Then, from Property 1 we have

$$s(f, \boldsymbol{x}_{\mathcal{A}}(\boldsymbol{x}_{n+1}), y) \geq s(f, \tilde{\boldsymbol{x}}_{n+1}, y). \tag{20}$$

This means that the distribution of the score function corresponding to $(\boldsymbol{x}_{\mathcal{A}}(\boldsymbol{x}), y)$ is the worst possible one among all the ones that can be generated from the one of $(\boldsymbol{x}, y)$ under the set of attacks $\mathcal{A}$. As a consequence we can state that

$$\mathbb{P}\left\{y_{n+1} \in \tilde{\mathcal{C}}(\tilde{\boldsymbol{x}}_{n+1})\right\} = \mathbb{P}\left\{s(f, \tilde{\boldsymbol{x}}_{n+1}, y_{n+1}) \leq Q_{1-\alpha}(\boldsymbol{s}_{\mathcal{A}})\right\} \tag{21}$$

$$\geq \mathbb{P}\left\{s(f, \boldsymbol{x}_{\mathcal{A}}(\boldsymbol{x}_{n+1}), y_{n+1}) \leq Q_{1-\alpha}(\boldsymbol{s}_{\mathcal{A}})\right\} \tag{22}$$

$$\geq 1 - \alpha, \tag{23}$$

where the last inequality is the classical CP results (Papadopoulos et al., 2002) on the worst case distribution. ■

Theorem 1 demonstrates that by constructing a worst-case distribution of attacked samples, derived from the set of attacks and the exchangeable samples of the original distribution, we can recalibrate the quantile to account for the worst-case scenario. Specifically, by enumerating admissible adversarial perturbations, attacks on the classifier reduce to attacks on the score function. This leads to a closed-form recalibration of the conformal quantile on the attacked score distribution, which effectively restores finite-sample coverage that randomized or verification-based schemes either relax or secure only at prohibitive computational cost.

Note finally, that, in our setting, there is still a computational asymmetry between defender an attacker, as noted in (Marro and Lombardi, 2023), even though this asymmetry is less pronounced. In fact, the attacker has to test all $\mathtt{A}_{\boldsymbol{\theta}} \in \mathcal{A}$ to try maximize its effectiveness. The defender, instead, need to perform all the attacks $\mathtt{A}_{\boldsymbol{\theta}} \in \mathcal{A}\ \forall(\boldsymbol{x}, y) \in \mathcal{D}_{\mathcal{C}}$. This means that computational cost of the defense is linearly, in the cardinality of $\mathcal{D}_{\mathcal{C}}$, more expensive than the attack.

5. Experiments

In this section, we present the results of applying PERCP to a diverse collection of datasets and models subjected to adversarial evasion attacks under a variety of configurations. By systematically varying key parameters, we assess its performance with respect to (i) marginal coverage of CP, (ii) the average size of the resulting prediction sets, and (iii) the mean computation time required to construct these CP sets. Moreover, we compare PERCP with state-of-the-art methods for robust conformal prediction. All the experiments were run on a server with 2x Intel Xeon Silver 4214R (24 cores, 24 threads, 2.4GHz), 94GB of RAM, and an NVIDIA GeForce RTX 4090 24GB GPU and the code to reproduce the experiments can be retrieved from our public repository[8].

Table 1 reports the marginal coverage, average prediction–set size, and average computation time on CIFAR10, CIFAR100, and TinyImageNet for each method, Vanilla (vanilla CP without caring about the adversary), RSCP+, RSCP+ (PTT), VRCP-I, VRCP-C, and PERCP, using exactly the experimental protocol of Table 1 in Jeary et al. (2024) (in particular, $|\mathcal{D}_{\mathcal{C}}| = 4,500$ and $|\mathcal{D}_{\text{test}}| = 5,000$). The only additional hyperparameter we must specify concerns the attack set for PERCP: we use projected-gradient descent (PGD) constrained to an L_2 ball with $T = 100$, $\gamma = 1/255$, and $\varepsilon = 0.02$ for CIFAR100 and TinyImageNet, and $\varepsilon = 0.03$ for CIFAR10 (this is exactly what has been done in Jeary et al. (2024)). Recall that Jeary et al. (2024) employs a comparatively lightweight architecture, since the verification routines of VRCP-I and VRCP-C do not scale to deeper networks. From Table 1, we observe that PERCP is both provable - its marginal coverage consistently remains near the target level $1 - \alpha = 0.90$ - and efficient - yielding prediction sets that are, on average, smaller than those produced by state-of-the-art baselines and comparable to those of the Vanilla procedure. Regarding computation time, PERCP behaves as expected: it matches the runtime of Vanilla while markedly outperforming the other methods. Finally, PERCP is

