RPP: A Certified Poisoned-Sample Detection Framework for Backdoor Attacks under Dataset Imbalance

Anonymous authors
Paper under double-blind review

Abstract

Deep neural networks are highly susceptible to backdoor attacks, yet most defense methods to date rely on balanced data, overlooking the pervasive class imbalance in real-world scenarios that can amplify backdoor threats. This paper presents the first in-depth investigation of how the *dataset imbalance* amplifies backdoor vulnerability, showing that (i) the imbalance induces a majority-class bias that increases susceptibility and (ii) conventional defenses degrade significantly as the imbalance grows. To address this, we propose **Randomized Probability Perturbation (RPP)**, a *certified poisoned-sample detection framework* that operates in a black-box setting using only model output probabilities. For any inspected sample, RPP determines whether the input has been backdoor-manipulated, while offering provable within-domain detectability guarantees and a probabilistic upper bound on the false positive rate. Extensive experiments on five benchmarks (MNIST, SVHN, CIFAR-10, TinyImageNet and ImageNet10) covering 10 backdoor attacks and 12 baseline defenses show that RPP achieves significantly higher detection accuracy than state-of-the-art defenses, particularly under dataset imbalance. RPP establishes a theoretical and practical foundation for defending against backdoor attacks in real-world environments with imbalanced data.

1 Introduction

Deep neural networks (DNNs) are increasingly embedded in safety-critical applications ranging from autonomous driving (Kong et al., 2020; Wen & Jo, 2022) to facial recognition (Yang et al., 2021; Anusudha et al., 2024). However, they are known to be vulnerable to backdoor attacks (Gu et al., 2019; Goldblum et al., 2022; Zeng et al., 2023; Le Roux et al., 2024). The core concept of a neural backdoor involves embedding a unique pattern, known as a trigger, into the training data. The resulting model behaves normally on clean inputs but, when the trigger appears, reliably redirects predictions to an attacker-chosen target class. For example, a small sticker on a stop sign can induce a traffic system to read "go" (Liu et al., 2018b).

One major way of defending against backdoor attack, called prior-training defense, is to detect and remove poisoned samples from the dataset before the training process, evidenced by a number of defenses (Tran et al., 2018; Chen et al., 2021; Qi et al., 2023b; Pan et al., 2023; Guo et al., 2023; Pal et al., 2024). However, most existing methods typically assume an idealized scenario where defenses operate on balanced training sets (or their derived models)—a condition rarely found in real-world domains. In practice, data are often highly imbalanced, with some classes far more prevalent than others (He & Garcia, 2009; Fernández et al., 2018; Johnson & Khoshgoftaar, 2019; Aguiar et al., 2024). For example, in medical diagnosis datasets, healthy samples might outnumber cancer cases by more than 1000:1. Likewise, in autonomous driving, minority classes (e.g., rare traffic signs or road anomalies) are inherently underrepresented.

In Sec. 3, we conducted a comprehensive investigation that reveals two key phenomena in this imbalance context: (1) We identify a significant correlation between the efficacy of backdoor attacks and the degree of dataset imbalance, demonstrating that imbalanced datasets are inherently more susceptible to backdoor attacks. (2) We further observe that the performance of existing backdoor defense mechanisms degrades significantly as data imbalance intensifies. On the other hand, existing approaches lack formal theoretical guarantees and remain susceptible to adaptive adversaries (Athalye et al., 2018; Liu et al., 2022; Zhong

et al., 2025), thereby perpetuating a cat-and-mouse dynamic between attackers and defenders, which is not affordable for safety-critical applications. *Therefore, there is an urgent need for a certified prior-training defense that is resilient to imbalanced data.*

To this end, we propose *Randomized Probability Perturbation (RPP)*, **a certified prior-training poisoned sample detection that maintains its robustness across both balanced and imbalanced data.** Unlike existing defenses that rely on distribution-level signals (such as clustering or class-level metrics to separate benign and poisoned samples) which often fail with imbalanced data, RPP examines individual sample behaviors under noise, making it resilient to distribution shifts. Specifically, we observe that poisoned inputs exhibit distinctive, consistent patterns in their prediction probabilities as random perturbations are applied—patterns that are noticeably different from those of benign samples. By quantifying the per-input stability of the predictive probability vector and calibrating detection thresholds via a conformal prediction framework, RPP identifies suspicious samples even when they are rare in imbalanced datasets. This *sample-level* perspective thus bypasses the heavy reliance on uniform or abundant class representations, making RPP naturally suited to both balanced and imbalanced scenarios. Moreover, Sec. 5 derives a certified condition under which backdoored samples are guaranteed to be detectable, yielding provable performance guarantees and enabling practical deployment, especially under severe class imbalance.

Our **contributions** are summarized as follows: **(1)** We investigate and report a strong correlation between the degree of data imbalance and the success rate of backdoor attacks, demonstrating that highly imbalanced datasets are intrinsically more vulnerable to such threats. **(2)** We show that conventional backdoor defense mechanisms degrade significantly in performance as data imbalance intensifies, highlighting the need for new defensive approaches tailored to skewed class distributions. **(3)** Guided by these observations, we introduce RPP, a prior-training poisoned sample detection technique that adapts effectively to severe class imbalance. We further provide certified detection guarantees and a rigorous theoretical foundation, offering a robust framework for future backdoor defense strategies.

2 Related Work

Backdoor attacks can be categorized into two primary types based on their threat models.

(1) Model-Manipulation Backdoor. The first type, model manipulation attacks, operates under a strong but less practical assumption where attackers can control the training process, such as Input-aware Nguyen & Tran (2020), and LIRA Doan et al. (2021). A range of *empirical* works have been proposed to detect or repair trained models to mitigate backdoor threats. Currently, common approaches include Neural Cleanse (Wang et al., 2019) and Artificial Brain Stimulation (Liu et al., 2019), which recover potential triggers to erase the backdoor, as well as fine-tuning on clean auxiliary data (Liu et al., 2017) or directly pruning backdoor-related neurons (Liu et al., 2018a). Some certified defense methods also be proposed (Wang et al., 2020a; Xie et al., 2021; Xiang et al., 2023). *In this work, we do not consider such model manipulation backdoor.*

(2) Data-Poisoning Backdoor. In our work, we focus on the second type with a more practical assumption, data poisoning attacks, which presents a reasonable scenario where attackers have no access to the training process, but can only poison a small portion of the training data by acting as malicious data providers. This category ranges from simple pixel or blend triggers (e.g., BadNets (Gu et al., 2019), Blend (Chen et al., 2017)) to more stealthy or geometric triggers (e.g., ISSBA (Li et al., 2021b), WaNet (Nguyen & Tran, 2021)) and clean-label strategies (e.g., CL (Turner et al., 2019), Sleeper Agent (Souri et al., 2022), SIG (Barni et al., 2019), Narcissus (Zeng et al., 2023)).

Empirical Defenses. The research community has reported a number of empirical defenses aimed at detecting and removing poisoned samples during the training process. These defenses span feature-outlier pruning (SS (Tran et al., 2018)), reverse-engineering optimization for training-set cleansing(RE (Xiang et al., 2021)), maximum-margin activation clipping (MMDF (Wang et al., 2024b)), unlearning to break trigger–label ties (ABL (Li et al., 2021a)), simulator-discrepancy filtering with a clean proxy (De-Pois (Chen et al., 2021)), correlation decoupling via random labels (CT (Qi et al., 2023b)), test-time entropy/perturbation screening (STRIP, BBCaL (Gao et al., 2019; Hu et al., 2024)), and topology/margin heuristics (MM-BD, TED, IBD-PSC (Wang et al., 2024a; Mo et al., 2024; Hou et al., 2024)).

Certified Defenses. However, these empirical approaches lack formal guarantees and are vulnerable to adaptive attacks (Athalye et al., 2018; Liu et al., 2022; Zhong et al., 2025), fueling a cat-and-mouse dynamic that is unacceptable in safety-critical settings. Inspired by certified defenses for adversarial examples (Rosenfeld et al., 2020; Levine & Feizi, 2020; Jia et al., 2022; Weber et al., 2023) that provide guarantee robustness to small input perturbations rather than backdoors, CBD (Xiang et al., 2023) is a *certified* detector of backdoored models. In specific, CBD targets *post-training* model inspection to determine whether a trained model is compromised, rather than detecting and removing poisoned training samples *before* training. A detailed comparison between CBD and our method, RPP, is provided in App. D. *To the best of our knowledge, no prior work provides certified identification and removing of mailicous training samples in data-poisoning backdoor attacks.*

Moreover, existing countermeasures implicitly assume an idealized scenario where defenses operate on balanced training sets—an assumption rarely met in real-world domains. In practice, datasets are often highly imbalanced, with some classes far more than others (He & Garcia, 2009; Fernández et al., 2018; Johnson & Khoshgoftaar, 2019; Aguiar et al., 2024). Our threat investigation (Sec. 3) demonstrates that imbalanced datasets not only significantly increase attack risks (i.e., high attack success rates with low attack budgets) but also drastically reduce the defense performance of existing countermeasures (i.e., low detection accuracy and high false positive rates). In sharp contrast, our proposed method, RPP, is a certified poisoned sample detector that maintains its performance across both balanced and imbalanced data (Sec. 4).

3 Threat Investigation

Threat Model. We focus on backdoor attacks in classification settings under a representative *data-poisoning* threat model, where the adversary can inject poisoned training samples but cannot manipulate the training procedure or final model (Oprea et al., 2022; Cinà et al., 2024). We further consider a practical yet challenging aspect of real-world settings: dataset imbalance. We assume that the attacker is aware of the general distribution of the victim's dataset and how the majority bias in imbalanced datasets affects the success rate of backdoor attacks. As such, the attacker strategically poisons data in minority classes by altering their labels to those of majority classes, which has proven to enhance the attack effectiveness and reduce the attack budget (Sec. 3). *In this paper, we do not consider multi-trigger or multi-target backdoors (Xue et al., 2020), as even empirical detection remains challenging (Xiang et al., 2023).*

Formal Backdoor Attack Setting. We consider K-class classification with inputs $\mathcal{X} \subset \mathbb{R}^d$ and labels $\mathcal{Y} = \{1, \ldots, K\}$. A classifier $f(\cdot|w)$ maps $x \in \mathcal{X}$ to a predictive distribution $\boldsymbol{p}(x|w) \in [0,1]^K$. The y-th entry, $p_y(x|w)$, is defined as $p_y(x|w) = \mathbb{P}(f(x|w) = y), \forall y \in \mathcal{Y}$.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be the collected dataset. In a backdoor attack, an adversary chooses a trigger δ and target class y_t, producing a poisoned set $\mathcal{D}(x+\delta)$ by stamping δ and assigning label y_t, so that $f(x|w) = y$ for benign inputs while $f(x+\delta|w) = y_t$ for triggered ones. We also assume a clean calibration set $\mathcal{V} = \{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from the same distribution as $\mathcal{D}$.

