GAME-THEORETIC DEFENSES FOR ROBUST CONFORMAL PREDICTION AGAINST ADVERSARIAL ATTACKS IN MEDICAL IMAGING

Rui Luo*[1], Jie Bao†[2], Zhixin Zhou‡[1], Chuangyin Dang§[1]

[1]City University of Hong Kong
[2]Huaiyin Institute of Technology

Abstract. Adversarial attacks pose significant threats to the reliability and safety of deep learning models, especially in critical domains such as medical imaging. This paper introduces a novel framework that integrates conformal prediction with game-theoretic defensive strategies to enhance model robustness against both known and unknown adversarial perturbations. We address three primary research questions: constructing valid and efficient conformal prediction sets under known attacks (RQ1), ensuring coverage under unknown attacks through conservative thresholding (RQ2), and determining optimal defensive strategies within a zero-sum game framework (RQ3). Our methodology involves training specialized defensive models against specific attack types and employing maximum and minimum classifiers to aggregate defenses effectively. Extensive experiments conducted on the MedMNIST datasets—PathMNIST, OrganAMNIST, and TissueMNIST—demonstrate that our approach maintains high coverage guarantees while minimizing prediction set sizes. The game-theoretic analysis reveals that the optimal defensive strategy often converges to a singular robust model, outperforming uniform and simple strategies across all evaluated datasets. This work advances the state-of-the-art in uncertainty quantification and adversarial robustness, providing a reliable mechanism for deploying deep learning models in adversarial environments.

1. Introduction

Adversarial attacks [1] significantly undermine the reliability and safety of deep learning models, particularly in high-stakes domains such as medical diagnostics [2] and autonomous driving [3]. These attacks introduce subtle perturbations to input data, causing models to produce incorrect and potentially harmful predictions. In critical applications like healthcare and autonomous systems, ensuring both adversarial robustness and reliable uncertainty quantification is paramount. Conformal Prediction (CP) [4, 5, 6] offers a robust framework for uncertainty quantification by generating prediction sets that encompass the true label with a predefined confidence level [7].

The interplay between adversarial robustness and conformal prediction [8, 9] is both crucial and underexplored. While extensive research has focused on enhancing adversarial robustness through various adversarial training methodologies [10, 11, 12, 13], the impact of these methods on the efficiency and reliability of conformal prediction sets remains inadequately understood. Specifically, prior studies have concentrated on adapting the scoring functions to attacks on either the test data or the calibration data, or both. However, these approaches often encounter validity issues, thereby compromising the reliability of CP sets under adversarial conditions.

This paper aims to bridge this knowledge gap by constructing robust conformal prediction sets for medical imaging datasets under diverse adversarial attack scenarios. We employ various attack types and partition the training data to develop specialized models resistant to each specific attack. Our objective is to formulate CP sets that uphold coverage guarantees while minimizing and optimizing prediction set sizes, even when subjected to both known and novel adversarial attacks.

To address the validity issues inherent in previous methodologies, we construct conservative prediction sets based on the largest estimated quantile derived from calibration data under different attacks. This approach ensures that the prediction sets maintain their validity guarantees despite adversarial perturbations. Central to our investigation is a pivotal question: given that validity is preserved, what constitutes the severest attack an adversary can orchestrate when they can choose any attack without the defender's knowledge? Furthermore, how can the defender effectively mitigate such attacks under this adversarial scenario? Addressing these questions leads naturally to a zero-sum game formulation,

*ruiluo@cityu.edu.hk
†1486103897@qq.com
‡zhixin0825@gmail.com
§mecdang@cityu.edu.hk

where we consider both maximum and minimum classifiers to strategize optimal defenses against the most potent attacks.

By integrating game-theoretic principles with conformal prediction, this study endeavors to enhance the robustness and reliability of deep learning models in adversarial environments. Our approach not only maintains high coverage rates but also ensures minimal prediction set ambiguity, advancing the safety and efficacy of AI deployments in critical applications such as healthcare.

2. Related Work

2.1. Adversarial Attack.

Poisoning attacks represent one of the most direct threats to the training process of machine learning models [14]. Current defenses against data poisoning attacks mainly fall into two categories. The first category focuses on anomaly detection using techniques such as nearest neighbors [15], training loss [16], singular value decomposition [17, 18], clustering [19], taxonomy [20], or logistic regression [21], as well as other related methods [22, 23] to filter out anomalies through data-driven approaches. While these countermeasures can mitigate the impact of poisoning to some extent, they exhibit clear shortcomings in terms of effectiveness, cost, and accuracy. The second category focuses on model-driven strategies aimed at enhancing model robustness through techniques such as randomized smoothing [24], ensemble methods [25], data augmentation [26], and adversarial training [27], among others [28].

However, poisoning attacks are more difficult to detect in deep learning models. Consequently, a wider range of methods has been developed to enhance model robustness against such attacks. For instance, improvements have been made by modifying the model architecture [29], designing robust loss functions [30, 31], and optimizing loss functions [32].

2.2. Conformal Prediction under Adversarial Attack.

Uncertainty estimation is crucial for the robustness of deep learning models. Conformal Prediction (CP), introduced by Vovk et al. [7], provides distribution-free coverage guarantees but faces challenges when confronted with data poisoning and adversarial attacks. Gendler et al. [33] proposed RSCP, enhancing CP's robustness through randomized smoothing by replacing the scoring function to defend against ℓ_2-norm adversarial perturbations, though its formal guarantees remain limited. Yan et al. [34] introduced RSCP+, which adds a hyperparameter β and Hoeffding bounds to correct Monte Carlo errors, while incorporating RCT and PTT techniques to improve efficiency. However, these modifications often result in overly conservative prediction sets and increased reliance on the holdout set.

In contrast, Ghosh et al. [35] proposed PRCP, which approaches adversarial perturbations from a probabilistic standpoint by relaxing robustness guarantees under a predefined distribution. Cauchois et al. [36] addressed distributional shifts using an f-divergence constraint, though the resulting optimization is overly conservative [37]. Angelopoulos et al. [8] extended CP to control expected loss but did not provide algorithms for computing worst-case adversarial loss. Einbinder et al. [9] demonstrated that standard CP is robust to random label noise, though adversarial label perturbations were not explored. Furthermore, traditional defenses based on anomaly detection, loss function adjustments, and data clustering [15, 16, 17, 19, 22] struggle to safeguard CP's unique properties, especially against sophisticated poisoning attacks targeting uncertainty mechanisms.

Building on these limitations, Liu et al. [12] proposed an uncertainty reduction-based adversarial training method (AT-UR), which combines Beta-weighted loss and entropy minimization to enable models to maintain robustness while generating larger and more meaningful prediction sets. Jeary et al. [13] introduced Verifiably Robust Conformal Prediction (VRCP), which supports perturbations bounded by various norms (including ℓ_1, ℓ_2, and ℓ_∞) and regression tasks, using recent neural network verification techniques to restore coverage guarantees under adversarial attacks. Li et al. [11] developed a new class of black-box data poisoning attacks targeting Conformal Prediction (CP), where the attacker aims to manipulate prediction uncertainty rather than induce misclassification. They also proposed an optimization framework to defend against such attacks, demonstrating its effectiveness across a range of scenarios. Zargarbashi et al. [10] proposed using the cumulative distribution function (CDF) of smoothed scores to derive tighter upper bounds on worst-case score variations, leading to more efficient prediction sets while preserving robustness guarantees. Their method addresses both continuous and discrete/sparse data, offering guarantees for evasion (perturbing test inputs) and contamination attacks (perturbing calibration data). Scholten et al. [38] introduced Reliable Prediction Sets (RPS), which aggregate predictions from classifiers trained on different partitions of the training data and calibrated on disjoint subsets, making CP more resilient to data poisoning attacks.

Recent studies have also highlighted CP's resilience to label noise. Penso and Goldberger [39] introduced Noise-Robust Conformal Prediction (NR-CP), which estimates noise-free scores from noisy data, constructing smaller and more efficient prediction sets while maintaining coverage guarantees. NR-CP significantly outperforms other methods in noisy environments. Another study [9] explored the inherent robustness of CP, showing that even with noisy labels, CP tends to conservatively maintain coverage, although some adversarial noise conditions may compromise its reliability. Both studies emphasize CP's resilience to label noise, a perturbation type that, while not intentionally malicious, introduces variability similar to adversarial attacks.

Our work's connection with existing works. Building upon the foundations laid by prior research, such as RSCP [33], PRCP [35], and VRCP [13], our work advances the robustness of CP systems in adversarial settings. While previous studies have focused on specific attack vectors targeting either test data [11] or calibration data [10], our approach considers a comprehensive adversarial model where the attacker can manipulate both aspects simultaneously without the defender's knowledge. By constructing a conservative prediction set based on the maximum estimated quantile from attacked calibration data, we ensure validity under a wider range of attack scenarios, addressing limitations noted in works such as RSCP+ [34] and PRCP [35], which may be overly conservative or insufficiently robust against certain attacks.

Furthermore, by framing the interaction between attacker and defender as a zero-sum game, akin to methodologies in adversarial training [27] and robust optimization [40], our work not only identifies optimal attack strategies but also devises corresponding defense mechanisms that preserve CP system integrity. This game-theoretic perspective extends the robustness guarantees provided by prior works [8, 38], offering a unified and strategic framework to enhance CP resilience against diverse and adaptive adversarial threats. Additionally, our approach synergizes with uncertainty reduction techniques [12] and efficient bound derivations [10], integrating these methodologies into a cohesive strategy that fortifies CP systems beyond the state-of-the-art methods.

In summary, our research complements and extends existing defense strategies by integrating game theory to model adversarial interactions and proposing novel mechanisms to fortify CP systems. This approach not only builds on the strengths of previous studies [35, 13, 11, 10] but also addresses their limitations, providing enhanced robustness and reliability for CP in high-stakes, adversarial environments.