8. https://github.com/AlbiCarle/PERCP

Table 1: Marginal coverage, average set size, and average time on CIFAR10, CIFAR100, and TinyImageNet for the different methods (Vanilla, RSCP+, RSCP+(PTT), VRCP-I, VRCP-C, and PERCP) under the same setting of Table 1 in Jeary et al. (2024).

Method	Marginal Coverage	Average Set Size	Average Time (s)
CIFAR10			
Vanilla	0.89 ± 0.01	1.75 ± 0.07	0.01 ± 0.02
RSCP+	1.00 ± 0.00	10.00 ± 0.00	0.10 ± 0.30
RSCP+ (PTT)	0.96 ± 0.02	5.92 ± 2.22	0.11 ± 0.33
VRCP-I	0.99 ± 0.00	4.51 ± 0.06	0.14 ± 0.03
VRCP-C	1.00 ± 0.00	5.06 ± 0.09	0.14 ± 0.01
PERCP	0.90 ± 0.03	$\mathbf{1.67 \pm 0.79}$	0.01 ± 0.02
CIFAR100			
Vanilla	0.88 ± 0.07	7.82 ± 0.27	0.08 ± 0.02
RSCP+	1.00 ± 0.00	100.00 ± 0.00	0.14 ± 0.11
RSCP+ (PTT)	0.93 ± 0.03	37.91 ± 26.56	0.14 ± 0.11
VRCP-I	0.98 ± 0.00	25.13 ± 0.53	0.18 ± 0.10
VRCP-C	0.99 ± 0.00	27.96 ± 1.73	0.18 ± 0.10
PERCP	0.90 ± 0.07	$\mathbf{4.17 \pm 0.99}$	0.08 ± 0.07
TinyImageNet			
Vanilla	0.88 ± 0.01	35.91 ± 2.00	0.15 ± 0.02
RSCP+	1.00 ± 0.00	200.00 ± 0.00	0.19 ± 0.02
RSCP+ (PTT)	0.93 ± 0.03	88.63 ± 47.65	0.21 ± 0.02
VRCP-I	0.96 ± 0.01	69.57 ± 2.44	0.22 ± 0.02
VRCP-C	0.97 ± 0.01	77.83 ± 3.42	0.22 ± 0.01
PERCP	0.91 ± 0.02	$\mathbf{12.28 \pm 1.36}$	0.15 ± 0.04

model-agnostic; it makes no assumptions about the architecture of the underlying classifier and therefore scales gracefully to the more complex models evaluated in Table 2. In contrast, verification-based algorithms incur prohibitive—sometimes even infeasible—computational costs as model complexity grows: even for the comparatively simple architectures examined by Jeary et al. (2024), VRCP can require up to $10\times$ the runtime of PERCP.

Figure 1 reports the marginal coverage and average set size in exactly the same settings of Table 1, but varying ε, for RSCP+, RSCP+(PTT), VRCP-I, VRCP-C, and PERCP. In the case of RSCP+, RSCP+(PTT), VRCP-I, and VRCP-C the adversarial samples have been generated as described in (Jeary et al., 2024) using PGD100, namely PGD with $T = 100$, and $\gamma = 1/255$. In the case of PERCP, we report the case in which the set of attacks is PGD1, PGD10, PGD100, or the Best of them with $\gamma = 1/255$.