Investigation. Previous studies (Wang et al., 2020b; Pang et al., 2024) have shown that minority classes in long-tail distribution are particularly susceptible to data poisoning attacks. Building on these insights, we investigate whether the inherent classification bias toward majority classes further amplifies this vulnerability in different distributions: specifically, if an attacker deliberately targets underrepresented classes to be misclassified as a majority class, does the existing bias increase the likelihood of the attack's success?

To simulate imbalanced scenarios, we modify the training set in two ways: *long-tailed* (Cui et al., 2019) and *step imbalance* (Buda et al., 2018). The imbalance ratio ρ quantifies the degree of imbalance, defined as the ratio between the largest and smallest class sizes:$\rho = \frac{\max_i\{n_i\}}{\min_i\{n_i\}}$. Unless otherwise specified, we consider $\rho = 2, 10, 100, 200$ (see Fig. 6). In the long-tailed setting, the sizes of the classes decay exponentially from head to tail, leading to a progressive decline across different classes. In the step imbalance setting, we assign a uniform, reduced sample size to all minority classes. We define μ as the proportion of minority classes, typically set at 0.9 (i.e., 9 out of 10 classes are designated as minority) across all experiments to balance the distribution and guarantee a thorough evaluation. More details are provided in App. B.1. In both cases, we apply Badnets (Gu et al., 2019) attacks by randomly selecting an equal number of samples from the minority

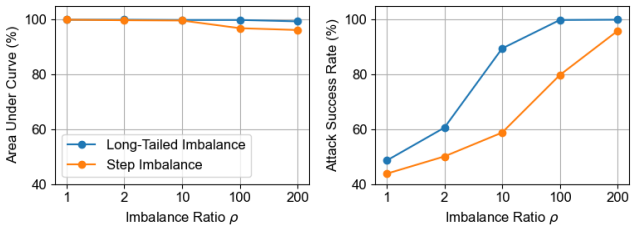

Figure 1: AUC/ASR of Badnets on MNIST with balanced ($\rho = 1$) and imbalanced ($\rho = 2, 10, 100, 200$) training sets; long-tailed ($r = 45$, $p = 0.4\%$) and step-imbalanced ($r = 18$, $p = 0.3\%$).

	AC		ASSET		SCALE-UP	
	TPR	FPR	TPR	FPR	TPR	FPR
ρ=2	61.8	10.4	88.1	38.0	98.9	17.5
ρ=10	57.2	8.9	97.5	53.1	100.0	56.8
ρ=100	31.5	0.8	33.3	52.7	100.0	84.3
ρ=200	0.1	0.1	0.0	0.0	0.1	0.2

Table 1: Comparison with AC, ASSET, and SCALE-UP on imbalanced MNIST ($\mu = 0.9$, $\rho = 2, 10, 100, 200$) against BadNets.

classes (classes 1–9) in each imbalanced training set (or preserving the same poisoning ratio), then relabel them as the majority class (class 0).

Observation 1: Imbalanced datasets are more susceptible to backdoor attacks. Fig. 1 shows that, under the same number of poisoned samples r, increasing the imbalance ratio ρ consistently raises attack success rate (ASR). Even with a fixed poisoning rate p, imbalanced datasets are more vulnerable than balanced ones (App. C.1), corroborating the hypothesis that majority-class bias amplifies backdoor susceptibility.

Observation 2: Existing defenses deteriorate under imbalance due to distribution-level signals. Both empirical and certified defenses, e.g., AC (Chen et al., 2018), ASSET (Pan et al., 2023), and SCALE-UP (Guo et al., 2023), exhibit declining performance as ρ grows (Tab. 1; App. C.13; App. C.14). A key reason is that many defenses rely on *distribution-level* signals, such as classwise statistics, clustering structure, or global thresholds. They become biased and unstable under class imbalance as majority classes dominate these estimates while minority classes yield noisy, underpowered signals, leading to missed triggers. Overrepresentation of majority classes thus steers models and defenses toward majority patterns, under-detecting subtle triggers in minority classes (Tab. 2). Recent work (Wang et al., 2025) adapts a backdoor defense to address post-training bias under data imbalance; its finding on imbalance-induced margin distortion supports our observation of degraded defense reliability.

In practical, datasets often exhibit significant imbalance (Fernández et al., 2018; Johnson & Khoshgoftaar, 2019; Aguiar et al., 2024). Hence, there is a pressing need for backdoor defenses that remain robust to such imbalance. Our proposed RPP method addresses this challenge by providing a certified poisoned samples detection using sample-level signal and calibrating detection thresholds via a conformal prediction framework, whose performance holds in both balanced and imbalanced settings.

4 RPP Detection Using Sample-Level Signal

4.1 Overview of RPP Detection

Key Intuition of Sample-Level Detection under Imbalance. Backdoor triggers are engineered to be *robust*: once present, they reliably drive a poisoned image to the target class and are largely insensitive to simple input transformations (e.g., random noise or blurring) (Chen et al., 2017; Liu et al., 2018b), keeping its predicted probability vector relatively stable. In contrast, clean images rely on naturally occurring correlations without artificially reinforced features; mild perturbations therefore induce noticeably larger changes in their probability vectors.

Motivated by the stability gap above and the limitations under imbalance, we introduce RPP—a *sample-level* robustness metric that measures the expected change in an image's probability vector under injected random noise (formalized in Sec. 4.2). Because RPP evaluates robustness *per instance* rather than from global distributional cues, it remains effective even with severe class imbalance. We further integrate RPP with a conformal prediction framework to calibrate detection thresholds in a distribution-adaptive manner, ensuring

principled performance even under severe class imbalance. Empirically, poisoned samples yield consistently lower RPP than clean ones across diverse imbalance ratios; see Fig. 2 and additional results in App. C.3.

Goal of RPP Detection. Our objective is to determine whether a sample is backdoor-poisoned using RPP, which quantifies the expected perturbation of its predicted probability vector under random noise. Empirically (Fig. 2; App. C.3), RPP exhibits a clear separation between clean and poisoned samples across a range of imbalance ratios, enabling effective identification. Beyond empirical detection, we seek a *certified* condition guaranteeing correct identification of poisoned samples. Crucially, certification must be coupled with control of the false positive rate (FPR); thus, we explicitly target a balance between detection power and a pre-specified FPR for practical deployment.

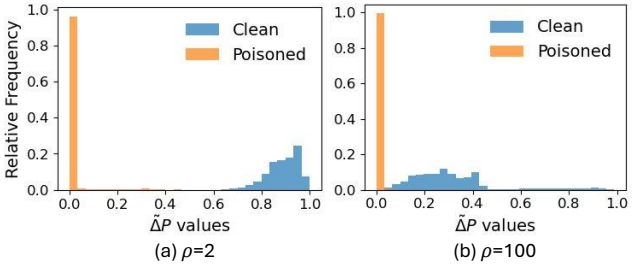

Figure 2: Relative frequency distribution of RPP ($\tilde{\Delta}P$) score on the SVHN dataset with $\mu = 0.9$ under imbalance ratios: (a) $\rho = 2$ and (b) $\rho = 100$.

Outline of RPP Detection. Suppose $\mathcal{D}(x + \delta)$ denotes the potentially poisoned dataset to be inspected, and $\mathcal{V}$ represents a clean calibration set. The detection pipeline is shown in Fig. 3. ***(1) Preliminary Model Training.*** First, we perform a brief training phase on the dataset to be inspected to obtain a preliminary model with acceptable performance. If the dataset contains poisoned samples, we assume that the model will have already learned the backdoor trigger (indicated by a relatively high ASR). ***(2) Threshold Calculation.*** For each x in the clean calibration set $\mathcal{V}$, we inject noise J times, feed the perturbed inputs to the preliminary model $f(\cdot|w)$, and compute the empirical $\tilde{\Delta}P$ (Sec. 4.2) from the resulting probability vectors. The threshold is specified by the quantile linked to a significance level α over the calibration set of size n. ***(3) Poisoned Sample Detection.*** Next, we compare the empirical $\tilde{\Delta}P$ value computed for each sample in $\mathcal{D}(x + \delta)$ to the threshold. If $\tilde{\Delta}P$ is at or below this threshold, x is labeled as poisoned; otherwise, it is classified as clean. The detection threshold is obtained via split conformal prediction (Aljanabi et al., 2023), providing distribution-free calibration that remains valid under class imbalance (Sec. 4.3). The step-by-step procedure is summarized as pseudocode in App. C.12.

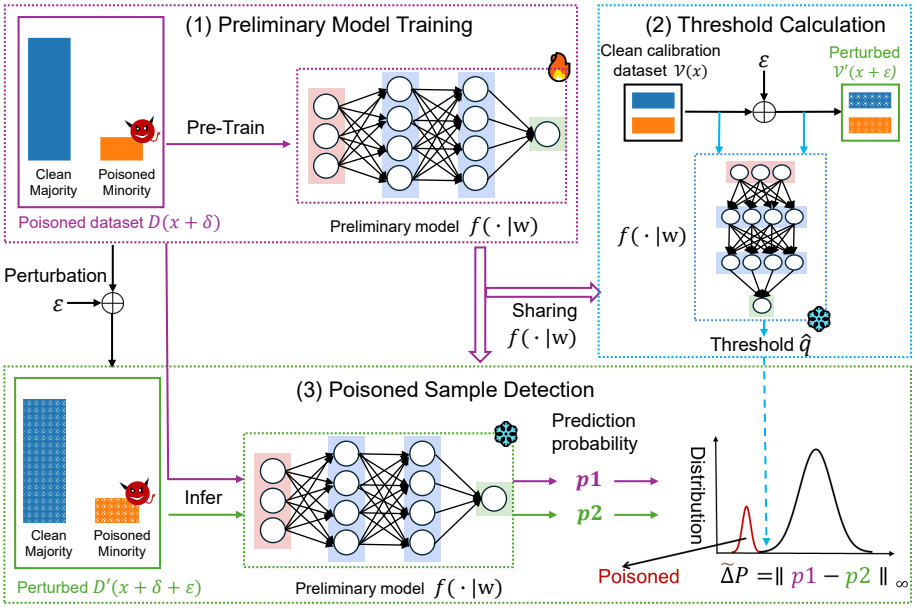

Figure 3: Overview of RPP method.

4.2 Definition of RPP

Inspired by (Xiang et al., 2023), we first define the Samplewise Probability Vector (SPV), representing the prediction probability vector for each sample. Building on this, we define RPP and its empirical version based on the probability vector after noise injection.