3. Preliminary and Problem Setup

In this section, we establish the foundational notation and problem setup for evaluating and enhancing the robustness of Conformal Prediction (CP) systems against adversarial attacks. By defining the adversarial attack strategies, defensive models, conformal prediction framework, score functions, and metrics, and by formulating key research questions, we set the stage for developing and analyzing robust prediction mechanisms that maintain valid uncertainty estimates in the presence of deliberate perturbations. The notation used throughout the paper is given in Table 1.

Notation	**Description**
$\{(x_i, y_i)\}_{i\in\mathcal{I}}$	Images and labels where i belongs to the index set $\mathcal{I}$.
$\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{cal}}, \mathcal{I}_{\text{eval}}, \mathcal{I}_{\text{test}}$	The index set for training, calibration, evaluation, and testing.
f_0	The normal (pre-trained) model trained on clean data.
g_j, $j = 1, \dots, m$	Adversarial attack functions, where g_j maps the original inputs and labels to perturbed inputs, i.e., $g_j : (X, Y, f_0) \to X'$.
f_j, $j = 1, \dots, m$	Defensive models trained to defend specific attacks, where $f_j(x_i)$ denotes the defensive model f_j's prediction on input x_i.
ϵ	Maximum allowable perturbation for adversarial attacks.
$\mathcal{L}(f(x_i), y_i)$	Loss function of model f with respect to input x_i and true label y_i.
$\Gamma(x_i)$	Conformal prediction set for input x_i.
$q_{1-\alpha}$, $\alpha \in (0, 1)$	Quantile threshold for the desired coverage $1 - \alpha$.
$\mathcal{Y}$	Set of all possible labels.

Table 1. Notation Used Throughout the Paper

3.1. **Adversarial Attacks and Defensive Models.**

3.1.1. *Adversarial Attacks.* Adversarial attacks are strategies employed to intentionally perturb input data in a way that deceives machine learning models into making incorrect predictions. We consider the following adversarial attacks:

Fast Gradient Sign Method (FGSM). FGSM is one of the earliest and most straightforward adversarial attack techniques introduced by Goodfellow et al. (2014) [41]. It generates adversarial examples by perturbing the input data in the direction of the gradient of the loss function with respect to the input. The perturbation magnitude is controlled by a small factor ϵ, ensuring that the modifications remain imperceptible to human observers. Mathematically, the adversarial example x_i' is computed as:

$$x_i' = x_i + \epsilon \cdot \operatorname{sign}\left(\nabla_{x_i} \mathcal{L}(f_0(x_i), y_i)\right)$$

Here, x_i is the original input sample, and $\nabla_{x_i} \mathcal{L}(f_0(x_i), y_i)$ represents the gradient of the loss with respect to the input x_i.

FGSM is computationally efficient due to its single-step nature, making it suitable for rapid adversarial example generation. However, its simplicity can limit its effectiveness against models trained with robust defenses.

Projected Gradient Descent (PGD). PGD is an iterative extension of FGSM that applies multiple small perturbations, projecting the adversarial example back into the allowed perturbation space after each iteration. Introduced by Madry et al. (2017) [42], PGD enhances the effectiveness of the attack by allowing more precise adjustments to the input data through multiple refinement steps. The adversarial example at iteration $t+1$, denoted as $x_i'^{(t+1)}$, is computed as:

$$x_i'^{(t+1)} = \Pi_{\mathcal{B}(x_i,\epsilon)} \left\{ x_i'^{(t)} + \alpha \cdot \operatorname{sign}\left(\nabla_{x_i'^{(t)}} \mathcal{L}(f_0(x_i'^{(t)}), y_i) \right) \right\}$$

In this equation, α is the step size for each perturbation, and $\Pi_{\mathcal{B}(x_i,\epsilon)}$ is the projection operator that ensures $x_i'^{(t+1)}$ remains within the ϵ-ball around the original input x_i.

PGD performs adversarial perturbations in multiple iterations, each time adjusting the input based on the gradient of the loss function. This iterative process leads to stronger adversarial examples that are more likely to deceive robust models, making PGD a standard benchmark for evaluating model robustness.

Simultaneous Perturbation Stochastic Approximation (SPSA). SPSA [43] is a gradient-free adversarial attack that estimates the gradient of the loss function using random perturbations, making it suitable for black-box scenarios where the attacker does not have access to the model's internal parameters. Unlike FGSM and PGD, which rely on exact gradient information, SPSA approximates the gradient by simultaneously perturbing multiple dimensions of the input data with small random noise.

Mathematically, the SPSA attack updates the adversarial example x_i' as follows:

$$\widehat{\nabla} \mathcal{L}(f_0(x_i'), y_i) \approx \frac{\mathcal{L}(f_0(x_i' + \Delta_i), y_i) - \mathcal{L}(f_0(x_i' - \Delta_i), y_i)}{2\delta} \Delta_i,$$
$$x_i'^{(t+1)} = \Pi_{\mathcal{B}(x_i,\epsilon)} \left\{ x_i'^{(t)} + \alpha \cdot \operatorname{sign}\left(\widehat{\nabla} \mathcal{L}(f_0(x_i'^{(t)}), y_i) \right) \right\}$$

Here, Δ_i is a small random perturbation vector, δ is a smoothing parameter, and α is the step size. The projection operator $\Pi_{\mathcal{B}(x_i,\epsilon)}$ ensures that the adversarial example remains within the permissible perturbation space.

SPSA is particularly effective in black-box settings where gradient information is not readily available, as it requires only the evaluation of the loss function at perturbed inputs to approximate the gradient. This makes SPSA a versatile and powerful attack method against models with limited attack surface information.

3.1.2. *Defensive Models.* Defensive models are pre-trained neural network models designed to withstand specific adversarial attacks. In this framework, we consider:

Normal Model (f_0). A standard model trained solely on clean (non-adversarial) data.

Attack-Specific Models (f_j). : Separate models, each adversarially trained against a specific type of attack g_j (e.g., FGSM, PGD, SPSA).

Each defensive model f_j has been pre-trained and saved prior to experimentation. These models serve as the foundation for constructing a robust prediction mechanism through weighted combinations.

3.2. **Conformal Prediction Framework. Conformal Prediction (CP)** is a statistical framework that provides valid measures of uncertainty for machine learning predictions in a distribution-free manner. We start by assuming that a classification algorithm provides $\widehat{p}_y(x)$, which approximates $P(Y = y|X = x)$. While our method and theoretical analysis do not depend on the accuracy of this approximation, it is beneficial to assume that higher values of $\widehat{p}_y(x)$ indicate a greater likelihood of sample with feature x having label y. We conduct the training procedure on a set $\mathcal{I}_{\text{train}}$, which is separate from the calibration and test processes. Details will be provided in Section 4. This separation ensures that $\widehat{p}_y(x)$ remains independent of the conformal prediction procedure discussed in this paper.

Algorithm 1: Split Conformal Prediction

```
Require: Data $\{(x_i, y_i)\}_{i\in\mathcal{I}}$, $\{x_i\}_{i\in\mathcal{I}_{\text{test}}}$, pre-determined coverage probability $1-\alpha$
Ensure: A prediction set $\Gamma(x_i)$ for each $i \in \mathcal{I}_{\text{test}}$
1: Randomly split $\mathcal{I}$ into $\mathcal{I}_{\text{train}}$ and $\mathcal{I}_{\text{cal}}$.
2: Train a model $\widehat{p}_y(x)$ on $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{train}}}$.
3: $q_{1-\alpha} \leftarrow$ the $\lceil(1+|\mathcal{I}_{\text{cal}}|)(1-\alpha)\rceil$-th smallest non-conformity score $s(x_i, y_i)$ for $i \in \mathcal{I}_{\text{cal}}$.
4: For each $x_i \in \mathcal{I}_{\text{test}}$, set $\Gamma(x_i) \leftarrow \{y \in \mathcal{Y} \mid s(x_i, y) \leq q_{1-\alpha}\}$.
```

Let us first consider a single *non-conformity score function* $s(x, y)$. This function is defined such that smaller values of $s(x, y)$ indicate a higher priority of believing x has label y. A common choice for the non-conformity score is $s(x, y) = -\widehat{p}_y(x)$ [44]. In this context, conformal prediction for classification problems can be described by the following algorithm. To summarize, we first find a threshold in a set of labeled data, so that $s(x_i, y_i) \leq q_{1-\alpha}$ holds for at least $1-\alpha$ proportion in the set $\mathcal{I}_{\text{train}}$. Then we use this threshold to define the prediction set for any x_i in the test set, which is the lower level set of the function $y \mapsto s(x_i, y)$.

Algorithm 1 constructs prediction sets $\Gamma(x_i)$ for each $i \in \mathcal{I}_{\text{test}}$, based on the non-conformity scores and a threshold determined by the desired coverage probability $1-\alpha$. The algorithm splits the training data into two subsets: $\mathcal{I}_{\text{train}}$ for calculating the non-conformity scores, and $\mathcal{I}_{\text{cal}}$ for calibration. The threshold $q_{1-\alpha}$ is chosen to ensure the desired coverage probability.

A key property of this method is that, under the assumption of exchangeability of the data points (a weaker assumption than i.i.d.), it guarantees that the resulting prediction sets will contain the true label with probability at least $1-\alpha$:

$$\mathbb{P}\left(y_i \in \Gamma(x_i)\right) \geq 1-\alpha. \tag{1}$$

This property holds regardless of the accuracy of the underlying classification algorithm, making conformal prediction a powerful tool for uncertainty quantification.