From Figure 1 we observe that the marginal coverage of PERCP stays consistently close to the target level $1 - \alpha = 0.90$, whereas the competing methods rapidly saturate at 1.0. As confirmed by the accompanying plot of average set size, this saturation leads to uninformative - i.e., trivial - conformal sets. In contrast, PERCP produces sets whose size grows smoothly with the attack magnitude ε: the stronger the attack, the larger the conformal sets, faithfully reflecting the increased uncertainty (note that the blue curve, corresponding

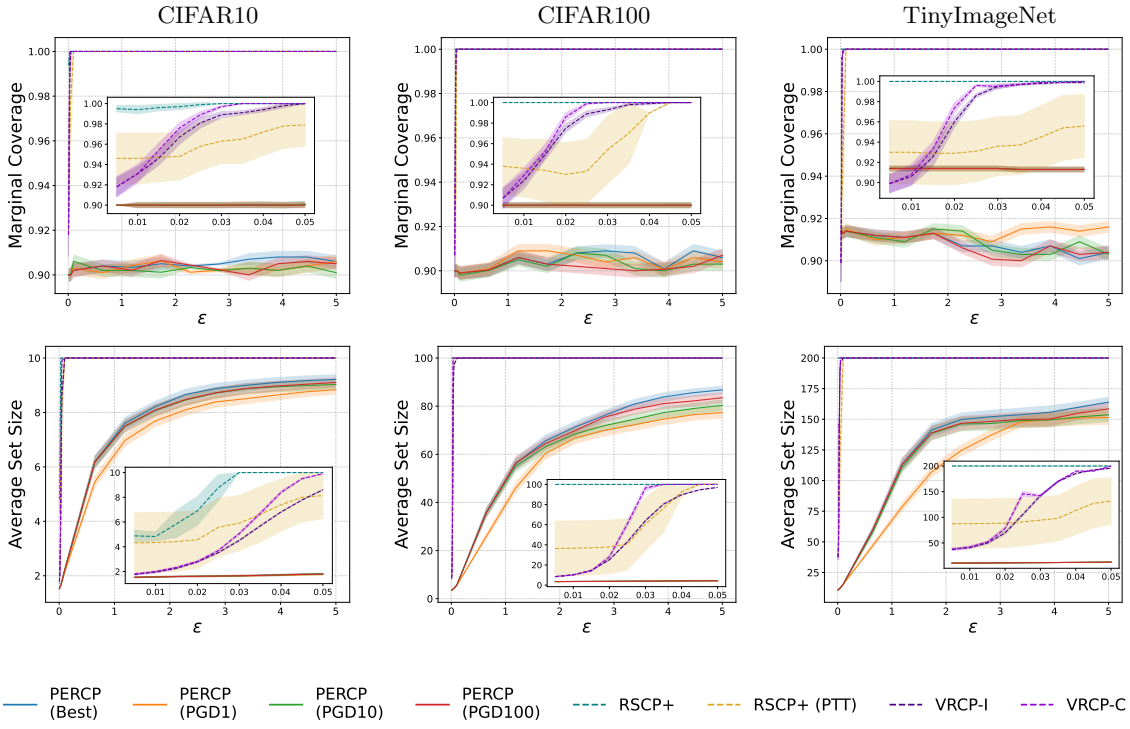

Figure 1: Marginal coverage and average set size for an L_2 ball attack of varing radius ε of RSCP+, RSCP+(PTT), VRCP-I, VRCP-C, and PERCP. In the case of RSCP+, RSCP+(PTT), VRCP-I, and VRCP-C the adversarial samples have been generated as described in (Jeary et al., 2024) using PGD100, namely PGD with $T = 100$, and $\gamma = 1/255$. In the case of PERCP, we report the case in which the set of attacks is PGD1, PGD10, PGD100, or the Best of them with $\gamma = 1/255$.

to the most effective attack, is the highest). For the other methods, however, the average set size quickly reaches the maximum permitted value (10, 100, and 200 for CIFAR10, CIFAR100, and TinyImageNet respectively), again resulting in uninformative predictions. For clarity, Figure 1 also provides a zoomed view of the interval $\varepsilon \in [0.01, 0.05]$ - the range considered in (Jeary et al., 2024) - where one can see how the state-of-the-art curves rise and promptly saturate at their upper bounds in both marginal coverage and average set size.