Definition 4.1 (Samplewise Probability Vector (SPV))**.** *Let $f(\cdot|w) : \mathcal{X} \to \mathcal{Y}$ be a pre-trained K-class model parameterized by w, where $\mathcal{Y} = \{1, \ldots, K\}$. For any input $x \in \mathcal{X}$, the* SPV *is the predictive probability vector $\mathbf{p}(x|w) \in [0,1]^K$, whose y-th component $p_y(x|w) = \mathbb{P}(f(x|w) = y)$ denotes the probability that $f(x|w)$ assigns x to class y. When isotropic Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is added, the corresponding noisy prediction is $p_y(x+\epsilon|w) = \mathbb{P}(f(x+\epsilon|w) = y)$. If a trigger δ is embedded into the input sample x, the prediction probability for the target class y_t is given by $p_{y_t}(x+\delta|w) = \mathbb{P}\big(f(x+\delta|w) = y_t\big)$, where y_t denotes the attacker-specified target class.*

Definition 4.2 (Randomized Probability Perturbation (RPP) and empirical RPP)**.** *Based on the SPV defined in Def. 4.1, let $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ denote isotropic Gaussian noise with $\sigma > 0$. For a pre-trained model $f(\cdot|w)$ and input x, the RPP is defined as*

$$\Delta P(x|w, \sigma) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I})} \|\boldsymbol{p}(x|w) - \boldsymbol{p}(x+\epsilon|w)\|_\infty \tag{1}$$

Moreover, if $\epsilon_j \in \mathbb{R}^d, j \in [J]$ are J samples independently drawn from the Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$, the empirical RPP (eRPP) is defined as

$$\tilde{\Delta} P(x|w, \sigma) = \frac{1}{J} \sum_{j=1}^{J} \|\boldsymbol{p}(x|w) - \boldsymbol{p}(x+\epsilon_j|w)\|_\infty \tag{2}$$

If the input sample x is embedded with a trigger δ, the corresponding eRPP is given by $\tilde{\Delta} P(x+\delta|w, \sigma) = \frac{1}{J}\sum_{j=1}^{J} \|\boldsymbol{p}(x+\delta|w) - \boldsymbol{p}(x+\delta+\epsilon_j|w)\|_\infty$.

Remarks: The RPP quantifies the bias of the probability vector by adding the randomization on the input. It is trivial that $\Delta P(x|w, \sigma) \to 0$ as $\sigma \to 0$ for any x and w. In practice, we estimate the RPP by it empirical version.

4.3 Threshold Determination of RPP

Although the $\tilde{\Delta} P$ distributions for benign and poisoned samples are distinct (Fig. 2 and App. C.3), determining an appropriate threshold under data imbalance settings for identifying a malicious example remains challenging. To address this challenge, we adapt conformal prediction techniques (Vovk et al., 2005; Angelopoulos et al., 2023), which are particularly well-suited for backdoor detection in imbalanced settings because they require only a single model fit and a small calibration dataset that shares a similar distribution with the test points—a requirement that aligns well with our threat model settings. The specific implementation details of our approach are described below.

Conformal prediction is traditionally used to construct prediction sets for any model (Angelopoulos et al., 2023). Given an i.i.d. calibration set $\{(x_i, y_i)\}_{i=1}^n$, it computes a score function $s(x_i, y_i)$ for each point, then selects a threshold $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of these scores. For a new test example $(x_{\text{test}}, y_{\text{test}})$, the corresponding prediction set $\mathcal{T}(x_{\text{test}})$ satisfies $\mathbb{P}(y_{\text{test}} \in \mathcal{T}(x_{\text{test}})) \geq 1 - \alpha$, where $\alpha \in [0,1]$ is the user-specified error rate. In our approach, instead of producing prediction sets, we focus on the distribution of $\Delta P(x|w, \sigma)$, which allows us to directly apply conformal methodology without constructing class-conditioned sets. Unlike standard conformal prediction, which uses the upper $(1-\alpha)$-quantile to guarantee coverage, our detector uses the lower α-quantile because poisoned samples are expected to yield unusually small RPP scores. We further preserve the i.i.d. assumption by storing a small, balanced, and clean calibration dataset offline, which can be readily satisfied during the initial model deployment phase, where the calibration data is pre-loaded alongside the model; we then re-sample from this dataset—adjusting to the local data distribution—to form a calibration subset that remains contamination-free.

Under this framework, if x_{test} is benign, the usual conformal properties hold; we prove in Sec. 5.2 that $\mathbb{P}(\Delta P(x_{\text{test}}|w, \sigma) \geq \hat{q}) \geq \min\left(1,\ 1 + \frac{1}{n+1} - \alpha\right)$, where $\hat{q}$ is the $\frac{\lceil \alpha(n+1) \rceil}{n}$ quantile of the calibration scores. Equivalently, if $\Delta P(x_{\text{test}}|w, \sigma) \leq \hat{q}$, then with high probability x_{test} has been poisoned. *This makes conformal prediction well-suited for threshold selection under data imbalance, offering a principled basis for backdoor trigger detection in skewed datasets.*

RPP for Imbalanced Dataset. Unlike traditional methods reliant on distribution-level characteristics, RPP evaluates each sample individually based on its stability under noise perturbations. This sample-centric evaluation inherently mitigates the adverse effects of dataset imbalance. The conformal prediction framework further enhances this approach in imbalanced settings by calibrating the detection threshold using empirical RPP distributions from clean validation data. When model bias arises due to imbalance, both RPP values and the corresponding threshold adjust accordingly, preserving detection validity. This ***self-normalizing property*** enables RPP to adapt automatically to varying class distributions without requiring explicit adjustment. As shown in Fig. 2 and App. C.3, the threshold $\hat{q}$ of RPP distributions of clean and poisoned samples are clearly distinguishable with approximately 0.001 for imbalanced datasets with $\rho = 200$, while around 0.5 for balanced datasets.

5 Theoretical Discussion

5.1 RPP Certification

Beyond detection, we provide a certification that guarantees detectability of poisoned samples under specific conditions, with detailed proofs in App. A.

Assumption (A.1) Let y_t be the target class of the attack. We assume the following inequality holds:

$$p_{y_t}(x + \delta + \epsilon|w) < p_{y_t}(x + \delta|w) \tag{3}$$

We note that Assumption (A.1) is a reasonable and common assumption that aligns with standard practices in the field, as evidenced by its adoption in numerous defense mechanisms (Cohen et al., 2019; Xiang et al., 2023). Experiments also used to support the assumption, shown in App. C.4.

Theorem 5.1. *Let $f(\cdot|w) : \mathcal{X} \to \mathcal{Y}$ be the classifier with the parameter w. Let y_t be the target class of the attacker. Define*

$$p(x) = p_{y_t}(x + \delta|w) \tag{4}$$

Suppose that for any specific $x \in \mathcal{X}$ and classes $y_t \in \mathcal{Y}$, there exist $\overline{p_t} \in (0, 1)$ such that

$$\mathbb{P}(f(x + \epsilon|w) = y_t) \leq \overline{p_t}, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is an injected Gaussian noise.

Let $\zeta(x, \delta)$ represent the upper bound of $\Delta P(x + \delta|w, \sigma)$ and suppose Assumption (A.1) is satisfied.

A sample attacked by a backdoor with trigger δ and the target class y_t is guaranteed to be detected if

$$\|\delta\|_2 \geq \sigma\left(\Phi^{-1}(p(x) - \zeta(x, \delta)) - \Phi^{-1}(\overline{p_t})\right) \tag{6}$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function.

Thm. 5.1 establishes a lower bound on the trigger size necessary to detect the poisoned sample. Specifically, a sample attacked by a backdoor is guaranteed to be detected with trigger size $\|\delta\|_2 \geq R = \sigma\left(\Phi^{-1}(p(x) - \zeta(x, \delta)) - \Phi^{-1}(\overline{p_t})\right)$. Notably, the lower bound R in Thm. 5.1 is relatively small and ensures that most poisoned samples can be detected. In practice, if the trigger size is too small ($\|\delta\|_2 \leq R$), the attack success rate remains low under normal settings and attack budgets (see App. C.2).

We provide an upper bound of the trigger size in Cor. 5.2. As long as the backdoor is successfully injected, it implies that $p(x) \to 1$ and $\overline{p_t} \to 0$ (note that $\overline{p_t}$ is the upper bound of $\mathbb{P}(f(x + \epsilon|w) = y_t)$ in equation 5), leading to $\left(\Phi^{-1}(p(x)) - \Phi^{-1}\left(\overline{p_t}\right)\right) \to \infty$ for any given x and w. This means that the upper bound of $\|\delta\|_2$ is an approximately infinite number.

Corollary 5.2. *Suppose all conditions in Thm. 5.1 are satisfied and Assumption (A.1) is satisfied. The trigger δ of a backdoor attack satisfies that*

$$\|\delta\|_2 \leq \sigma\big(\Phi^{-1}(p(x)) - \Phi^{-1}(\overline{p_t})\big) \tag{7}$$

According to Thm. 5.1 and Cor. 5.2, we can establish the range of trigger sizes that guarantee detection. This requires estimating $p(x), \zeta(x, \delta)$, and $\overline{p_t}$. Here, $p(x)$ denotes the probability that the sample $x + \delta$ is classified into the target class. Since the upper bound $\zeta(x, \delta)$ is not directly accessible, we estimate it using the $\frac{\lceil \alpha(n+1) \rceil}{n}$-quantile of the calibration set $\mathcal{S}$. The term $\overline{p_t}$ represents the upper bound of $\mathbb{P}(f(x + \epsilon_j|w) = y_t)$. We let y_{maj} denote the class that appears most frequently among J noise-injected copies of x, where $\epsilon_1, \ldots, \epsilon_J \sim \mathcal{N}(0, \sigma^2 I)$. Specifically, we process each $x + \epsilon_j$ through the pre-trained classifier $f(\cdot|w)$ and count the frequency of each class across the J predictions. The class with the highest frequency is selected as y_{maj} and the count of its occurrences is denoted by n_{maj}. For a given α, we compute $\overline{p_t}$ using one-sided $(1 - \alpha)$ upper confidence intervals of the binomial distribution $B(n_{\text{maj}}, J)$, as implemented in the function $\textsc{UpperConfBound}(n_{\text{maj}}, J, 1 - \alpha)$. The complete algorithm procedure is outlined in App. C.12.

5.2 Theoretical Guarantees

Conformal Calibration Guarantee: We formally present the calibration guarantee, ensuring that a benign example exceeds the threshold with high probability.

Theorem 5.3. *Let $\mathcal{V} = \{(x_i, y_i)\}_{i=1}^n$ be a calibration data set that x_i is benign and i.i.d. drawn from from $\mathcal{D}$. If the test data x_{test} is clean and $\hat{q}$ be the $\frac{\lceil \alpha(n+1) \rceil}{n}$ quantile as*

$$\hat{q} = \inf \left\{ q : \frac{|\{i : s_i \leq q\}|}{n} \geq \frac{\lceil \alpha(n+1) \rceil}{n} \right\}. \tag{8}$$

Then we have that

$$\mathbb{P}(\Delta P(x_{test}|w, \sigma) \geq \hat{q}) \geq \min \left(1,\ 1 + \frac{1}{n+1} - \alpha \right). \tag{9}$$

Performance Guarantee: The theorem establishes a performance guarantee for our proposed robust poisoned sample detection method. Specifically, Thm. 5.4 derives an upper bound for the FPR and outlines its asymptotic behavior. A numerical validation of Thm. 5.4 is provided in App. C.18.