3.3. **Problem Formulation.** We aim to address the following research questions to construct a robust conformal prediction system under adversarial attack scenarios:

(1) **RQ1**: Given that a known attack g_j is applied to the test dataset $\{x_i\}_{i\in\mathcal{I}_{\text{test}}}$, how can we use a defensive model f_j to construct a valid and efficient conformal prediction set $\Gamma(x_i)$?
(2) **RQ2**: Given that an unknown adversarial attack is applied to the test dataset $\{x_i\}_{i\in\mathcal{I}_{\text{test}}}$, how can we use a defensive model to construct a valid and efficient conformal prediction set $\Gamma(x_i)$?
(3) **RQ3**: Given that an unknown and potentially adversarial attack is applied to the test dataset $\{x_i\}_{i\in\mathcal{I}_{\text{test}}}$, how can we determine the optimal defensive strategy for the defender under a game-theoretic setting where the defender and adversarial attacker comprise a zero-sum game?

To formalize these questions, consider the following setup: Let f_0 be the normal model trained on clean data $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{train}}}$. An adversary with access to the true labels of the test dataset applies an attack function g_j to generate perturbed inputs $x_i' = g_j(x_i, y_i, f_0)$ for $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{test}}}$. Defensive models $\{f_j\}$ are trained on $\{(x_i', y_i)\}_{i\in\mathcal{I}_{\text{train}}}$ to resist specific attacks, where $x_i' = g_j(x_i, y_i, f_0)$, i.e., each f_j corresponds to defending against attack g_j.

The conformal predictor aims to construct prediction sets $\Gamma(x_i')$ for $i \in \mathcal{I}_{\text{test}}$ that satisfy the coverage guarantee (1):

$$\mathbb{P}\left(y_i \in \Gamma(x_i')\right) \geq 1-\alpha,$$

under the adversarial perturbations introduced by g_j.

Addressing these research questions is essential for enhancing the robustness of conformal prediction in adversarial settings. RQ1 and RQ2 focus on constructing valid and efficient prediction sets under known and unknown attacks, respectively, ensuring reliable uncertainty quantification. RQ3 explores optimal defensive strategies within a game-theoretic framework, providing strategic insights to mitigate adversarial threats. Together, these investigations advance the development of resilient conformal prediction systems capable of maintaining coverage guarantees even in the presence of adversarial perturbations.

4. Methodology

In this section, we detail our approach to addressing the three key research questions outlined in the problem formulation. Each research question is tackled in its respective subsection, where we describe the strategies that form the backbone of our robust CP system against adversarial attacks.

4.1. **Addressing RQ1: Known Adversarial Attack. RQ1**: Given that a known attack g_j is applied to the test dataset $\{x_i\}_{i\in\mathcal{I}_{\text{test}}}$, how can we use a defensive model f_j to construct a valid and efficient conformal prediction set?

To address RQ1, we leverage the knowledge of the attack g_j by applying it to the calibration set $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$, which is exchangeable with the test set. By replicating the attack g_j on the calibration data and obtaining $\{(x'_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$, we can compute the corresponding threshold $q^j_{1-\alpha}$. This threshold is then used to construct the prediction sets $\Gamma(x'_i)$ for each $i \in \mathcal{I}_{\text{test}}$. This approach follows the traditional split conformal prediction framework (Algorithm 1), with the key difference that both the calibration and test datasets are subjected to adversarial attacks. The exchangeability between the calibration and test sets ensures that the coverage guarantee (1) of conformal prediction remains valid even under these adversarial perturbations.

```
Algorithm 2: Constructing Conformal Prediction Sets under Known Attack (RQ1)
Require: Training dataset $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{train}}}$, Calibration dataset $\mathcal{I}_{\text{cal}} = \{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$,
         Test dataset $\mathcal{I}_{\text{test}} = \{x_i\}_{i\in\mathcal{I}_{\text{test}}}$, Known attack function $g_j$,
         Coverage probability $1-\alpha$
Ensure: Prediction sets $(\Gamma(x_i))_{i\in\mathcal{I}_{\text{test}}}$
▷ Train defensive models $f_k$ for each attack using attacked training set:
1: for each attack function $g_k, k \in 1, \dots, m$ do
       Apply attack $g_k$ to the training set to obtain:
           $x'_i = g_k(x_i, y_i, f_0), i \in \mathcal{I}_{\text{train}}$.
       Train the defensive model $f_k$ on the attacked training set $\{(x'_i, y_i)\}_{i\in\mathcal{I}_{\text{train}}}$.
▷ Calibrate each defensive model by applying the known attack $g_j$ to the calibration set:
2: Apply the known attack $g_j$ to the calibration set $\mathcal{I}_{\text{cal}}$ to obtain $\{(x'_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$, where $x'_i = g_j(x_i, y_i, f_0)$.
3: Compute non-conformity scores $s(x_i, y_i)$ based on $f_j(x_i)$ for each $(x_i, y_i) \in \mathcal{I}_{\text{cal}}$.
4: Determine the quantile threshold $q_\alpha$ as the $\lceil(1 + |\mathcal{I}_{\text{cal}}|)(1-\alpha)\rceil$-th smallest score in $\mathcal{I}_{\text{cal}}$.
▷ Construct prediction sets using the quantile estimated on the attacked calibration set
5: for each test instance $x_i$, $i \in \mathcal{I}_{test}$ do
       Construct the prediction set:
           $\Gamma(x_i) = \{y \in \mathcal{Y} \mid s(x_i, y) \le q_\alpha\}$
```

Procedure: The steps outlined in Algorithm 2 systematically apply the known adversarial attack to the calibration set, train an adversarially robust defensive model f_j, compute non-conformity scores based on the defensive model's predictions, and determine the appropriate quantile threshold to construct valid and efficient prediction sets for the test data.

Intuition: By applying the known attack g_j to the calibration set and training the defensive model f_j accordingly, we create a scenario where the calibration set remains exchangeable with the adversarially perturbed test set. This alignment guarantees that the coverage condition $\mathbb{P}(y_i \in \Gamma(x_i)) \ge 1-\alpha$ holds, while the adversarially trained classifier f_j optimizes the efficiency of the prediction sets.

4.2. **Addressing RQ2: Unknown Adversarial Attack. Research Question 2 (RQ2)**: Given that an unknown adversarial attack is applied to the test dataset $\mathcal{I}_{\text{test}}$, how can we use a defensive model to construct a valid and efficient conformal prediction set $\Gamma(x_i)$?

In scenarios where the adversarial attack is unknown, we cannot rely on applying the same attack strategy to the calibration set to maintain exchangeability. To ensure the coverage condition under these circumstances, we adopt a conservative approach by considering the worst-case adversarial scenario.

Procedure: The steps in Algorithm 3 involve applying each potential adversarial attack from the set $\mathcal{G}$ to the calibration set, training defensive models against each attack, computing corresponding quantile thresholds, and selecting the maximum quantile threshold to ensure coverage under the worst-case scenario when the exact nature of the attack is unknown.

Intuition: By evaluating multiple potential adversarial attacks and selecting the most stringent quantile threshold, we ensure that the prediction sets remain valid even when the exact nature of the attack is unknown. This conservative approach guarantees coverage but may result in larger prediction sets due to the maximization step.

4.3. **Addressing RQ3: Game-Theoretic Adversarial Attack. Research Question 3 (RQ3)**: Given that an unknown and potentially adversarial attack is applied to the test dataset $\mathcal{I}_{\text{test}}$, how can we determine the optimal defensive strategy for the defender under a game-theoretic setting?

Algorithm 3: Constructing Conformal Prediction Sets under Unknown Attack (RQ2)

```
Require: Training dataset $\{(x_i, y_i)\}_{i \in \mathcal{I}_{\text{train}}}$, Calibration dataset $\{(x_i, y_i)\}_{i \in \mathcal{I}_{\text{cal}}}$,
         Test dataset $\{x_i\}_{i \in \mathcal{I}_{\text{test}}}$, Set of potential attack functions $g_j, j = 1, \ldots, m$,
         Coverage probability $1 - \alpha$
Ensure: Prediction sets $(\Gamma(x_i)), i \in \mathcal{I}_{\text{test}}$
Based on the training method for $f_k$ outlined in Algorithm 2, we obtain the pre-trained defensive models
  $f_k$, $k = 1, \ldots, m$.
▷ Determine the largest quantile score across various attacks $g_j$ on the calibration set:
for each attack function $g_j, j \in 1, \ldots, m$ do
    Apply attack $g_j$ to the calibration set to obtain:
        $x'_i = g_j(x_i, y_i, f_0), i \in \mathcal{I}_{\text{cal}}$.
    Use $f_k(x'_i)$ to compute non-conformity scores $\{s^j(x'_i, y_i)\}_{i \in \mathcal{I}_{\text{cal}}}$.
    Compute the quantile threshold as:
        $q^j_{1-\alpha}$ as the $\lceil (1 + |\mathcal{I}_{\text{cal}}|)(1 - \alpha) \rceil$-th smallest score in $s^j(x'_i, y_i)_{i \in \mathcal{I}_{\text{cal}}}$.
Determine the maximum quantile threshold $q_{1-\alpha} = \max_{j=1,\ldots,m} q^j_{1-\alpha}$.
▷ Construct conservative prediction sets using the largest quantile score to ensure coverage:
for each test instance $x_i$, $i \in \mathcal{I}_{test}$ do
    Construct the conservative prediction set:
        $\Gamma(x_i) = \{y \in \mathcal{Y} \mid s(x_i, y) \le q_{1-\alpha}\}$
```

To address RQ3, we model the interactions between the attacker and the defender as a zero-sum game, where the attacker aims to maximize the size of the prediction sets (thereby inducing uncertainty), and the defender seeks to minimize the prediction set sizes while maintaining the coverage guarantee. Specifically, the defender selects a defensive model from a predefined set of defensive models[5] $\{f_k\}_{k=1}^{p}$, while the attacker chooses an attack from a set of possible attacks $\{g_j\}_{j=1}^{m}$. Formally, the defender's objective is to minimize the maximum possible adversarial impact by solving the following optimization problem:

$$\min_{f_k, q_k} \max_{g_j} \sum_{i \in \mathcal{I}_{\text{test}}} \mathbb{E}\left[|\Gamma(x'_i)|\right],$$

subject to the coverage constraint:

$$\mathbb{P}\left(y_i \in \Gamma(x_i)\right) \ge 1 - \alpha,$$

where $x'_i = g_j(x_i, y_i, f_0)$ represents the adversarially perturbed input under attack g_j, and the non-conformity score associated with the perturbed input $s^k(x'_i, y_i)$ is computed based on f_k, i.e., the outcome is determined by both parties. Here, $\Gamma(x'_i)$ is the prediction set constructed using the defensive model f_k and the corresponding threshold q_k. As illustrated in Algorithm 3, q_k is obtained as the maximum of all $q^j_{1-\alpha}$ of the scores computed using the defensive model f_k on the calibration set attacked by g_j.