Finally, Table 2 reports the performance marginal coverage and average set size on CIFAR10, CIFAR100, and ImageNet, for PERCP for different L_p balls. We report the case in which the set of attacks is PGD ($T = 100$ and $\gamma = \varepsilon/T$), FGSM, BIM ($T = 100$ and $\gamma = \varepsilon/T$), DeepFool ($T = 100$ and $\gamma = \varepsilon/T$), CW ($c = 0.5$, $T = 100$ and $\gamma = \varepsilon/T$), and the Best of them. We tested the most and least reliable model in RobustBench[4] with the corresponding configurations of ε. In order to generate the attacks we used the SecML[9] library. Note that, CW is not implemented in SecML for the L_∞ attack.

9. https://secml.readthedocs.io/

Table 2: Marginal coverage and average set size on CIFAR10, CIFAR100, and ImageNet, for PERCP for different L_p balls. We report the case in which the set of attacks is PGD ($T = 100$ and $\gamma = \varepsilon/T$), FGSM, BIM ($T = 100$ and $\gamma = \varepsilon/T$), DeepFool ($T = 100$ and $\gamma = \varepsilon/T$), CW ($c = 0.5$, $T = 100$ and $\gamma = \varepsilon/T$), and the Best of them. We tested the most and least reliable model in RobustBench[4] (reporting, in order, the model's name, the model's architecture and the citation, if any) with the corresponding configurations of ε.

Attack	CIFAR10		CIFAR100		ImageNet	
	Marginal Coverage	**Average Set Size**	**Marginal Coverage**	**Average Set Size**	**Marginal Coverage**	**Average Set Size**
L_2 - $\varepsilon = 0.5$						
	Most Robust Model					
	Wang2023Better WRN-70-16 (Wang et al., 2023b)		Diffender2021_Winning_LRR WRN-18-2 (Diffenderfer et al., 2021)		Tian2022Deeper DeiT Base (Tian et al., 2022)	
PGD	0.90 ± 0.03	1.00 ± 0.00	0.90 ± 0.01	1.21 ± 0.05	0.90 ± 0.24	12.76 ± 2.37
FGSM	0.90 ± 0.02	1.00 ± 0.00	0.90 ± 0.03	1.16 ± 0.04	0.90 ± 0.25	10.51 ± 1.64
BIM	0.90 ± 0.04	1.00 ± 0.00	0.90 ± 0.03	1.04 ± 0.02	0.90 ± 0.22	4.34 ± 0.49
DeepFool	0.90 ± 0.02	1.00 ± 0.00	0.90 ± 0.00	1.19 ± 0.05	0.90 ± 0.24	2.09 ± 0.11
CW	0.90 ± 0.02	1.92 ± 0.02	0.90 ± 0.02	2.14 ± 0.08	0.91 ± 0.25	3.72 ± 0.25
Best	0.90 ± 0.02	1.92 ± 0.03	0.90 ± 0.03	2.59 ± 0.01	0.92 ± 0.24	18.72 ± 6.09
	Least Robust Model					
	Standardly trained WRN-28-10		Gowal2022Uncovering WRN-70-16 (Gowal et al., 2020)		AlexNet AlexNet (Krizhevsky et al., 2012)	
PGD	0.91 ± 0.01	9.27 ± 0.20	0.90 ± 0.03	2.22 ± 0.12	0.91 ± 0.27	19.83 ± 1.63
FGSM	0.89 ± 0.03	5.97 ± 0.32	0.90 ± 0.01	1.88 ± 0.11	0.90 ± 0.29	13.70 ± 3.59
BIM	0.92 ± 0.03	9.08 ± 0.22	0.90 ± 0.00	1.24 ± 0.05	0.90 ± 0.29	32.59 ± 3.23
DeepFool	0.88 ± 0.03	6.28 ± 0.32	0.91 ± 0.03	1.52 ± 0.07	0.90 ± 0.29	16.94 ± 1.29
CW	0.90 ± 0.03	8.11 ± 0.28	0.92 ± 0.00	1.24 ± 0.05	0.90 ± 0.30	20.93 ± 1.47
Best	0.91 ± 0.03	9.12 ± 0.22	0.91 ± 0.02	3.61 ± 0.13	0.91 ± 0.27	15.02 ± 2.60
L_∞ - $\varepsilon = 8/255$ for CIFAR10 and CIFAR10 - $\varepsilon = 4/255$ for ImageNet						
	Most Robust Model					
	Wang2023Better WRN-70-16 (Wang et al., 2023b)				Liu2023Comprehensive Swin-L (Liu et al., 2023)	
PGD	0.90 ± 0.08	6.30 ± 0.27	0.92 ± 0.00	9.75 ± 0.53	0.90 ± 0.02	8.35 ± 0.76
FGSM	0.90 ± 0.03	2.53 ± 0.11	0.92 ± 0.00	2.15 ± 0.12	0.90 ± 0.02	8.12 ± 0.76
BIM	0.91 ± 0.03	6.94 ± 0.27	0.90 ± 0.04	13.93 ± 0.73	0.90 ± 0.01	7.54 ± 0.69
DeepFool	0.90 ± 0.03	2.20 ± 0.09	0.90 ± 0.03	1.44 ± 0.06	0.90 ± 0.02	3.23 ± 0.22
Best	0.90 ± 0.03	7.08 ± 0.28	0.91 ± 0.00	15.39 ± 0.80	0.90 ± 0.01	8.40 ± 0.78
	Least Robust Model					
	Standardly trained WRN-28-10		Rice2020Overfitting PreActResNet-18 (Rice et al., 2020)		Standardly trained ResNet-50	
PGD	0.92 ± 0.08	9.14 ± 0.20	0.90 ± 0.03	34.97 ± 2.12	1.00 ± 0.00	1000 ± 0.00
FGSM	0.93 ± 0.03	8.57 ± 0.22	0.90 ± 0.03	14.63 ± 0.84	0.90 ± 0.03	499.42 ± 39.29
BIM	1.00 ± 0.00	10.00 ± 0.00	0.90 ± 0.03	32.75 ± 2.05	0.90 ± 0.02	540.51 ± 40.22
DeepFool	0.93 ± 0.02	8.58 ± 0.22	0.90 ± 0.03	7.60 ± 0.43	0.91 ± 0.02	38.98 ± 4.91
Best	1.00 ± 0.00	10.00 ± 0.00	0.90 ± 0.03	36.55 ± 2.24	1.00 ± 0.00	1000 ± 0.00