Theorem 5.4. *Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ be the calibration set where $s_i = \Delta P(x_i|w, \sigma)$ for $i = 1, \ldots, n$ and α be a pre-selected confidence level. Then, the FPR can be bounded as the following:*

$$\mathbb{P}\Big(\Delta P(x_{test}|w, \sigma) \leq \hat{q}|\mathcal{S}\Big) \leq \alpha + \frac{1}{n+1} \tag{10}$$

Moreover, if we denote $Z_n = \mathbb{P}\Big(\Delta P(x_{test}|w, \sigma) \leq \hat{q}|\mathcal{S}\Big)$, then for any $\xi > 0$ the following probability holds $\lim_{n\to\infty} \mathbb{P}(Z_n \leq \alpha + \xi) = 1$.

6 Experiment

6.1 Experiment Setting

Imbalanced Datasets. We converted the originally balanced MNIST, SVHN, CIFAR-10, TinyImageNet, and ImageNet10 training sets into long-tailed and step-imbalanced versions based on the imbalance ratio ρ and the fraction of minority classes μ, while keeping the test sets unchanged (see Fig.6). A clean calibration set of i.i.d. samples is drawn from the same distribution. More details about these datasets are deferred to App. B.1. The results for MNIST are presented in the App. C.8. To further evaluate the generalization of RPP in realistic long-tailed settings, we conducted experiments on the naturally imbalanced iNaturalist-2018 dataset (see App. C.9).

Backdoor Attacks. For *certification* performance, we consider two well-known backdoor attacks: Badnets (Gu et al., 2019) and the Blend attack (Chen et al., 2017). In both cases, a trigger δ is added to the original image x such that $\|\delta\|_2 \geq 0.8$ (i.e., $x \mapsto x + \delta$). Although other patterns could be used, our certified robustness depends primarily on the magnitude of the backdoor perturbation and the fraction of poisoned

training data, so these patterns suffice to illustrate our approach. Meanwhile, to *empirically* validate the RPP method's detection efficacy, we evaluate *ten* additional types of triggers, detailed in App. B.2.

Evaluation Metric. Detection performance is measured by TPR (correctly detected poisoned samples) and FPR (benign samples incorrectly flagged as poisoned).

Training. We use PreActResNet18 for SVHN and CIFAR10, ResNet34 for TinyImageNet and ViT-B/16 for ImageNet10 as classification models. The results for different architectures (VGG16, DenseNet161 and EfficientNetB0) are shown in App. C.6. Training details are provided in App. B.3. We ensured that the ASR in the pre-trained models remained above a certain threshold, e.g., 50% (App. C.17) and 90% (Tab. 2), confirming the effectiveness of the backdoor. All reported results are averaged over multiple independent runs to ensure statistical robustness.

6.2 Certification Performance

In Fig. 4, we present the TPR and FPR across dataset SVHN, CIFAR-10, TinyImageNet and ImageNet10 under balanced ($\rho = 1$) and various imbalance ratios ($\rho = 2, 10, 100, 200$). To ensure robustness, we apply isotropic Gaussian noise with dataset-specific standard deviations σ: 0.6 for SVHN, 1.0 for CIFAR-10, TinyImageNet and ImageNet10. The certification process employs two different values for the user-defined error rate parameter α: 0.05 and 0.1. For each input sample, we generate independent Gaussian noise samples ($J = 3$) to effectively measure classification probability variations while maintaining computational efficiency, more results using different J are shown in App. C.5.

Our certification method demonstrates strong effectiveness across all three datasets under various class balance conditions. When $\alpha = 0.05$, for SVHN, RPP achieves certification coverage ranges from 98.8% to 100.0% across all balance scenarios, with FPRs between 1.2% and 23.9%. The CIFAR-10 dataset shows similarly robust results, with certification rates ranging from 90.4% to 99.9% and FPRs between 5.9% and 29.6%. For TinyImageNet, RPP certifies up from 99.3% to 100.0% and FPRs between 2.0% and 31.2%. Even in ImageNet10, the method achieves certification from 84.5% to 96.7%. The corresponding FPR remains relatively low, ranging from 11.6% to 24.6% across different balance ratios. All these optimal results were achieved with $\alpha = 0.05$. When increasing α to 0.1, we observed that, while the TPR remains largely stable, the FPR increases. Based on this trade-off analysis, we selected $\alpha = 0.05$ as the optimal parameter for our subsequent experiments. Additionally, we demonstrate the performance of the RPP across varying trigger sizes, see App. C.17.

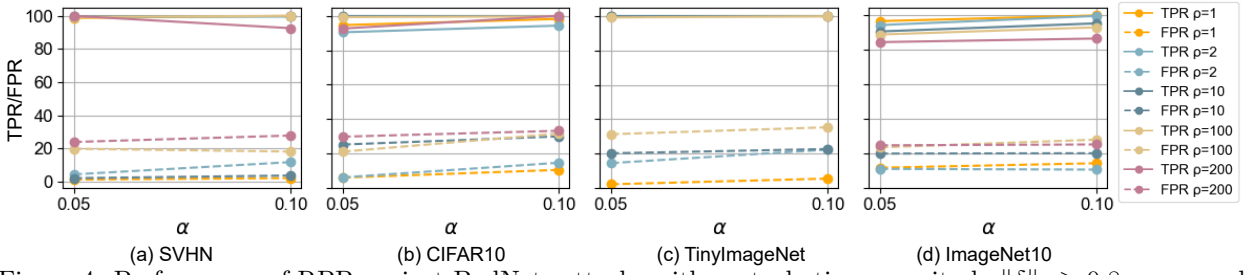

Figure 4: Performance of RPP against BadNets attacks with perturbation magnitude $\|\delta\|_2 \geq 0.8$, measured by TPR and FPR across balanced ($\rho = 1$) and imbalanced ($\rho = 2, 10, 100, 200$) settings, with $n = 100$ and varying α on SVHN, CIFAR-10, TinyImageNet and ImageNet10.

6.3 Empirical Performance

Here, we present the detection performance of RPP against 10 types of backdoor attacks with $\alpha = 0.05$, $n = 100$, and $\sigma = 1.0$. For all experiments, we ensure that the poisoned dataset achieves ASR above 90%. We compare RPP with 12 empirical backdoor defense methods, as shown in Tab. 2 and Tab. 14. For all baselines, we use the official implementations when available and follow the hyperparameters recommended in the original papers. All methods are evaluated on the same datasets, model architectures, poisoning settings, and imbalance ratios, without additional imbalance-specific tuning unless it is part of the original method.

		Badnets		Blend		Trojan		SIG		ISSBA		WaNet		Sleeper Agent		AdaPatch		AdaBlend		Narci.	
	Defenses	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
$\rho = 2$	SS	55.6	1.2	57.9	9.6	54.7	2.0	62.1	25.2	39.5	1.9	60.0	8.2	60.0	2.3	11.7	2.8	16.1	3.9	8.3	7.7
	AC	56.1	16.1	42.7	13.5	24.0	49.6	66.7	16.3	44.1	9.6	58.3	6.7	32.1	20.2	20.7	17.3	21.3	12.6	3.6	21.6
	RE	73.9	17.0	72.7	20.5	56.5	16.9	38.7	11.0	41.5	16.4	50.4	12.3	52.9	11.7	57.8	17.5	50.4	20.8	10.1	12.3
	ABL	23.6	2.6	36.2	8.0	15.2	0.3	10.9	3.2	33.7	0.6	60.3	11.5	4.9	1.3	12.1	0.6	10.3	4.7	0.0	0.5
	CT	75.1	12.6	84.3	10.2	82.1	3.2	90.1	37.6	91.6	18.3	**92.5**	15.7	75.3	10.4	68.7	0.6	77.2	8.5	10.6	16.3
	ASSET	53.3	31.4	55.1	27.6	71.9	37.1	62.2	36.4	80.4	15.6	81.1	16.4	30.5	29.4	32.9	36.9	65.9	22.8	89.0	6.2
	SCALE-UP	95.8	44.0	75.7	28.4	55.7	38.9	**99.0**	**40.1**	90.6	29.8	85.4	21.1	32.6	4.6	76.7	69.2	72.4	39.7	60.5	17.2
	MSPC	88.2	45.1	90.7	20.4	59.8	25.9	80.1	38.9	78.3	18.8	89.8	9.6	67.6	33.8	64.3	34.8	70.3	27.8	77.5	27.1
	BBCaL	95.7	18.4	**94.2**	19.3	72.8	13.0	89.4	19.6	**93.1**	17.7	81.8	20.3	80.4	17.1	70.1	24.4	67.3	20.5	84.0	16.1
	STRIP	90.2	18.9	77.8	26.3	20.7	10.5	88.6	30.7	40.2	22.8	10.5	33.2	17.6	10.3	85.4	24.0	80.9	26.3	0.0	0.8
	TED	96.2	4.8	90.9	8.2	76.5	12.1	88.0	12.9	90.3	18.4	88.4	11.9	89.9	13.3	86.0	13.6	83.9	14.0	90.3	14.0
	IBD-PSC	95.3	6.8	93.6	9.5	**90.3**	10.0	85.2	12.6	90.3	13.6	88.5	12.1	64.3	14.4	85.3	9.7	85.6	14.9	90.5	22.9
	Ours	**96.7**	3.2	89.5	9.3	89.9	12.9	93.0	4.7	89.0	10.9	90.6	8.8	**93.8**	24.7	**87.4**	4.9	**86.7**	11.2	**95.6**	16.9
$\rho = 100$	SS	29.7	5.3	29.5	6.7	58.3	4.2	10.0	5.0	20.8	18.7	30.6	5.5	33.6	20.5	6.7	3.2	20.3	5.2	0.0	1.2
	AC	39.2	0.4	21.2	2.9	0.0	0.1	0.0	0.2	30.0	7.9	32.5	6.8	10.3	0.3	4.6	0.1	10.5	4.6	0.0	0.0
	RE	13.6	22.5	10.4	14.3	20.2	11.0	14.4	15.9	16.4	15.9	17.0	9.4	10.8	8.1	9.7	2.8	8.9	8.4	0.0	3.1
	ABL	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	2.1	10.2	1.6	0.0	2.6	0.0	1.5	0.0	1.8	0.0	1.6
	CT	0.0	14.0	0.0	4.4	0.0	2.4	0.0	0.1	6.1	7.3	21.6	5.2	0.0	12.5	0.0	0.0	9.3	9.3	0.0	9.3
	ASSET	0.0	0.0	12.1	4.7	0.0	0.0	0.0	8.7	10.1	6.4	13.3	5.8	0.0	4.2	21.9	22.8	17.6	6.4	5.2	2.2
	SCALE-UP	90.7	45.3	52.0	41.8	52.1	20.2	48.9	10.8	**63.7**	**32.9**	50.1	24.6	**76.3**	**42.1**	47.1	22.6	33.3	30.5	19.8	17.4
	MSPC	23.4	10.6	36.1	9.5	31.4	11.3	40.8	9.4	21.6	3.3	50.2	16.7	21.5	0.8	19.8	0.9	22.7	6.4	41.1	28.6
	BBCaL	67.6	29.9	68.1	30.6	50.7	19.9	52.0	30.7	58.4	28.9	46.8	31.3	60.3	22.0	39.8	30.1	37.2	28.4	55.7	21.3
	STRIP	49.5	22.4	37.6	29.1	12.0	17.4	42.2	31.1	18.2	20.6	7.9	16.2	8.2	10.4	39.9	32.2	27.6	29.9	0.0	0.5
	TED	76.9	9.3	71.4	16.3	38.7	22.6	70.3	20.4	60.7	20.1	54.3	17.7	61.2	16.9	50.5	20.0	49.4	22.3	65.5	19.5
	IBD-PSC	85.4	12.0	74.7	17.5	47.1	22.9	69.5	16.7	58.8	19.5	63.4	20.3	44.7	15.8	58.9	16.6	63.1	21.4	61.4	29.7
	Ours	**100.0**	9.2	**79.9**	16.3	**75.0**	3.4	**73.9**	12.9	52.2	11.1	**70.3**	24.9	59.2	3.8	**66.1**	15.9	**69.2**	15.4	**67.4**	18.7

Table 2: Comparison with SoTA defenses on imbalanced CIFAR-10 datasets ($\mu = 0.9$, $\rho = 2, 100$). Additional results for $\rho = 1, 10, 200$ are reported in Tab. 14.