This optimization framework seeks to identify the defensive model f_k and threshold q_k that offer the most robust defense against the worst-case attack g_j. By considering the defender and attacker within a game-theoretic paradigm, we systematically evaluate and enhance the resilience of conformal prediction sets against a diverse range of adversarial threats. This ensures that the coverage guarantees of conformal prediction remain intact even when subjected to strategic adversarial perturbations, thereby reinforcing the reliability and robustness of the predictive system in adversarial environments.

Procedure: The steps in Algorithm 4 involve a calibration and a evaluation set that are exchangeable with the test set. It works by applying each potential adversarial attack to the calibration set, training defensive models against each attack, computing quantile thresholds and corresponding prediction set sizes on the evaluation set to construct the payoff matrix. The zero-sum game is then solved to identify the optimal defensive strategy that minimizes the maximum prediction set size. Finally, the selected quantile threshold is used to construct conservative prediction sets for the test data.

Intuition: By modeling the interaction between the attacker and the defender as a zero-sum game and evaluating the expected prediction set sizes across different attack and defense strategies, we can identify the defensive strategy that offers the most robust protection against the worst-case adversarial attacks. Selecting the largest quantile threshold q_k for each selected model f_k ensures that the coverage condition is maintained, even in the presence of the most challenging adversarial perturbations.

[5] We can have more defensive models than attacks, as exemplified by the Maximum classifier and the Minimum classifier introduced in Section 5.2.3.

Algorithm 4: Game-Theoretic Optimal Defensive Strategy (RQ3)

```
Require: Calibration dataset $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$, Evaluation dataset $\mathcal{I}_{\text{eval}} = \{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{eval}}}$,
         Test dataset $\{x_i\}_{i\in\mathcal{I}_{\text{test}}}$, Set of potential attack functions $g_j, j = 1, \dots, m$,
         Set of defensive models $f_k, k = 1, \dots, p$, Coverage probability $1 - \alpha$
Ensure: Optimal defensive strategy for constructing prediction sets $(\Gamma(x_i))_{x_i\in\mathcal{I}_{\text{test}}}$
▷ Determine the largest quantile score $q_k$ for each defensive model $f_k$ across various attacks:
for each defensive model $f_k, k = 1, \dots, p$ do
    for each attack function $g_j, j = 1, \dots, m$ do
        Apply attack $g_j$ to the calibration set to obtain:
            $x'_i = g_j(x_i, y_i, f_0), i \in \mathcal{I}_{\text{cal}}$.
        Use $f_k(x'_i)$ to compute non-conformity scores $\{s^j(x'_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$.
        Compute the quantile threshold as:
            $q^j_{1-\alpha}$ as the $\lceil(1 + |\mathcal{I}_{\text{cal}}|)(1 - \alpha)\rceil$-th smallest score in $\{s^j(x'_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$.
    Determine the maximum quantile threshold $q_k = \max_{j=1,\dots,m} q^j_{1-\alpha}$ for defensive model $f_k$.
▷ Estimate the prediction set sizes on the evaluation set:
for each defensive model $f_k, k = 1, \dots, p$ do
    for each attack function $g_j, j = 1, \dots, m$ do
        Compute the average prediction set size on the evaluation set:
            $\frac{1}{|\mathcal{I}_{\text{eval}}|} \sum_{i\in\mathcal{I}_{\text{eval}}} |\Gamma(x'_i)|$,
        where $\Gamma(x'_i) = |\{y \in \mathcal{Y} \mid s^j(x'_i, y) \le q_k\}|$, with $x'_i$ the perturbed input attacked by $g_j$ and
        $s^j(x'_i, y)$ computed using $f_k$.
Construct the payoff matrix $P$ where each entry $P_{k,j}$ represents the expected prediction set size when the
    attacker selects attack $g_j$ and the defender selects defensive model $f_k$.
Solve the zero-sum game using the payoff matrix $P$ to identify the optimal defensive strategy.
Select defensive models $f_k$, $k = 1, \dots, p$ according to the optimal defensive strategies.
▷ Construct the prediction sets using the optimal defensive strategy:
for each test instance $x_i, i \in \mathcal{I}_{test}$ do
    Construct the conservative prediction set:
        $\Gamma(x_i) = \{y \in \mathcal{Y} \mid s(x_i, y) \le q_k\}$
    with probability equal to the probability of selecting $f_k$ in the (potentially mixed) strategy.
```

5. Experiment

5.1. **Experimental Design.**

5.1.1. *Dataset Preparation.* The experiments utilize the MedMNIST dataset suite, selected via the `-dataset` argument (e.g., PathMNIST). Three specific datasets are employed, each with unique characteristics. PathMNIST [45] comprises 97176 images of colon pathology categorized into nine classes, with each image sized at 28 × 28 pixels. OrganAMNIST [45] consists of 34561 abdominal CT images distributed across eleven classes, also with a resolution of 28 × 28 pixels. TissuMNIST [46, 45] contains 236386 images of human kidney cortex cells organized into eight categories. Each grayscale image in TissuMNIST is originally $32 \times 32 \times 7$ pixels, with 2D projections obtained by taking the maximum pixel value along the axial axis of each pixel and subsequently resized to 28 × 28 grayscale images [47]. These datasets are preprocessed to ensure uniformity across models, including normalization and resizing as required by the ResNet18 architecture.

5.1.2. *Data Splitting and Experimental Setup.* For each dataset and every defensive model, including the normal (non-adversarial) model, the data is split into training, validation, and test sets. Specifically, 50% of the training dataset is used to train the defensive models, while 10% is allocated for validation purposes. The remaining data constitutes the test set, which is further divided based on the research question being addressed. For RQ1 and RQ2, the test dataset is split equally into a calibration set and a test set, each comprising 50% of the original test data. In contrast, for RQ3, the test dataset is partitioned into three subsets: 25% for calibration, 25% for evaluation, and 50% for the final test set. This stratified splitting ensures that each class is appropriately utilized for training, validating, and evaluating the defensive models under different adversarial attack scenarios.

The experimental setup comprises several key components, including the model architecture, training parameters, and adversarial example generation. We employ ResNet18 adapted for single-channel input and appropriate output layers as our model architecture. Training parameters include optimization using Stochastic Gradient Descent (SGD) with learning rate scheduling, a batch size set to 128, and training conducted over 60 epochs. Adversarial examples are generated using TorchAttacks, specifically implementing FGSM, PGD, and SPSA attacks during both the training and evaluation phases.

5.1.3. *Adversarial Attack Types.* Adversarial attacks are implemented using TorchAttacks [48]. The selected attacks include FGSM, PGD, and SPSA. Each attack type is designed to target the defensive model against which it is trained, facilitating comprehensive evaluation of model robustness across a variety of adversarial strategies. Employing multiple attack methodologies allows for a thorough assessment of a model's vulnerabilities and ensures that defenses are not tailored to only one specific type of adversarial manipulation.

5.1.4. *Conformal Prediction Calibration and Evaluation.* Conformal Prediction is performed using TorchCP [49]. To assess the performance of conformal prediction methods, we employ three primary metrics: Coverage, Size, and Size-Stratified Coverage Violation (SSCV).

The **Coverage** measures the proportion of test instances in $\mathcal{I}_{\text{test}}$ where the true label is contained within the prediction set $\Gamma(x_i)$, and is defined as

$$\text{Coverage} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \mathbb{1}\left(y_i \in \Gamma(x_i)\right). \tag{2}$$

A higher coverage indicates that the prediction sets reliably contain the true labels.

The **Size** metric calculates the average number of labels in the prediction sets across all test instances,

$$\text{Size} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} |\Gamma(x_i)|, \tag{3}$$

where smaller sizes denote more precise and informative predictions.

The **Size-Stratified Coverage Violation** (SSCV) [50] evaluates the consistency of coverage across different prediction set sizes. It is defined as

$$\text{SSCV}(\Gamma, \{S_j\}_{j=1}^{s}) = \sup_{j \in [s]} \left| \frac{|\{i \in J_j : y_i \in \Gamma(x_i)\}|}{|J_j|} - (1-\alpha) \right|, \tag{4}$$

where $\{S_j\}_{j=1}^{s}$ partitions the possible prediction set sizes, and $J_j = \{i \in \mathcal{I}_{\text{test}} : |\Gamma(x_i)| \in S_j\}$. A smaller SSCV indicates more stable coverage across different set sizes.