In Table 2 we again observe a clear relationship between robustness and the average CP set size: the more robust the model, the tighter the conformal sets. Regardless of the strength of the attack, the marginal coverage remains close to the prescribed confidence level $1 - \alpha = 0.90$. As an example, consider the CIFAR100–L_2 scenario, where the method of Diffenderfer et al. (2021) achieves a robust accuracy of 71.08%, surpassing that of Gowal et al. (2018) of 49.46%. The greater robustness of the former corresponds to smaller average CP sets. This trend becomes even more pronounced when comparing robust models with standard (non-defended) models—for instance on CIFAR10 under both L_2 and L_∞, and on ImageNet under L_∞. More generally, robust models yield compact prediction sets in most cases—especially against relatively weaker attacks such as FGSM and DeepFool - signalling high model confidence. By contrast, non-robust models produce considerably larger sets, particularly under stronger iterative attacks like PGD and BIM, reflecting larger predictive uncertainty in adversarial settings. On ImageNet, average set sizes are larger overall owing to the task's higher complexity and dimensionality, with the widest sets appearing for non-robust architectures such as AlexNet. Finally, the "Best" row aggregates the worst-case set size and corresponding coverage across all attacks, serving as a conservative robustness estimate. For robust models, the set size increase is controlled, whereas for non-robust models it becomes excessive (e.g., 1000 on ImageNet with the standardly trained model).

6. Contributions and Limitations of this Work

We revisited robust conformal prediction through the lens of realistic threat modeling. By explicitly enumerating admissible adversarial perturbations, attacks on the predictor reduce to attacks on the score function. This insight enables a closed-form recalibration of the conformal quantile, restoring finite-sample coverage that randomized or verification-based schemes either relax or secure only at prohibitive computational cost.