RPP achieves comparable or higher TPRs than SOTA detection methods for most trigger types. In several cases, methods such as SCALE-UP and MSPC obtain higher TPRs, as highlighted in red in Tab. 2 and Tab. 14, but at the cost of substantially increased FPR. The sharp degradation of several baselines under severe imbalance is closely related to their reliance on distribution-level signals, such as class-wise statistics, clustering structures, prediction-consistency distributions, or global thresholds. Under severe imbalance, these signals are dominated by majority classes, while minority classes provide noisy and underpowered estimates. This weakens the separation between clean and poisoned samples, leading to reduced TPR or increased FPR. In contrast, RPP evaluates each sample individually through its probability stability under random perturbations and uses clean conformal calibration for thresholding, reducing its dependence on unreliable class-level statistics under imbalance. A detailed per-baseline explanation is provided in App. C.13.

6.4 Additional Experiment

Impact of Noise Level. To assess the impact of isotropic Gaussian noise standard deviation (σ) on RPP, experiments were conducted with fixed $\alpha = 0.05$ and $n = 100$. Fig. 5 illustrates the TPR and FPR across four datasets. Optimal performance requires dataset-specific σ ranges: 0.2 to 1.2 for SVHN, 0.1 to 3.0 for CIFAR10, 0.1 to 2.0 for TinyImageNet and ImageNet10, with variability under 20%, demonstrating RPP's robustness to noise variations.

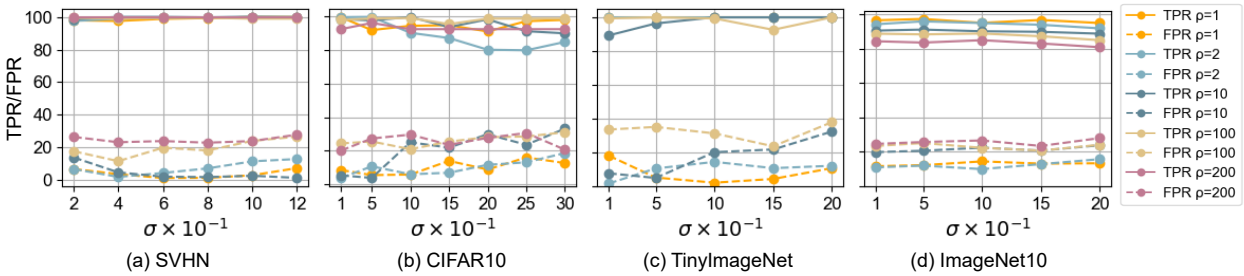

Figure 5: Performance of RPP against BadNets attacks with $\|\delta\|_2 \geq 0.8$, measured by TPR and FPR on SVHN, CIFAR-10, TinyImageNet and ImageNet10 under balanced ($\rho = 1$) and imbalanced ($\rho = 2, 10, 100, 200$) settings for varying σ, with $\alpha = 0.05$ and $n = 100$.

Adaptive Attacks on RPP. We further evaluate RPP under two defense-aware adaptive strategies that directly target the probability-stability signal used by our detector. These adaptive attacks remain within our data-poisoning threat model: the adversary may know the RPP scoring rule and choose the trigger strength,

	ρ=1	ρ=2	ρ=10	ρ=100	ρ=200
TPR	99.7	92.0	76.4	71.0	46.5
FPR	15.9	12.4	22.6	22.8	23.0

Table 3: Performance of RPP against a defense-aware small-norm BadNets attack with $\|\delta\|_2 < 0.8$ on CIFAR-10. The attacker increases the poisoning rate to $p = 10\%$ to maintain ASR above 90%.

	ρ=1	ρ=2	ρ=10	ρ=100	ρ=200
TPR	90.0	81.5	79.0	61.8	59.7
FPR	6.8	13.1	12.5	16.9	26.4

Table 4: Performance of RPP against the hybrid-label adaptive BadNets attack on CIFAR-10. The attack combines label-flipped poisoned samples with noisy-trigger samples that retain their original labels.

poisoning rate, and labeling strategy, but cannot manipulate the training algorithm, modify the learned model after training, or contaminate the clean calibration set used for conformal thresholding. We emphasize that these experiments are intended to test representative evasion strategies against RPP, rather than to claim robustness against all possible adaptive attacks.

(1) Small-norm trigger attack. In realistic backdoor scenarios, attackers typically keep the poisoning rate low to preserve stealth while maintaining a high ASR. This assumption is consistent with common backdoor settings (Souri et al., 2022; Zeng et al., 2023), where small perturbation budgets are used together with low poisoning rates. In our standard setting, a trigger satisfying $\|\delta\|_2 \geq 0.8$ can already achieve a high ASR under low poisoning rates; for example, under step imbalance, a poisoning rate of 0.3% yields ASR above 95% (Fig. 1). In contrast, when the trigger norm is reduced below this regime while keeping the same low poisoning rate, the ASR drops substantially, e.g., to around 20% (App. C.2).

To test whether an adaptive adversary can evade RPP by weakening the trigger, we consider a defense-aware small-norm trigger attack with $\|\delta\|_2 < 0.8$. Since such weak triggers are less effective under low poisoning rates, the attacker increases the poisoning rate to $p = 10\%$ to maintain ASR above 90%. As shown in Tab. 3, RPP still empirically detects a substantial fraction of poisoned samples under this adaptive setting, although the detection performance decreases as imbalance becomes more severe. This result indicates empirical resistance to this specific small-norm evasion strategy. Moreover, extremely small triggers often require much larger poisoning budgets to remain effective, which can reduce their practicality in stealth-constrained threat scenarios.

(2) Hybrid-label adaptive attack. We also evaluate a hybrid-label adaptive attack designed to weaken the probability-stability gap between clean and poisoned samples. In this setting, the adversary poisons 80% of the selected samples through standard label flipping, while the remaining 20% are stamped with triggers combined with Gaussian noise but keep their original labels. This design aims to introduce prediction ambiguity and increase the RPP scores of poisoned samples, thereby making them appear closer to clean samples. Despite this strategy, RPP remains robust (Tab. 4), as the clean-only conformal calibration effectively filters out samples with atypical confidence fluctuations, preserving strong detection performance.

The above evaluation focuses on two adaptive attack strategies that remain within our data-poisoning threat model. Stronger adaptive settings, such as adversaries that can poison the clean calibration set, manipulate the training process against the RPP score, go beyond the threat model considered in this work. We therefore do not further evaluate these settings and leave them as important directions for future work.

More Experiments. In subsequent ablation studies, we systematically investigate several key factors that may affect the performance of RPP. Specifically, we vary the size of the calibration set (App. C.7) and examine the impact of using imbalanced calibration sets (App. C.10). We further explore RPP's robustness when the backdoor target label belongs to a minority or intermediate class (App. C.15), and assess its effectiveness using class-imbalance mitigation techniques such as logits adjustment (App. C.16).

7 Conclusion

In this paper, we show that imbalanced datasets are more vulnerable to backdoor attacks than balanced ones, and existing defenses, designed for balanced data, perform poorly under imbalance. To address this, we propose RPP, the first certified poisoned samples detector that quantifies changes in prediction probabilities

caused by random noise. RPP not only enables accurate detection but also provides a certified condition under which backdoors are guaranteed to be detectable. Our analysis shows that if the backdoor perturbation exceeds a threshold, ensuring high ASR with minimal poisoning, RPP can reliably distinguish poisoned from clean samples. Extensive experiments on five benchmarks confirm its strong detection and certification performance, even in imbalanced scenarios.

8 Broader Impact Statement

RPP offers a robust poisoned samples detection against backdoor threats in machine learning pipelines. It effectively maintain high performance even under challenging adversarial conditions, reducing the risk of malicious exploits while ensuring safer deployment of AI technologies. Consequently, this work has the potential to set new benchmarks in content moderation and ethical AI practices, fostering more secure and socially responsible applications across diverse domains.

References

Gabriel Aguiar, Bartosz Krawczyk, and Alberto Cano. A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *Machine learning*, 113 (7):4165–4243, 2024.

Mohammad Aljanabi, Alaa Hamza Omran, Maad M Mijwil, Mostafa Abotaleb, El-Sayed M El-kenawy, Sahar Yousif Mohammed, and Abdelhameed Ibrahim. Data poisoning: issues, challenges, and needs. In *Proceedings of the 7th IET Smart Cities Symposium (SCS 2023)*, volume 2023, pp. 359–363. IET, 2023.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

K Anusudha et al. Real time face recognition system based on yolo and insightface. *Multimedia Tools and Applications*, 83(11):31893–31910, 2024.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the International conference on machine learning (ICML)*, pp. 274–283, 2018.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 101–105, 2019.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. De-pois: An attack-agnostic defense against data poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 16:3412–3425, 2021.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Machine learning security against data poisoning: Are we there yet? *Computer*, 57(3):26–34, 2024.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the International conference on machine learning (ICML)*, pp. 1310–1320, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9268–9277, 2019.

Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 11966–11976, 2021.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference (ACSAC)*, pp. 113–125, 2019.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Aleksander Song, Dawn, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. *arXiv preprint arXiv:2405.09786*, 2024.

Mengxuan Hu, Zihan Guan, Junfeng Guo, Zhongliang Zhou, Jielu Zhang, and Sheng Li. Bbcal: Black-box backdoor detection under the causality lens. *Transactions on Machine Learning Research*, 2024.

Jinyuan Jia, Yupei Liu, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of nearest neighbors against data poisoning and backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9575–9583, 2022.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54, 2019.

Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14254–14263, 2020.

Quentin Le Roux, Eric Bourbao, Yannick Teglia, and Kassem Kallas. A comprehensive survey on backdoor attacks and their defenses in face recognition systems. *IEEE Access*, 12:47433–47468, 2024.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defense against general poisoning attacks. *arXiv preprint arXiv:2006.14768*, 2020.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021a.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 16463–16472, 2021b.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of the International symposium on research in attacks, intrusions, and defenses (RAID)*, pp. 273–294, 2018a.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of the 25th Annual Network And Distributed System Security Symposium (NDSS)*, 2018b.

Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1265–1282, 2019.

Yinpeng Liu, Yao Cheng, Li Gao, Xiaoyu Liu, Qiang Zhang, and Junchi Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15105–15114, 2022.

Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *Proceedings of the 2017 IEEE International Conference on Computer Design (ICCD)*, pp. 45–48, 2017.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Xiaoxing Mo, Yechao Zhang, Leo Yu Zhang, Wei Luo, Nan Sun, Shengshan Hu, Shang Gao, and Yang Xiang. Robust backdoor detection for deep learning via topological evolution dynamics. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (S&P)*, pp. 2048–2066, 2024.

Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020.

Alina Oprea, Anoop Singhal, and Apostol Vassilev. Poisoning attacks against machine learning: Can machine learning be trustworthy? *Computer*, 55(11):94–99, 2022.

Soumyadeep Pal, Yuguang Yao, Ren Wang, Bingquan Shen, and Sijia Liu. Backdoor secrets unveiled: Identifying backdoor data with optimized scaled prediction consistency. *arXiv preprint arXiv:2403.10717*, 2024.

Minzhou Pan, Yi Zeng, Lingjuan Lyu, Xue Lin, and Ruoxi Jia. {ASSET}: Robust backdoor data detection across a multiplicity of deep learning paradigms. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2725–2742, 2023.

Lu Pang, Tao Sun, Weimin Lyu, Haibin Ling, and Chao Chen. Long-tailed backdoor attack using dynamic data augmentation operations. *arXiv preprint arXiv:2410.12955*, 2024.

Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023a.

Xiangyu Qi, Tinghao Xie, Jiachen T Wang, Tong Wu, Saeed Mahloujifar, and Prateek Mittal. Towards a proactive {ML} approach for detecting backdoor poison samples. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1685–1702, 2023b.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *Proceedings of the International conference on machine learning (ICML)*, pp. 8230–8241, 2020.

Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35:19165–19178, 2022.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8769–8778, 2018.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian conference on machine learning (ACML)*, pp. 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong, et al. On certifying robustness against backdoor attacks via randomized smoothing. *arXiv preprint arXiv:2002.11750*, 2020a.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the 2019 IEEE symposium on security and privacy (S&P)*, pp. 707–723, 2019.

Hang Wang, Zhen Xiang, David J. Miller, and George Kesidis. Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In *Proceedings of the 45th IEEE Symposium on Security and Privacy (S&P)*, pp. 1994–2012, 2024a. doi: 10.1109/SP54263.2024.00015.

Hang Wang, Zhen Xiang, David J Miller, and George Kesidis. Improved activation clipping for universal backdoor mitigation and test-time detection. In *Proceedings of the IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2024b.

Hang Wang, David J Miller, and George Kesidis. Maximum margin based activation clipping for post-training overfitting mitigation in dnn classifiers. *IEEE Transactions on Artificial Intelligence*, 2025.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in neural information processing systems*, 33:16070–16084, 2020b.

Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. Rab: Provable robustness against backdoor attacks. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy (S&P)*, pp. 1311–1328, 2023.

Li-Hua Wen and Kang-Hyun Jo. Deep learning-based perception systems for autonomous driving: A comprehensive survey. *Neurocomputing*, 489:255–270, 2022.

Zhen Xiang, David J Miller, and George Kesidis. Detection of backdoors in trained classifiers without access to the training set. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3):1177–1191, 2020.

Zhen Xiang, David J Miller, and George Kesidis. Reverse engineering imperceptible backdoor attacks on deep neural networks for detection and training set cleansing. *Computers & Security*, 106:102280, 2021.

Zhen Xiang, Zidi Xiong, and Bo Li. Cbd: A certified backdoor detector based on local dominant probability. *Advances in Neural Information Processing Systems*, 36:4937–4951, 2023.

Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. Crfl: Certifiably robust federated learning against backdoor attacks. In *Proceedings of the International conference on machine learning (ICML)*, pp. 11372–11382, 2021.

Mingfu Xue, Can He, Jian Wang, and Weiqiang Liu. One-to-n & n-to-one: Two advanced backdoor attacks against deep learning models. *IEEE Transactions on Dependable and Secure Computing*, 19(3):1562–1578, 2020.

Xiaolong Yang, Xiaohong Jia, Dihong Gong, Dong-Ming Yan, Zhifeng Li, and Wei Liu. Larnet: Lie algebra residual network for face recognition. In *Proceedings of the International conference on machine learning (ICML)*, pp. 11738–11750, 2021.

Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 771–785, 2023.

Zhuang Zhong, Wen Bao, Jun Wang, Jie Chen, Lingjuan Lyu, and W. Y. B. Lim. Sacfl: Self-adaptive federated continual learning for resource-constrained end devices. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. Early Access.

Appendices

Contents

- **Section A: Proof of Theorems**
  - A.1: Proof of Theorem 5.1
  - A.2: Proof of Corollary 5.2
  - A.3: Proof of Theorem 5.3
  - A.4: Proof of Theorem 5.4
- **Section B: Details for the Experimental Setting**
  - B.1: Details for the Imbalanced Datasets
  - B.2: Examples of Poisoned Samples
  - B.3: Details for the Model Training
- **Section C: More Experiments**
  - C.1: Attack Success Rate with Same Poisoning Ratio
  - C.2: Attack Success Rate with Smaller Trigger Size
  - C.3: Relative Frequency Distributions of RPP
  - C.4: Empirical Validation for Assumption (A.1)
  - C.5: Performance of Different Noise Samples
  - C.6: Performance on Different Model Architectures
  - C.7: Impact of Calibration Set Size
  - C.8: Performance on MNIST
  - C.9: Performance on iNaturalist
  - C.10: Performance under Imbalanced Calibration Set
  - C.11: Impact of Preliminary Model Quality
  - C.12: Algorithm for RPP Detection and Certification
  - C.13: Baselines for Backdoor Sample Detection
  - C.14: Certified Defenses under Imbalanced Dataset
  - C.15: Impact of Target Label Class
  - C.16: Performance under Logits-Adjusted Training
  - C.17: Performance against Different Trigger Sizes

  - C.18: Validation of the FPR Upper Bound Guarantee

- **Section D: Differences between RPP and CBD**
- **Section E: Computational Complexity**
- **Section F: Limitations & Future Work**
- **Section G: Use of Large Language Models (LLMs)**

A Proof of Theorems

A.1 Proof of theorem 5.1

We restate theorem 5.1 as theorem A.1 to enhance clarity and convenience. The proof is motivated by the approach in Cohen et al. (2019), which leverages Lemma 4. For completeness, we restate this lemma as theorem A.2 and introduce theorem A.3 (Assumption A.1) below.

Theorem A.1 (theorem 5.1)**.** *Let $f(\cdot|w) : \mathcal{X} \to \mathcal{Y}$ be the classifier with the parameter w. Let y_t be the target class of the attacker. Define*

$$p(x) = p_{y_t}(x + \delta|w) \tag{11}$$

Suppose that for any specific $x \in \mathcal{X}$ and classes $y_t \in \mathcal{Y}$, there exist

$$\mathbb{P}(f(x + \epsilon|w) = y_t) \leq \overline{p_t} \tag{12}$$

Let $\zeta(x, \delta)$ represent the upper bound of $\Delta P(x + \delta|w, \sigma)$ and suppose theorem A.3 is satisfied. A sample attacked by a backdoor with a trigger δ and the target class y_t is guaranteed to be detected if

$$\|\delta\|_2 \geq \sigma\big(\Phi^{-1}(p(x) - \zeta(x, \delta)\big) - \Phi^{-1}(\overline{p_t})) \tag{13}$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function.

Lemma A.2 (Lemma 4, (Cohen et al., 2019))**.** *Let $X \sim \mathcal{N}(x, \sigma^2 I)$ and $Y \sim \mathcal{N}(x+\delta, \sigma^2 I)$. Let $h : \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function. Then:*

1. *If $S = \{z \in \mathbb{R}^d : \delta^\top z \leq \beta\}$ for some β and $\mathbb{P}(h(X) = 1) \geq \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \geq \mathbb{P}(Y \in S)$.*
2. *If $S = \{z \in \mathbb{R}^d : \delta^\top z \geq \beta\}$ for some β and $\mathbb{P}(h(X) = 1) \leq \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \leq \mathbb{P}(Y \in S)$.*

Assumption A.3. *Let y_t be the target class of the attack. We assume the following inequality holds:*

$$p_{y_t}(x + \delta + \epsilon|w) < p_{y_t}(x + \delta|w) \tag{14}$$

Proof of theorem A.1. We define the following half-spaces:

$$A := \{z : \delta^\top(z - x) \geq \sigma\|\delta\|_2 \Phi^{-1}(1 - \overline{p_t})\} \tag{15}$$

The following lemma will be used in the proof, with its proof provided at the end.

Lemma A.4. *Let $\overline{p_t} \in (0, 1)$ is a number defined in equation 12 and A are defined by equation 15. The following equality holds:*

$$\mathbb{P}(x + \epsilon \in A) = \overline{p_t} \tag{16}$$

According to theorem A.4 and the definitions of $\overline{p_t}$, we have that

$$\mathbb{P}(f(x + \epsilon|w) = y_t) \leq \mathbb{P}(x + \epsilon \in A) \tag{17}$$

which implies the following:

$$p_{y_t}(x + \delta + \epsilon|w) = \mathbb{P}(f(x + \delta + \epsilon|w) = y_t) \leq \mathbb{P}(x + \delta + \epsilon \in A) \tag{18}$$

by applying theorem A.2 with $h(z) := \mathbf{1}[f(z) = y_t]$.

Now,

$$\begin{aligned}
\mathbb{P}(x+\delta+\epsilon \in A) &= \mathbb{P}(\delta^\top(x+\delta+\epsilon-x) \geq \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_t})) \\
&= \mathbb{P}(\delta^\top\mathcal{N}(0,\sigma^2 I) + \|\delta\|_2^2 \geq \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_t})) \\
&= \mathbb{P}(\sigma\|\delta\|_2 Z \geq \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_t}) - \|\delta\|_2^2) \\
&= \mathbb{P}\left(Z \geq \Phi^{-1}(1-\overline{p_t}) - \frac{\|\delta\|_2}{\sigma}\right) \\
&= \Phi(\Phi^{-1}(\overline{p_t}) + \frac{\|\delta\|_2}{\sigma})
\end{aligned} \tag{19}$$

where $Z \sim \mathcal{N}(0,1)$ is the standard Gaussian distribution.