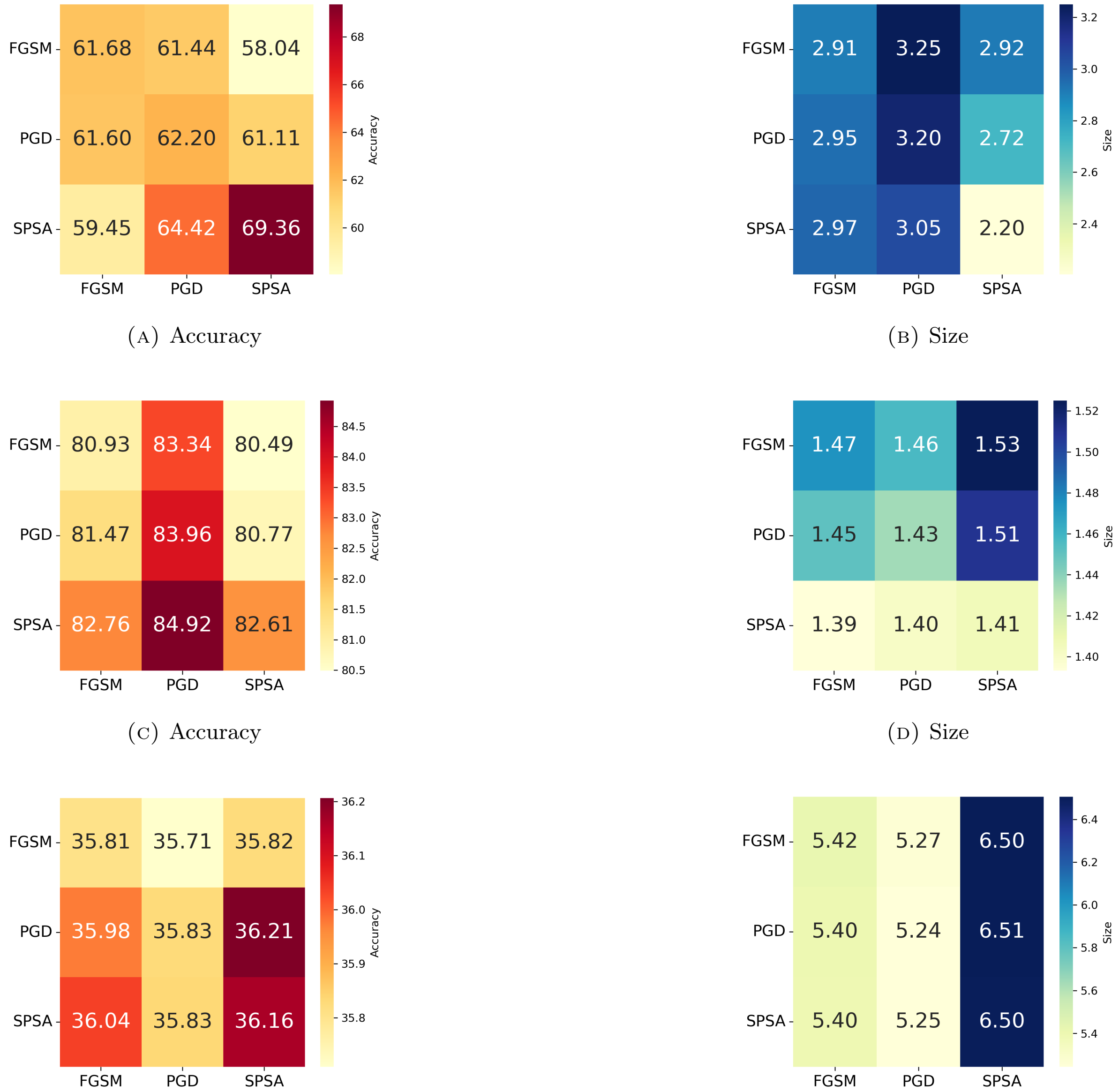
These metrics balance the trade-off between achieving the desired coverage probability and maintaining informative prediction sets, while ensuring that the coverage guarantees of conformal prediction hold regardless of the underlying model's accuracy.

5.2. **Results.** The trained models are evaluated against both known and unknown adversarial attacks to assess their robustness and the effectiveness of the conformal prediction sets. For RQ1, models are tested against the known attack applied during calibration. For RQ2, models are evaluated against unseen attacks to determine their generalizability. For RQ3, the game-theoretic approach is analyzed to identify optimal defensive strategies under adversarial conditions. For all experiments, we report the average and standard deviation across a number of splits to ensure statistical robustness.

5.2.1. *Results for RQ1.* In the first experiment, an attack of a known type is applied to the test data. We train defensive models f_j against each attack g_j, for $j = 1, \ldots, m$, and a normal model f_0. We evaluate both the prediction accuracy and the conformal prediction set's performance in terms of coverage, size, and SSCV. Specifically, Figure 1 illustrates the confusion matrices for both prediction accuracy and prediction set size for each defensive model and each attack. Comprehensive performance results are presented in Tables 2, 3, and 4. The key findings from RQ1 are as follows:

Defensive Model Accuracy. Although intuitively, the defensive model adversarially trained against a specific attack should exhibit the highest accuracy against that attack, this is not always the case. Various factors could be the reason, including the inherent complexity of certain attacks, the capacity of the defensive models, and the diversity of adversarial perturbations. Consequently, some defensive models demonstrate superior performance even against attacks they were not explicitly trained on.

Conformal Prediction Set Efficiency. The model with the best accuracy does not necessarily produce the most efficient conformal prediction sets. This discrepancy may be influenced by the chosen nonconformity score function. In our experiments, we employed the APS (Adaptive Prediction Set) [51] as the score function, which considers the cumulative probability of labels that have the same or lower estimated probabilities compared to the label of interest. As a result, APS emphasizes the tail probabilities, potentially leading to larger prediction sets for models focused primarily on accuracy.



(A) Accuracy

(B) Size

(C) Accuracy

(D) Size

(E) Accuracy

(F) Size

Figure 1. Confusion Matrices for Accuracy and Size across Datasets. Each row represents a dataset (**PathMNIST**, **OrganAMNIST**, **TissueMNIST**) and each column represents a metric (Accuracy, Size).

Table 2. **RQ1:** Mean and Standard Deviation of Coverage, Size, SSCV, and Accuracy for **OrganAMNIST**

Attack	Defensive Model	Coverage (%)	Size	SSCV	Accuracy (%)
FGSM	FGSM	89.79 ± 0.48	1.47 ± 0.02	0.06 ± 0.01	80.93 ± 0.30
	PGD	89.96 ± 0.49	**1.46 ± 0.01**	0.04 ± 0.01	**83.34 ± 0.28**
	SPSA	90.01 ± 0.36	1.53 ± 0.01	0.06 ± 0.01	80.49 ± 0.15
	Normal	89.96 ± 0.46	8.02 ± 0.05	0.13 ± 0.01	27.36 ± 0.24
PGD	FGSM	89.86 ± 0.37	1.45 ± 0.01	0.05 ± 0.01	81.47 ± 0.20
	PGD	89.93 ± 0.49	**1.43 ± 0.01**	0.04 ± 0.01	**83.96 ± 0.25**
	SPSA	90.01 ± 0.34	1.51 ± 0.01	0.06 ± 0.01	80.77 ± 0.26
	Normal	89.72 ± 0.36	8.91 ± 0.03	0.43 ± 0.03	14.86 ± 0.20
SPSA	FGSM	89.88 ± 0.30	**1.39 ± 0.01**	0.06 ± 0.01	82.76 ± 0.18
	PGD	89.92 ± 0.46	1.40 ± 0.02	0.04 ± 0.01	**84.92 ± 0.22**
	SPSA	89.97 ± 0.45	1.41 ± 0.01	0.07 ± 0.01	82.61 ± 0.35
	Normal	89.97 ± 0.29	7.72 ± 0.04	0.10 ± 0.01	31.70 ± 0.23

5.2.2. *Results for RQ2.* In the second experiment, an unknown attack is applied to the test data, with the assumption that it belongs to one of the three predefined attack types: FGSM, PGD, or SPSA. We

Table 3. **RQ1:** Mean and Standard Deviation of Coverage, Size, SSCV, and Accuracy for **PathMNIST**

Attack	**Defensive Model**	**Coverage (%)**	**Size**	**SSCV**	**Accuracy (%)**
FGSM	FGSM	90.19 ± 0.80	**2.91 ± 0.06**	0.05 ± 0.01	**61.68 ± 0.56**
	PGD	90.22 ± 0.48	3.25 ± 0.03	0.04 ± 0.01	61.44 ± 0.29
	SPSA	90.33 ± 0.56	2.92 ± 0.04	0.04 ± 0.01	58.04 ± 0.55
	Normal	89.89 ± 0.79	5.86 ± 0.10	0.60 ± 0.02	22.28 ± 0.34
PGD	FGSM	89.94 ± 0.39	2.95 ± 0.04	0.06 ± 0.01	61.60 ± 0.42
	PGD	90.31 ± 0.43	3.20 ± 0.04	0.03 ± 0.01	**62.20 ± 0.40**
	SPSA	90.25 ± 0.39	**2.72 ± 0.03**	0.04 ± 0.01	61.11 ± 0.58
	Normal	91.00 ± 0.63	6.86 ± 0.07	0.72 ± 0.02	12.76 ± 0.10
SPSA	FGSM	89.69 ± 0.61	2.97 ± 0.04	0.06 ± 0.01	59.45 ± 0.59
	PGD	90.17 ± 0.46	3.05 ± 0.04	0.03 ± 0.01	64.42 ± 0.35
	SPSA	90.03 ± 0.54	**2.20 ± 0.04**	0.03 ± 0.01	**69.36 ± 0.51**
	Normal	89.82 ± 0.86	4.82 ± 0.14	0.39 ± 0.01	46.57 ± 0.62

Table 4. **RQ1:** Mean and Standard Deviation of Coverage, Size, SSCV, and Accuracy for **TissueMNIST**

Attack	**Defensive Model**	**Coverage (%)**	**Size**	**SSCV**	**Accuracy (%)**
FGSM	FGSM	90.03 ± 0.17	5.42 ± 0.02	0.19 ± 0.07	35.81 ± 0.16
	PGD	90.03 ± 0.19	**5.27 ± 0.02**	0.21 ± 0.27	35.71 ± 0.11
	SPSA	90.02 ± 0.20	6.50 ± 0.00	0.13 ± 0.10	**35.82 ± 0.14**
	Normal	90.00 ± 0.21	5.91 ± 0.01	0.00 ± 0.00	14.91 ± 0.01
PGD	FGSM	89.95 ± 0.30	5.40 ± 0.04	0.18 ± 0.06	35.98 ± 0.13
	PGD	90.03 ± 0.19	**5.24 ± 0.03**	0.17 ± 0.26	35.83 ± 0.13
	SPSA	90.07 ± 0.22	6.51 ± 0.01	0.13 ± 0.07	**36.21 ± 0.19**
	Normal	90.06 ± 0.12	5.91 ± 0.01	0.00 ± 0.00	14.82 ± 0.02
SPSA	FGSM	89.99 ± 0.28	5.40 ± 0.03	0.18 ± 0.05	36.04 ± 0.14
	PGD	90.08 ± 0.20	**5.25 ± 0.02**	0.26 ± 0.34	35.83 ± 0.11
	SPSA	89.98 ± 0.22	6.50 ± 0.01	0.17 ± 0.09	**36.16 ± 0.14**
	Normal	89.99 ± 0.16	5.91 ± 0.00	0.00 ± 0.00	14.91 ± 0.02

employ Algorithm 3 to determine the largest threshold across the calibration sets under different attacks, thereby constructing a conservative prediction set that ensures the coverage condition is met.