Experiments on standard classification benchmarks show that our method provide a good compromise between marginal coverage, robustness, and efficiency: prediction sets remain tight while guaranteeing coverage against the certified threat set. Importantly, these guarantees are falsifiable: when a previously unseen (zero-day) attack is discovered, it can be folded into the threat model and the score quantile patched without retraining the underlying model.

This work represents a first step toward provably and efficient conformal prediction, and several research directions remain open. Extending the framework to structured prediction tasks will require score functions tailored to high-dimensional outputs. Although patching is lightweight, automating the discovery and integration of novel attacks would further strengthen practical resilience, for example by leveraging online learning or continual adversarial training frameworks that adapt to emerging threats in real time. Finally, allowing the threat model to evolve with the data distribution could bridge the gap between certified robustness and real-world deployment, where adversaries are strategic and dynamic.

A key limitation of the current approach, PERCP, is its reliance on prior knowledge of the set of possible attacks, $\mathcal{A}$. This makes it less effective when the defender is unaware of a new attack. While the conformal guarantee can be restored by adding the new attack to the set $\mathcal{A}$, the approach is currently not designed for a-priori defense, which remains an

open challenge. Additionally, the method has been tested and is effective for classification tasks, but extending it to regression tasks is part of ongoing and future work.
In summary, by anchoring conformal prediction to a clear, falsifiable threat model, we chart a principled course toward prediction sets that are simultaneously efficient, provably robust, and amenable to continual hardening as adversarial tactics evolve.

Acknowledgments

This work is partially supported by (i) project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU and (ii) project FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU, (iii) REXASI-PRO H-EU project (HORIZON-CL4-2021-HUMAN-01-01, Grant Agreement ID: 101070028), (iv) "Fit4MedRob - Fit for Medical Robotics" Grant (PNC0000007), (v) project ELSA – European Lighthouse on Secure and Safe AI funded by the European Union's Horizon Europe under the grant agreement No. 101070617, and (vi) project FISA-2023-00128 funded by the MUR program "Fondo italiano per le scienze applicate"

References

R. Ahmad, I. Alsmadi, W. Alhamdani, and L. Tawalbeh. Zero-day attack detection: a systematic literature review. *Artificial Intelligence Review*, 56(10):10733–10811, 2023.

M. Akbari and T. N. A. Do. A systematic review of machine learning in logistics and supply chain management: current trends and future directions. *Benchmarking: An International Journal*, 28(10):2977–3005, 2021.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2013.

C. M. Bishop and H. Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023.

R. Bommasani, D. A Hudson, E. Adeli, R. Altman, and Others. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

C. Brix, M. N. Müller, S. Bak, T. T. Johnson, and C. Liu. First three years of the international verification of neural networks competition (VNN-COMP). *International Journal on Software Tools for Technology Transfer*, 25(3):329–339, 2023.

A. Carlevaro, T. Alamo, F. Dabbene, and M. Mongelli. Conformal predictions for probabilistically robust scalable machine learning classification. *Machine Learning*, 113(9): 6645–6661, 2024a.

A. Carlevaro, S. Narteni, F. Dabbene, T. Alamo, and M. Mongelli. A probabilistic scaling approach to conformal predictions in binary image classification. In *Conformal and Probabilistic Prediction with Applications*, 2024b.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.

A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, and Others. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, 55(13s):1–39, 2023.

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *International conference on machine learning*, 2019.

T. Cordier, V. Blot, L. Lacombe, T. Morzadec, A. Capitaine, and N. Brunel. Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library. In *Conformal and Probabilistic Prediction with Applications*, 2023.

J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

J. Diffenderfer, B. Bartoldson, S. Chaganti, J. Zhang, and B. Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. *Neural information processing systems*, 2021.

M. F. Dixon, I. Halperin, and P. Bilokon. *Machine learning in finance*. Springer, 2020.

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

S. Feldman, B. Einbinder, S. Bates, A. N. Angelopoulos, A. Gendler, and Y. Romano. Conformal prediction is robust to dispersive label noise. In *Conformal and Probabilistic Prediction with Applications*, 2023.

J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, and Others. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56:1513–1589, 2023.

A. Gendler, T. W. Weng, L. Daniel, and Y. Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

S. Ghosh, Y. Shi, T. Belkhouja, Y. Yan, J. Doppa, and B. Jones. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, 2023.