For any $x+\delta \in \mathcal{Y}$, we have that

$$\begin{aligned}
\|\boldsymbol{p}(x+\delta|w) - \boldsymbol{p}(x+\delta+\epsilon|w)\|_\infty &= \max_{1\leq k\leq K} |p_k(x+\delta|w) - p_k(x+\delta+\epsilon|w)| \\
&\geq |p_{y_t}(x+\delta|w) - p_{y_t}(x+\delta+\epsilon|w)| \\
&\overset{\text{(I)}}{=} p_{y_t}(x+\delta|w) - p_{y_t}(x+\delta+\epsilon|w) \\
&\overset{\text{(II)}}{\geq} p(x) - \mathbb{P}(x+\delta+\epsilon \in A) \\
&\overset{\text{(III)}}{\geq} p(x) - \Phi(\Phi^{-1}(\overline{p_t}) + \frac{\|\delta\|_2}{\sigma})
\end{aligned} \tag{20}$$

where (I) holds because of theorem A.3, (II) follows from equation 18 and (III) follows from equation 19. Taking the expectation over $\epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I})$, we obtain that

$$p(x) - \Phi(\Phi^{-1}(\overline{p_t}) + \frac{\|\delta\|_2}{\sigma}) \leq \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\|\boldsymbol{p}(x+\delta|w) - \boldsymbol{p}(x+\delta+\epsilon|w)\|_\infty \tag{21}$$

Let $\zeta = \zeta(x,\delta)$ be the upper bound of $\Delta P(x+\delta|w,\sigma)$. From equation 20, we have that

$$0 < p(x) - \Phi(\Phi^{-1}(\overline{p_t}) + \frac{\|\delta\|_2}{\sigma}) \leq \zeta(x,\delta) \tag{22}$$

which implies that

$$\|\delta\|_2 \geq \sigma\big(\Phi^{-1}(p(x) - \zeta(x,\delta)\big) - \Phi^{-1}(\overline{p_t})) \tag{23}$$

□

Proof of theorem A.4. The proof of theorem A.4 is similar to the proofs of two claims in Cohen et al. (2019, Theorem 2). We provide them here for completeness.

Recall that $\epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I})$ and $A = \{z : \delta^\top(z-x) \geq \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_t})\}$. Then,

$$\begin{aligned}
\mathbb{P}(x+\epsilon \in A) &= \mathbb{P}(\delta^\top(x+\epsilon-x) \geq \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_t})) \\
&= \mathbb{P}(\delta^\top\mathcal{N}(0,\sigma^2\mathbf{I}) \geq \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_t})) \\
&= \mathbb{P}(\sigma\|\delta\|_2 Z \geq \sigma\|\delta\|_2\Phi^{-1}(1-\overline{p_t})) \\
&= \mathbb{P}(Z \geq \Phi^{-1}(1-\overline{p_t})) \\
&= 1 - \Phi(\Phi^{-1}(1-\overline{p_t})) = \overline{p_t}.
\end{aligned}$$

□

A.2 Proof of theorem 5.2

From the inequality of equation 22, we have that

$$\Phi(\Phi^{-1}(\overline{p_t}) + \frac{\|\delta\|_2}{\sigma}) \leq p(x) \tag{24}$$

which implies that

$$\|\delta\|_2 \leq \sigma\big(\Phi^{-1}(p(x)) - \Phi^{-1}(\overline{p_t})\big)$$

A.3 Proof of theorem 5.3

Let $s_i = \Delta P(x_i|w,\sigma)$ for $i = 1,\ldots,n$ and $s_0 = \Delta P(x_{\text{test}}|w,\sigma)$. By the exchangeability of x_i and x_{test} (they are i.i.d.), we have that

$$\mathbb{P}(s_0 \geq s_k) = \frac{n+2-k}{n+1}, \quad \forall k \in [n],$$

which is because s_0 is equally likely to fall in anywhere between $s_1,\ldots,s_n$. Hence, we can conclude that

$$\begin{aligned}\mathbb{P}(s_0 \geq s_{\lceil\alpha(k+1)\rceil}) &= \frac{n+2-\lceil\alpha(n+1)\rceil}{n+1} \\ &\geq 1 + \frac{1}{n+1} - \alpha.\end{aligned}$$

A.4 Proof of theorem 5.4

Let $\mathcal{S} = \{s_1,\ldots,s_n\}$ be the calibration set where $s_i = \Delta P(x_i|w,\sigma)$ for $i = 1,\ldots,n$. Moreover, we denote $s_0 = \Delta P(x_{\text{test}}|w,\sigma)$. Similar to the proof of theorem 5.3, by the exchangeability of x_i and x_{test} (assume x_{test} is benign), we have that

$$\mathbb{P}\Big(\Delta P(x_{\text{test}}|w,\sigma) \leq \hat{q}|\mathcal{S}\Big) = \mathbb{P}(s_0 \leq s_{\lceil\alpha(n+1)\rceil}|\mathcal{S}) = \frac{\lceil\alpha(n+1)\rceil}{n+1} \leq \alpha + \frac{1}{n+1},$$

which proves the upper bound. Next, we discuss the asymptotic property of FPR. In the classic conformal prediction, the distribution of coverage, $\mathbb{P}(y_{\text{test}} \in \mathcal{T}(x_{\text{test}}))$, has an analytic form and it was first introduced by Vladimir Vovk in Vovk (2012), i.e.,

$$\mathbb{P}(y_{\text{test}} \in \mathcal{T}(x_{\text{test}})|\mathcal{S}) \sim \text{Beta}(n+1-\lfloor\alpha(n+1)\rfloor, \lfloor\alpha(n+1)\rfloor).$$

The details of the proof of this fact see Vovk (2012). We note that the probability of coverage is actually equal to $\mathbb{P}(s_0 \leq s_{\lceil\alpha(k+1)\rceil})$. Indeed, during the proof of theorem 5.3, we just skip to show the probability of coverage and choose the $\frac{\lfloor\alpha(n+1)\rfloor}{n}$ quantile as the threshold while the $\frac{\lfloor(1-\alpha)(n+1)\rfloor}{n}$ quantile is used in the classic conformal prediction. Therefore, we have that $\mathbb{P}\Big(\Delta P(x_{\text{test}}|w,\sigma) \leq \hat{q}|\mathcal{S}\Big) \sim \text{Beta}(n+1-\lfloor(1-\alpha)(n+1)\rfloor, \lfloor(1-\alpha)(n+1)\rfloor)$. If we denote $Z_n = \mathbb{P}(s_0 \leq s_{\lceil\alpha(k+1)\rceil}|\mathcal{S})$, the mean and variance of Z_n are given by

$$\begin{aligned}\mathbb{E}[Z_n] &= \frac{n+1-\lfloor(1-\alpha)(n+1)\rfloor}{n+1} = \frac{\lceil\alpha(n+1)\rceil}{n+1} \\ \text{Var}[Z_n] &= \frac{(n+1-\lfloor(1-\alpha)(n+1)\rfloor)(\lfloor(1-\alpha)(n+1)\rfloor)}{(n+1)^2(n+2)}\end{aligned}$$

From the Chebyshev's inequality, we have that for any $\xi > 0$,

$$\mathbb{P}(Z_n \geq \frac{\lceil\alpha(n+1)\rceil}{n+1} + \xi) < \mathbb{P}(|Z_n - \frac{\lceil\alpha(n+1)\rceil}{n+1}| \geq \xi) \leq \frac{1}{\xi^2}\frac{\lceil\alpha(n+1)\rceil)(\lfloor(1-\alpha)(n+1)\rfloor)}{(n+1)^2(n+2)},$$

which implies that

$$\lim_{n\to\infty}\mathbb{P}(Z_n \leq \alpha+\xi) = \lim_{n\to\infty}\mathbb{P}(Z_n \leq \frac{\lceil\alpha(n+1)\rceil}{n+1} + \xi) = 1 - \lim_{n\to\infty}\mathbb{P}(Z_n \geq \frac{\lceil\alpha(n+1)\rceil}{n+1} + \xi) = 1.$$

B Details for the Experimental Setting

B.1 Details for the Imbalanced Datasets

The original MNIST, SVHN, and CIFAR-10 datasets consist of 60,000 training images and 10,000 test images (MNIST), 73,257 training images and 26,032 test images (SVHN), and 50,000 training images and 10,000 test images (CIFAR-10), with both datasets having 10 classes. For TinyImageNet, we selected 50 classes from TinyImageNet200 as a subset for our experiments, with each class containing 500 training images. For ImageNet-10, we constructed a 10-class subset of ILSVRC-2012 (ImageNet-1K) by selecting ten semantically diverse categories. We use the original per-class splits (approximately 1,300 training images and 50 validation images per class).To create imbalanced versions, we reduce the number of training samples per class while keeping the test sets unchanged.

We use two types of imbalance: long-tailed (Cui et al., 2019) and step imbalance(Buda et al., 2018). The imbalance ratio ρ measures the degree of imbalance, defined as the ratio between the largest and smallest class sizes:$\rho = \frac{\max_i\{n_i\}}{\min_i\{n_i\}}$.

In our studies, we designated imbalance ratios of $\rho = 2, 10, 100, 200$, as depicted in Fig. 6. For TinyImageNet, the ratios were set at $\rho = 2, 10, 100$. In scenarios with long-tailed distributions, the sizes of the classes decrease exponentially, leading to a progressive decline across different classes. Conversely, the step imbalance method assigns a uniform, reduced sample size to all minority classes, while maintaining larger sample sizes for more frequent classes, thus establishing a distinct separation between the two groups. We define μ as the proportion of minority classes, typically set at 0.9 across all experiments to balance the distribution and guarantee a thorough evaluation, except for TinyImageNet where $\mu = 0.98$.

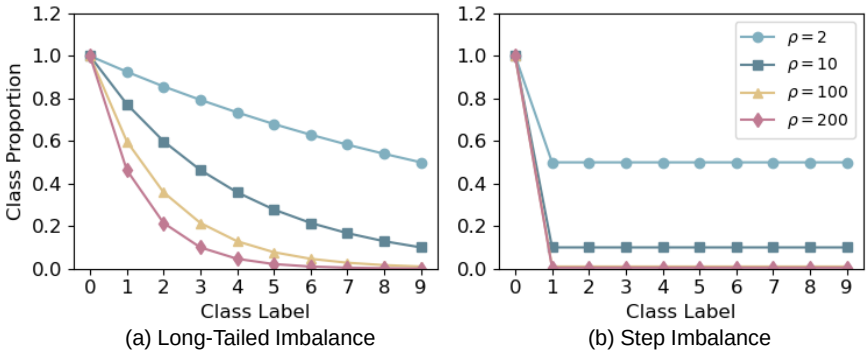

Figure 6: Class proportion types on the MNIST, SVHN, CIFAR10, TinyImageNet and ImageNet datasets. (a) Long-Tailed Imbalance ($\mu = 1/(\rho + 1)$, ρ = 2, 10, 100, 200). (b) Step Imbalance ($\mu = 0.9$, ρ = 2, 10, 100, 200).