We evaluate the conformal prediction set's performance in terms of coverage, size, and SSCV. The results are detailed in Tables 5, 6, and 7 for the respective datasets.

The findings from RQ2 indicate that the conservative approach effectively maintains the desired coverage level across all datasets, albeit with an increase in the average prediction set size due to the consideration of multiple potential attacks. In addition, the trade-off between coverage and prediction set size is evident, highlighting the balance between robustness and prediction precision.

Table 5. **RQ2:** Mean and Standard Deviation of Coverage, Size, and SSCV for **OrganAMNIST**

Attack	**Defensive Model**	**Coverage (%)**	**Size**	**SSCV**
FGSM	FGSM	89.83 ± 0.40	1.48 ± 0.02	0.06 ± 0.01
	PGD	89.89 ± 0.39	**1.46 ± 0.01**	0.04 ± 0.01
	SPSA	90.05 ± 0.47	1.53 ± 0.02	0.06 ± 0.01
	Normal	97.22 ± 0.13	9.52 ± 0.02	0.10 ± 0.00
PGD	FGSM	90.21 ± 0.42	1.48 ± 0.02	0.05 ± 0.01
	PGD	90.27 ± 0.22	**1.45 ± 0.01**	0.04 ± 0.01
	SPSA	90.27 ± 0.33	1.52 ± 0.01	0.06 ± 0.01
	Normal	89.79 ± 0.33	8.92 ± 0.03	0.42 ± 0.05
SPSA	FGSM	90.87 ± 0.38	1.45 ± 0.01	0.05 ± 0.01
	PGD	90.66 ± 0.27	**1.43 ± 0.01**	0.04 ± 0.01
	SPSA	91.26 ± 0.31	1.49 ± 0.01	0.05 ± 0.00
	Normal	97.50 ± 0.16	9.35 ± 0.03	0.10 ± 0.00

5.2.3. *Results for RQ3.* In the third experiment, an unknown and potentially adversarial attack is applied to the test data. The prediction set size corresponds to the attacker's payoff and the defender's cost, framing the interaction as a zero-sum game. To develop an optimal solution for the defender in a strategic manner, we preserve an additional evaluation set, which is exchangeable with both calibration and test sets. We use the prediction set size evaluated on this evaluation set as the estimated payoff matrix and derive the Nash equilibrium to determine the defender's optimal strategy. This strategy is then applied during the test phase.

In addition to the defensive models f_j and the normal model f_0, we implement a Maximum classifier $f_{\max}$ and a Minimum classifier $f_{\min}$. These classifiers aggregate the logits from multiple defensive models

Table 6. **RQ2:** Mean and Standard Deviation of Coverage, Size, and SSCV for **PathMNIST**

Attack	Defensive Model	Coverage (%)	Size	SSCV
FGSM	FGSM	91.08 ± 0.42	2.99 ± 0.03	0.05 ± 0.00
	PGD	90.08 ± 0.42	3.25 ± 0.04	0.03 ± 0.01
	SPSA	90.16 ± 0.39	**2.90 ± 0.04**	0.04 ± 0.01
	Normal	98.14 ± 0.31	7.65 ± 0.05	0.62 ± 0.07
PGD	FGSM	90.46 ± 0.49	2.98 ± 0.03	0.06 ± 0.01
	PGD	90.32 ± 0.64	3.21 ± 0.04	0.03 ± 0.01
	SPSA	90.85 ± 0.43	**2.79 ± 0.03**	0.04 ± 0.01
	Normal	90.86 ± 0.81	6.84 ± 0.08	0.71 ± 0.03
SPSA	FGSM	89.84 ± 0.70	2.96 ± 0.03	0.06 ± 0.01
	PGD	91.21 ± 0.54	3.18 ± 0.04	0.03 ± 0.01
	SPSA	94.18 ± 0.27	**2.65 ± 0.03**	0.06 ± 0.00
	Normal	98.98 ± 0.23	7.62 ± 0.05	0.36 ± 0.05

Table 7. **RQ2:** Mean and Standard Deviation of Coverage, Size, and SSCV for **TissueMNIST**

Attack	Defensive Model	Coverage (%)	Size	SSCV
FGSM	FGSM	89.98 ± 0.22	5.41 ± 0.02	0.18 ± 0.07
	PGD	90.00 ± 0.25	**5.26 ± 0.03**	0.24 ± 0.35
	SPSA	90.16 ± 0.29	6.51 ± 0.01	0.15 ± 0.09
	Normal	90.24 ± 0.09	5.93 ± 0.00	0.00 ± 0.00
PGD	FGSM	90.08 ± 0.16	5.42 ± 0.02	0.19 ± 0.07
	PGD	90.15 ± 0.22	**5.27 ± 0.03**	0.38 ± 0.39
	SPSA	90.22 ± 0.21	6.51 ± 0.01	0.09 ± 0.05
	Normal	90.06 ± 0.11	5.91 ± 0.00	0.20 ± 0.37
SPSA	FGSM	90.13 ± 0.20	5.42 ± 0.02	0.17 ± 0.04
	PGD	90.15 ± 0.23	**5.27 ± 0.03**	0.26 ± 0.35
	SPSA	90.14 ± 0.15	6.50 ± 0.01	0.16 ± 0.08
	Normal	90.31 ± 0.09	5.93 ± 0.00	0.00 ± 0.00

by taking the element-wise maximum and minimum across all models, respectively. Formally, for a given input x_i and class y, they are defined as:

$$\begin{aligned} f_{\max}(x_i)[y] &= \max_j f_j(x_i)[y], \\ f_{\min}(x_i)[y] &= \min_j f_j(x_i)[y], \end{aligned}$$

where $f_j(x_i)[y]$ denotes the predicted probability of the j-th defensive model for class y given input x_i.

The payoff matrix in this scenario has a shape of 6×3, accounting for the five defensive strategies (including $f_{\max}$ and $f_{\min}$) and the three attack types. Examples of pairs of estimated payoff matrices and their corresponding true payoff matrices are displayed in Figures 2, 3, and 4 for each of the three datasets.

We evaluate the performance of different defensive strategies, including the optimal strategy derived from the Nash equilibrium on the evaluation set, as well as simple strategies and a uniform strategy that assigns equal probability to each defensive model. The prediction set sizes on the test set using these strategies are reported in Figures 5a, 5b, and 5c. The results indicate that the optimal strategy consistently outperforms other strategies, often degenerating to a singular defensive strategy rather than a mixed strategy, thereby achieving the best performance across different datasets.

5.3. **Discussion.** The experiments validate that adversarial training tailored to specific attack types enhances the robustness of conformal prediction sets. In scenarios with unknown attacks, leveraging a validation set to optimally combine defenses ensures maintained coverage with efficient prediction sizes. This aligns with findings from Liu et al. [12], which emphasize the importance of considering CP efficiency during adversarial training.

However, there are limitations. The assumption that the attack type is one of the known types in Experiment 1 may not hold in all real-world scenarios. Additionally, the grid search in Experiment 2 introduces computational overhead, which could be mitigated with more efficient optimization techniques.