J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

C. Herley. Unfalsifiability of security claims. *Proceedings of the National Academy of Sciences*, 113(23):6415–6420, 2016.

L. Jeary, T. Kuipers, M. Hosseini, and N. Paoletti. Verifiably robust conformal prediction. In *Neural Information Processing Systems*, 2024.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

A. Krizhevsky. *Learning multiple layers of features from tiny images*. Toronto, ON, Canada, 2009.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Neural information processing systems*, 2012.

S. Kullback. *Kullback-leibler divergence*. Tech. Rep, 1951.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arxiv 2016. *arXiv preprint arXiv:1607.02533*, 2016.

A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 2018.

Y. Le and X. Yang. Tiny imagenet visual recognition challenge. In *Stanford Class CS 231N*, 2015.

L. Lindemann, Y. Zhao, X. Yu, G. J. Pappas, and J.V. Deshmukh. Formal verification and control with conformal prediction. *arXiv preprint arXiv:2409.00536*, 2024.

C. Liu, Y. Dong, W. Xiang, X. Yang, H. Su, J. Zhu, et al. A comprehensive study on robustness of image classification models: benchmarking and rethinking pp 1–36. *arXiv preprint arXiv:2302.14301*, 2023.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

S. Marro and M. Lombardi. Computational asymmetries in robust classification. In *International Conference on Machine Learning*, 2023.

S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE conference on computer vision and pattern recognition*, 2016.

R. Muthalagu, J. Malik, and P. M. Pawar. Detection and prevention of evasion attacks on machine learning models. *Expert Systems with Applications*, 266:126044, 2025.

A. Paleyes, R.G. Urma, and N. D. Lawrence. Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, 55(6):1–29, 2022.

H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, 2002.

N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas. A taxonomy and survey of attacks against machine learning. *Computer Science Review*, 34:100199, 2019.

Y. Qian, K. Siau, and F.F. Nah. Societal impacts of artificial intelligence: Ethical, legal, and governance issues. *Societal Impacts*, 3:100040, 2024. ISSN 2949-6977.

L. Rice, E. Wong, and Z. Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, 2020.

Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. In *Neural Information Processing Systems*, 2020.

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

H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Neural information processing systems*, 2019.

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

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

K. Stuurman and E. Lachaud. Regulating ai. a label to complete the proposed act on artificial intelligence. *Computer Law & Security Review*, 44:105657, 2022.

R. Tian, Z. Wu, Q. Dai, H. Hu, and Y. Jiang. Deeper insights into the robustness of vits towards common corruptions. *arXiv preprint arXiv:2204.12143*, 2022.

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. In *Neural information processing systems*, 2019.

E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Conference on fairness, accountability, and transparency*, 2020.

I. Vaccari, A. Carlevaro, S. Narteni, E. Cambiaso, and M. Mongelli. explainable and reliable against adversarial machine learning in data analytics. *IEEE Access*, 10:83949–83970, 2022.

V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10:988–999, 1999.

V. Vovk. Randomness, exchangeability, and conformal prediction. *arXiv preprint arXiv:2501.11689*, 2025.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Y. Wang, Z. Cui, and R. Ke. *Machine learning for transportation research and applications*. Elsevier, 2023a.

Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan. Better diffusion models further improve adversarial training. In *International conference on machine learning*, 2023b.

E. Wong and J. Z. Kolter. Learning perturbation sets for robust machine learning. In *International Conference on Learning Representations*, 2023.

E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, 2018.

W. Xiong and R. Lagerström. Threat modeling-a systematic literature review. *Computers & security*, 84:53–69, 2019.

K. Xu, H. Zhang, S. Wang, Y. Wang, S. Jana, X. Lin, and C. J. Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021.

G. Yan, Y. Romano, and T. W. Weng. Provably robust conformal prediction with improved efficiency. *arXiv preprint arXiv:2404.19651*, 2024.

H. S. Zargarbashi, M. S. Akhondzadeh, and A. Bojchevski. Robust yet efficient conformal prediction sets. In *International Conference on Machine Learning*, 2024.

H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning, and C. J. Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020.

K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. Fu. A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, 1(1):100003, 2023.