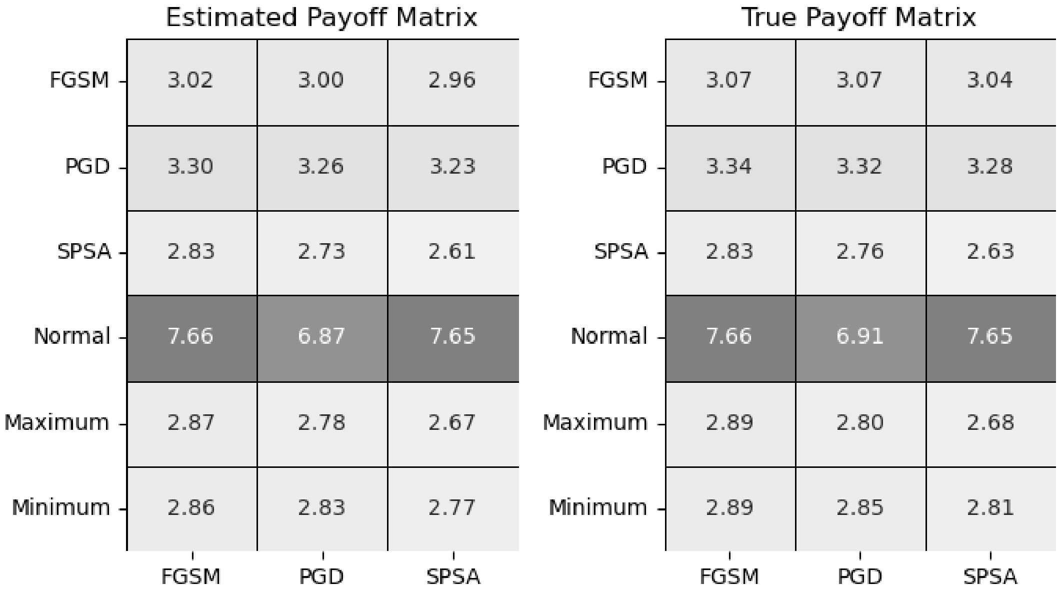


FIGURE 2. **RQ3:** Estimated and True Payoff Matrices for PathMNIST

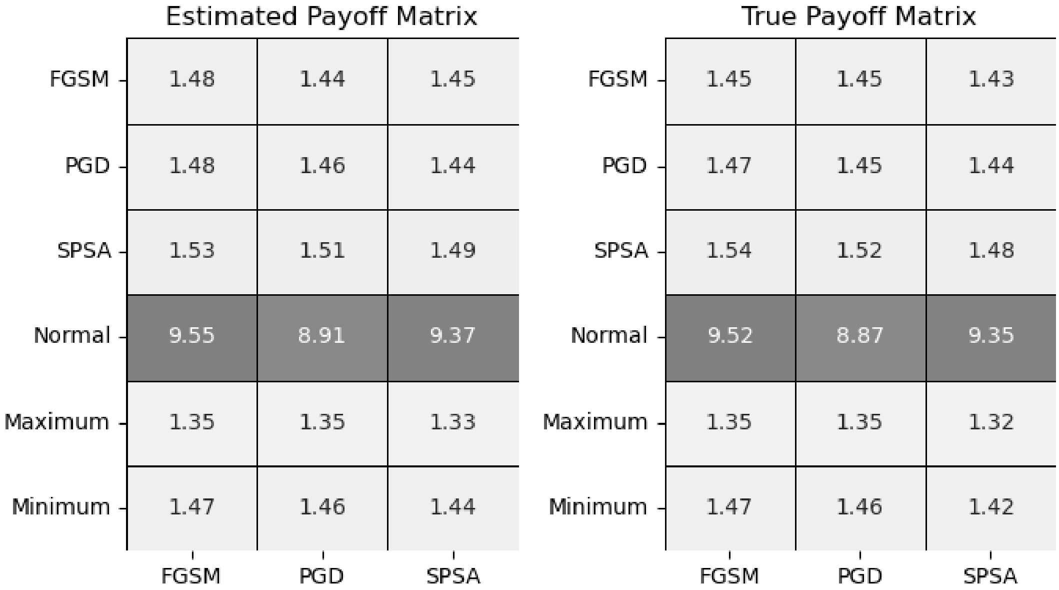


FIGURE 3. **RQ3:** Estimated and True Payoff Matrices for OrganAMNIST

Future research could focus on dynamically adapting to a wider variety of attacks, expanding the framework to include additional medical imaging datasets, and exploring the integration of uncertainty-reducing adversarial training techniques. Additionally, developing novel scoring functions [52], weighting strategies [53], and entropy reweighting methods [54] could further enhance the efficiency of CP.

6. CONCLUSION

This study presents a robust framework for constructing conformal prediction sets on medical imaging datasets under adversarial attacks. By training specialized models for distinct attack types and employing strategic model selection and weighting, we achieve high coverage guarantees with minimal prediction set sizes. Crucially, our methodology integrates game-theoretic principles to formulate and identify optimal defensive strategies within a zero-sum game framework between the attacker and defender. This game-theoretic approach ensures that the defensive strategies are not only resilient against known and unknown

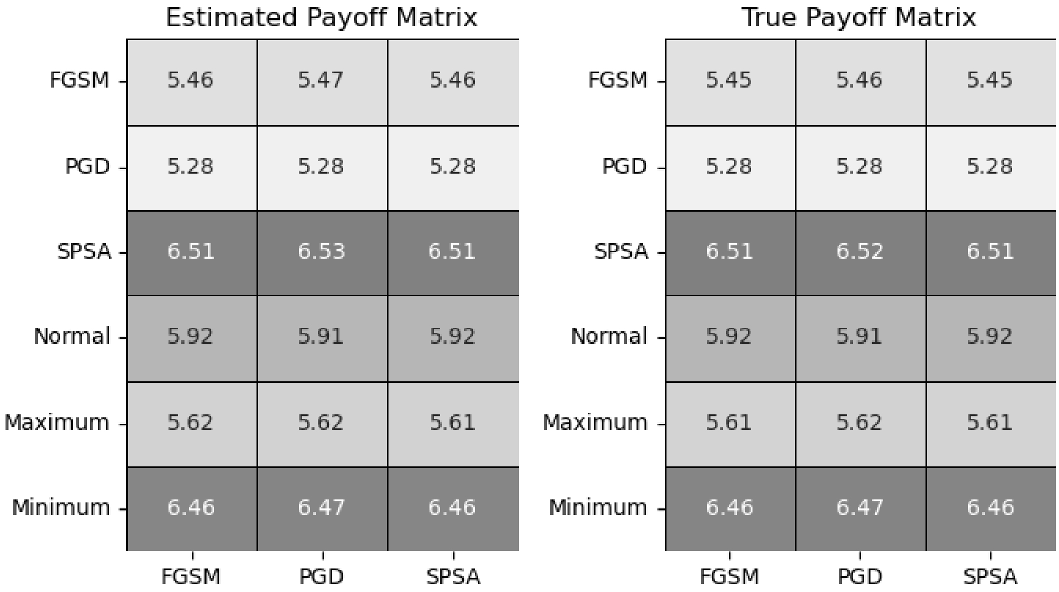


FIGURE 4. **RQ3:** Estimated and True Payoff Matrices for TissueMNIST

adversarial perturbations but also strategically optimized to mitigate the most severe threats posed by adaptive adversaries.

In summary, the proposed methodology synergizes adversarial robustness with uncertainty quantification through a weighted combination of defensive models and the conformal prediction framework. By systematically calibrating and evaluating different weight configurations within a game-theoretic context, our approach guarantees that the final prediction sets are both reliable and efficient, maintaining high coverage rates while minimizing ambiguity in predictions. The incorporation of game-theoretic defensive strategies enhances the ability of the conformal prediction system to adaptively respond to diverse and evolving adversarial attacks, thereby positioning our methodology as a comprehensive and strategic solution for deploying deep learning models in adversarial environments. This dual focus on resilience against adversarial perturbations and the provision of meaningful uncertainty estimates significantly advances the reliability and security of AI systems in critical applications such as healthcare.

REFERENCES

[1] N. Kumari, M. Singh, A. Sinha, H. Machiraju, B. Krishnamurthy, V. N. Balasubramanian, Harnessing the vulnerability of latent layers in adversarially trained models, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, pp. 2779–2785.

[2] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, F. Lu, Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recognition 110 (2021) 107332.

[3] B. Badjie, J. Cecílio, A. Casimiro, Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review, ACM Computing Surveys (2024).

[4] K. Katsios, H. Papadopoulos, Multi-label conformal prediction with a mahalanobis distance nonconformity measure, Proceedings of Machine Learning Research 230 (2024) 1–14.

[5] J. A. Meister, K. A. Nguyen, S. Kapetanakis, Z. Luo, A novel deep learning approach for one-step conformal prediction approximation, Annals of Mathematics and Artificial Intelligence (2023) 1–28.

[6] V. Balasubramanian, S.-S. Ho, V. Vovk, Conformal prediction for reliable machine learning: theory, adaptations and applications, Newnes, 2014.

[7] V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world, in: Algorithmic Learning in a Random World, 2005.

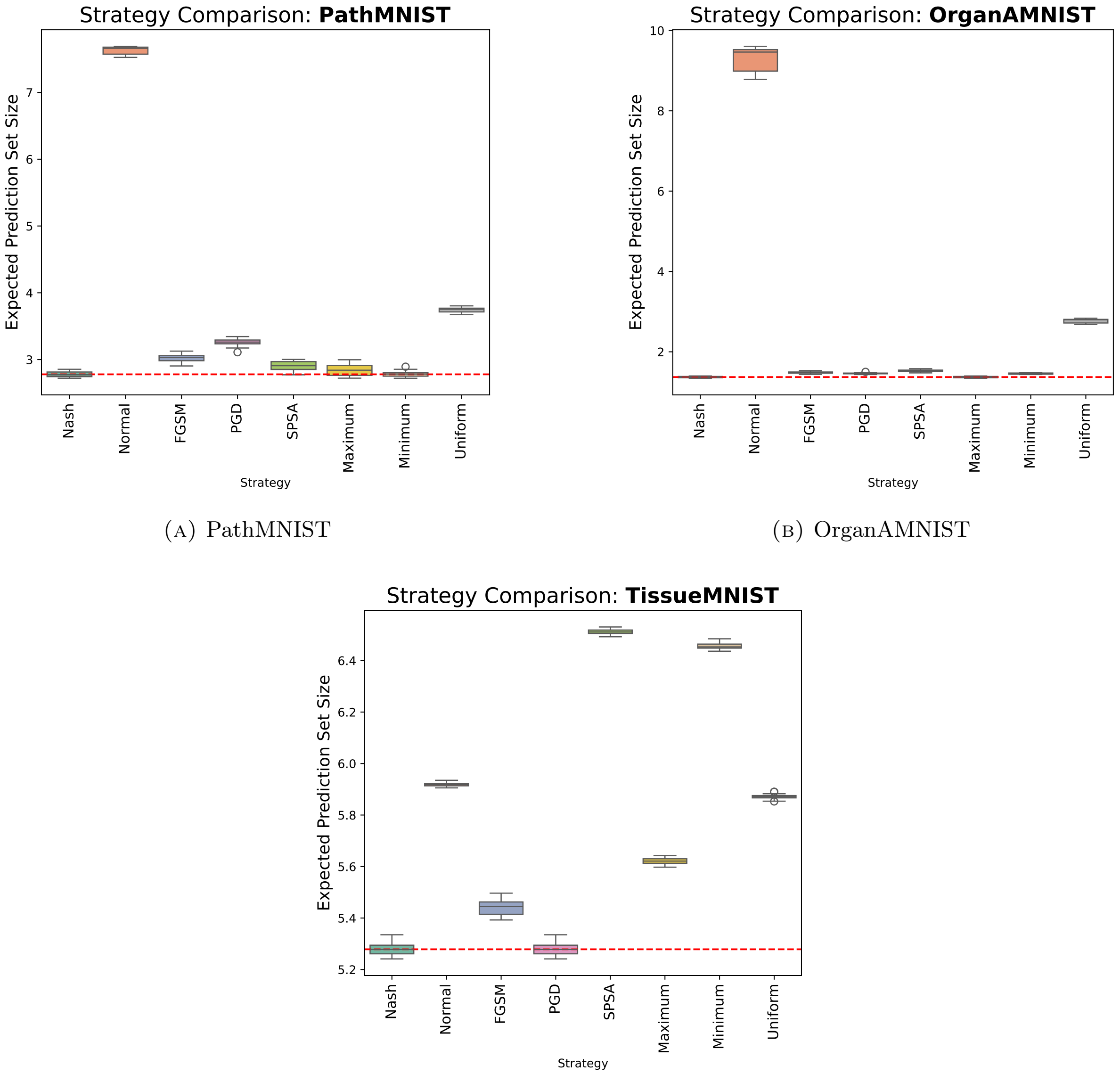
[8] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster, Conformal risk control, arXiv preprint arXiv:2208.02814 (2022).

[9] B.-S. Einbinder, S. Bates, A. N. Angelopoulos, A. Gendler, Y. Romano, Conformal prediction is robust to label noise, arXiv preprint arXiv:2209.14295 2 (2022).

[10] S. H. Zargarbashi, M. S. Akhondzadeh, A. Bojchevski, Robust yet efficient conformal prediction sets, arXiv preprint arXiv:2407.09165 (2024).

[11] Y. Li, A. Chen, W. Qian, C. Zhao, D. Lidder, M. Huai, Data poisoning attacks against conformal prediction, in: Forty-first International Conference on Machine Learning, 2024.

[12] Z. Liu, Y. Cui, Y. Yan, Y. Xu, X. Ji, X. Liu, A. B. Chan, The pitfalls and promise of conformal inference under adversarial attacks, arXiv preprint arXiv:2405.08886 (2024).



(a) PathMNIST

(b) OrganAMNIST

(c) TissueMNIST

Figure 5. **RQ3:** Comparison of Strategies Across Different Datasets

[13] L. Jeary, T. Kuipers, M. Hosseini, N. Paoletti, Verifiably robust conformal prediction, arXiv preprint arXiv:2405.18942 (2024).

[14] Z. Tian, L. Cui, J. Liang, S. Yu, A comprehensive survey on poisoning attacks and countermeasures in machine learning, ACM Computing Surveys 55 (8) (2022) 1–35.

[15] N. Peri, N. Gupta, W. R. Huang, L. Fowl, C. Zhu, S. Feizi, T. Goldstein, J. P. Dickerson, Deep k-nn defense against clean-label data poisoning attacks, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 55–70.

[16] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo, A. D. Keromytis, Casting out demons: Sanitizing training data for anomaly sensors, in: 2008 IEEE Symposium on Security and Privacy (sp 2008), IEEE, 2008, pp. 81–95.

[17] B. Tran, J. Li, A. Madry, Spectral signatures in backdoor attacks, Advances in neural information processing systems 31 (2018).

[18] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, A. Stewart, Sever: A robust meta-algorithm for stochastic optimization, in: International Conference on Machine Learning, PMLR, 2019, pp. 1596–1606.

[19] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, B. Srivastava, Detecting backdoor attacks on deep neural networks by activation clustering, arXiv preprint arXiv:1811.03728 (2018).

[20] M. Barreno, B. Nelson, A. D. Joseph, J. D. Tygar, The security of machine learning, Machine learning 81 (2010) 121–148.

[21] J. Feng, H. Xu, S. Mannor, S. Yan, Robust logistic regression and classification, Advances in neural information processing systems 27 (2014).

[22] J. Steinhardt, P. W. W. Koh, P. S. Liang, Certified defenses for data poisoning attacks, Advances in neural information processing systems 30 (2017).

[23] A. Paudice, L. Muñoz-González, A. Gyorgy, E. C. Lupu, Detection of adversarial training examples in poisoning attacks through anomaly detection, arXiv preprint arXiv:1802.03041 (2018).

[24] M. Weber, X. Xu, B. Karlaš, C. Zhang, B. Li, Rab: Provable robustness against backdoor attacks, in: 2023 IEEE Symposium on Security and Privacy (SP), IEEE, 2023, pp. 1311–1328.

[25] A. Levine, S. Feizi, Deep partition aggregation: Provable defense against general poisoning attacks, arXiv preprint arXiv:2006.14768 (2020).

[26] Y. Ma, X. Zhu, J. Hsu, Data poisoning against differentially-private learners: Attacks and defenses, arXiv preprint arXiv:1903.09860 (2019).

[27] L. Tao, L. Feng, J. Yi, S.-J. Huang, S. Chen, Better safe than sorry: Preventing delusive adversaries with adversarial training, Advances in Neural Information Processing Systems 34 (2021) 16209–16225.

[28] C. Liu, B. Li, Y. Vorobeychik, A. Oprea, Robust linear regression against training data poisoning, in: Proceedings of the 10th ACM workshop on artificial intelligence and security, 2017, pp. 91–102.

[29] J. Goldberger, E. Ben-Reuven, Training deep neural-networks using a noise adaptation layer, in: International conference on learning representations, 2017.

[30] A. Ghosh, H. Kumar, P. S. Sastry, Robust loss functions under label noise for deep neural networks, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 31, 2017.

[31] Y. Xu, P. Cao, Y. Kong, Y. Wang, L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise, Advances in neural information processing systems 32 (2019).

[32] D. Hendrycks, M. Mazeika, D. Wilson, K. Gimpel, Using trusted data to train deep networks on labels corrupted by severe noise, Advances in neural information processing systems 31 (2018).

[33] A. Gendler, T.-W. Weng, L. Daniel, Y. Romano, Adversarially robust conformal prediction, in: International Conference on Learning Representations, 2021.

[34] G. Yan, Y. Romano, T.-W. Weng, Provably robust conformal prediction with improved efficiency, arXiv preprint arXiv:2404.19651 (2024).

[35] S. Ghosh, Y. Shi, T. Belkhouja, Y. Yan, J. Doppa, B. Jones, Probabilistically robust conformal prediction, in: Uncertainty in Artificial Intelligence, PMLR, 2023, pp. 681–690.

[36] M. Cauchois, S. Gupta, A. Ali, J. C. Duchi, Robust validation: Confident predictions even when distributions shift, Journal of the American Statistical Association (2024) 1–66.

[37] K. D. Dvijotham, J. Hayes, B. Balle, Z. Kolter, C. Qin, A. Gyorgy, K. Xiao, S. Gowal, P. Kohli, A framework for robustness certification of smoothed classifiers using f-divergences, in: International Conference on Learning Representations, 2020.

[38] Y. Scholten, S. Günnemann, Provably reliable conformal prediction sets in the presence of data poisoning, arXiv preprint arXiv:2410.09878 (2024).

[39] C. Penso, J. Goldberger, Noise-robust conformal prediction for medical image classification, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2024, pp. 159–168.

[40] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, L. Qu, Making deep neural networks robust to label noise: A loss correction approach, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1944–1952.

[41] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 (2014).

[42] A. Madry, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).

[43] J. Uesato, B. O'donoghue, P. Kohli, A. Oord, Adversarial risk and the dangers of evaluating against weak attacks, in: International conference on machine learning, PMLR, 2018, pp. 5025–5034.

[44] M. Sadinle, J. Lei, L. Wasserman, Least ambiguous set-valued classifiers with bounded error levels, Journal of the American Statistical Association 114 (09 2016). doi:10.1080/01621459.2017.1395341.

[45] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification, Scientific Data 10 (1) (2023) 41.

[46] J. Yang, R. Shi, B. Ni, Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 191–195.

[47] A. Woloshuk, S. Khochare, A. F. Almulhim, A. T. McNutt, D. Dean, D. Barwinska, M. J. Ferkowicz, M. T. Eadon, K. J. Kelly, K. W. Dunn, et al., In situ classification of cell types in human kidney tissue using 3d nuclear staining, Cytometry Part A 99 (7) (2021) 707–721.

[48] H. Kim, Torchattacks: A pytorch repository for adversarial attacks, arXiv preprint arXiv:2010.01950 (2020).

[49] H. Wei, J. Huang, Torchcp: A library for conformal prediction based on pytorch, arXiv preprint arXiv:2402.12683 (2024).

[50] A. N. Angelopoulos, S. Bates, M. Jordan, J. Malik, Uncertainty sets for image classifiers using conformal prediction, in: International Conference on Learning Representations, 2021.

[51] Y. Romano, M. Sesia, E. Candes, Classification with valid and adaptive coverage, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 3581–3591.

[52] R. Luo, Z. Zhou, Trustworthy classification through rank-based conformal prediction sets, arXiv preprint arXiv:2407.04407 (2024).

[53] R. Luo, Z. Zhou, Weighted aggregation of conformity scores for classification, arXiv preprint arXiv:2407.10230 (2024).

[54] R. Luo, N. Colombo, Entropy reweighted conformal classification, arXiv preprint arXiv:2407.17377 (2024).