B.2 Examples of Poisoned Samples

For **certification** performance, we consider two well-known backdoor attacks: Badnets (Gu et al., 2019) and the Blend attack (Chen et al., 2017). In both cases, a trigger δ is added to the original image x such that $\|\delta\|_2 \geq 0.8$ (i.e., $x \mapsto x + \delta$). For Badnets, we use a chessboard pattern (Xiang et al., 2020) as the trigger; for the Blend attack, we adopt a "hello-kitty" image with a blending rate of 0.2. Although other patterns could be used, our certified robustness depends primarily on the magnitude of the backdoor perturbation and the fraction of poisoned training data, so these patterns suffice to illustrate our approach.

Meanwhile, to **empirically** validate the RPP method's detection efficacy, we evaluate 10 types of triggers. These include Chessboard (Xiang et al., 2020), Blend (Chen et al., 2017), Trojan (Liu et al., 2018b), Sleeper Agent (Souri et al., 2022), SIG (Barni et al., 2019),ISSBA(Li et al., 2021b)), WaNet(Nguyen & Tran, 2021), AdaPatch (Qi et al., 2023a), AdaBlend Patterns (Qi et al., 2023a) and Narcissus (Narci.) (Zeng et al., 2023), detailed in 6.3. Here, we showcase original clean samples alongside their corresponding poisoned counterparts, each embedded with one of these nine distinct triggers, as shown in Fig. 7.

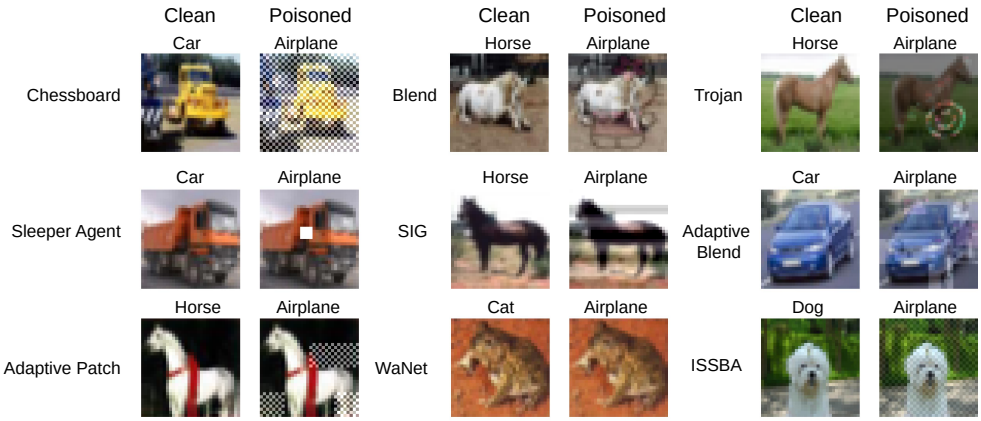

Figure 7: Examples of nine types of backdoor attacks on CIFAR10.

B.3 Details for the Model Training

We employ a standard two-layer convolutional neural network (Tab. 5) for MNIST, trained for 50 epochs with a batch size of 128 and a learning rate of 10^{-3} using the Adam optimizer.

For SVHN and CIFAR-10, we train PreActResNet18 (He et al., 2016) for 100 epochs at a batch size of 128 and a learning rate of 5×10^{-4} using Adam. The training images are augmented by random cropping and horizontal flipping.

For TinyImageNet, we train ResNet34 (He et al., 2016) for 100 epochs at a batch size of 128 and a learning rate of 5×10^{-4} using Adam. Data augmentation includes random horizontal flipping and random cropping.

For ImageNet10, we train ViT-B/16 (Dosovitskiy, 2020) for 100 epochs at a batch size of 128 and a learning rate of 5×10^{-4} using Adam. Data augmentation includes random horizontal flipping and random cropping.

Layer	# of Channels	Filter Size	Stride	Activation
Conv	32	3×3	1	ReLU
MaxPool	32	2×2	2	-
Conv	64	3×3	1	ReLU
MaxPool	64	2×2	2	-
FC	128	-	-	ReLU
FC	10	-	-	Softmax

Table 5: Model Architecture for MNIST.

C More Experiments

C.1 Attack Success Rate with Same Poisoning Ratio

In Fig. 1, we observe that when an equal number of minority class samples (classes 1-9) are randomly selected and subjected to Badnets backdoor attacks (Gu et al., 2019) with their labels changed to the majority class (class 0), the ASR in imbalanced datasets is higher than in balanced datasets. In Fig. 8, we preserved the same poisoning ratio (p) across balanced ($\rho = 1$) and imbalanced training set ($\rho = 200$), and implemented the same backdoor attacks. Our findings also suggest that imbalanced datasets are more vulnerable to backdoor attacks compared to their balanced counterparts.

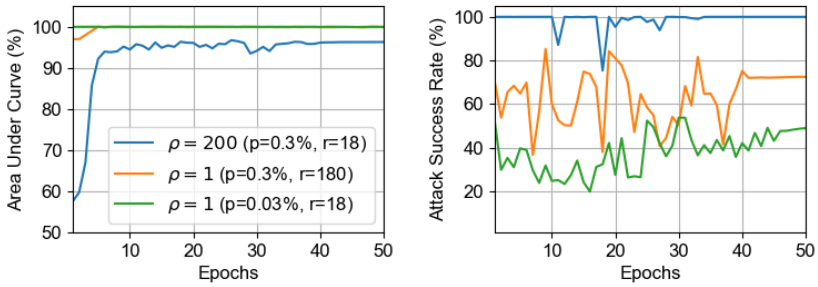

Figure 8: The AUC and ASR of the Badnets backdoor attacks on balanced ($\rho = 1$) and imbalanced ($\mu = 0.9, \rho = 200$) MNIST training datasets with the same poisoned ratios (p = 0.3%) or same number (r = 18) of backdoor samples.

Attack	$\rho = 1$		$\rho = 10$		$\rho = 200$	
	AUC	ASR	AUC	ASR	AUC	ASR
Badnets	100.0	13.8	100.0	14.5	98.9	21.8
Blend	100.0	16.3	100.0	18.8	99.0	22.9
Trojan	100.0	29.4	100.0	30.6	99.0	35.6
SIG	100.0	29.0	100.0	33.7	99.0	38.2
ISSBA	99.8	30.2	99.8	47.6	98.8	53.3
WaNet	100.0	35.8	99.7	46.0	98.9	55.1
Sleeper Agent	99.9	28.6	99.5	36.9	99.0	48.6
Adaptive patch	99.8	20.1	99.8	26.9	98.5	33.0
Adaptive Blend	99.9	22.7	99.0	26.1	98.9	30.4

Table 6: The ASR and AUC under $\|\delta\|_2 < 0.8$ on the balanced MNIST dataset ($\rho = 1$) and two types of imbalanced MNIST datasets ($\mu = 0.9$, $\rho = 10$, 200) with same poisoned ratio (p=0.3%).

C.2 Attack Success Rate with Smaller Trigger Size

Typically, to ensure stealth, attackers aim to minimize the amount of poisoning while maintaining a high success rate for the attack. Our preliminary findings indicate that in datasets with extreme imbalance, a poisoning rate of 0.3% in step imbalance can achieve an ASR of more than 95%, as shown in Fig. 1.

As demonstrated in Thm. 5.1, a sample attacked by a backdoor is guaranteed to be detected with trigger size $\|\delta\|_2 \geq R = \sigma\left(\Phi^{-1}(p(x) - \zeta(x,\delta)) - \Phi^{-1}(\overline{p_t})\right)$. In this section, we examine scenarios where the trigger size is smaller than $R_{\sigma,t}$, while keeping the poisoning rate constant. As shown in Tab. 6, all evaluated attacks struggle to maintain high ASR under such stealthy constraints, confirming the general difficulty of successful backdoor injection when both poisoning budget and perturbation size are limited. In practice, defensive measures are not required when the ASR drops below 50%.

C.3 Relative Frequency Distributions of RPP

In Sec. 4.1, we present the relative frequency distributions of $\tilde{\Delta}P$ for clean and malicious samples under datasets with imbalance ratios $\rho = 2$ and $\rho = 100$. In Fig. 9, the relative frequency distribution of $\tilde{\Delta}P$ for clean and malicious samples in balanced datasets ($\rho = 1$) as well as those with imbalance ratios $\rho = 10$ and $\rho = 200$ are depicted. It is observed that images subjected to backdoor attacks typically exhibit a lower $\tilde{\Delta}P$ compared to clean images. Although there is some overlap in the $\tilde{\Delta}P$ between poisoned data and clean samples when $\rho = 200$, the RPP method still achieves TPR of 90.0% with FPR below 30% (Fig. 4). Moreover, datasets with a step imbalance of $\rho = 200$ are also rare in real-world scenarios.

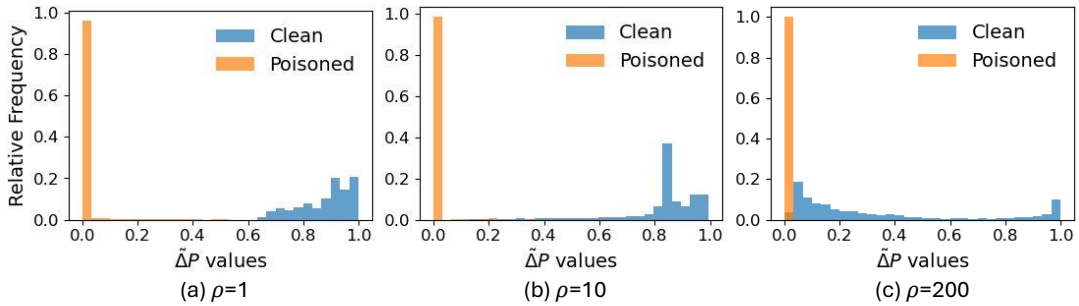

Figure 9: Relative frequency distribution of $\tilde{\Delta}P$ for clean and poisoned samples on SVHN dataset with imbalance ratios: (a) $\rho = 1$, (b) $\mu = 0.9$, $\rho = 10$ and (c) $\mu = 0.9$, $\rho = 200$.

Attack	$\rho = 1$	$\rho = 2$	$\rho = 10$	$\rho = 100$	$\rho = 200$
Badnets	100.0	100.0	100.0	100.0	99.8
Blend	100.0	100.0	100.0	100.0	99.6
Trojan	100.0	100.0	100.0	100.0	99.7
Narcissus	100.0	100.0	100.0	100.0	100.0

Table 7: Percentage of poisoned samples satisfying Assumption 1 under different imbalance ratios ρ with $\sigma = 1.0$ on CIFAR-10, evaluated on four backdoor attacks: Badnets, Blend, Trojan, and Narcissus.

C.4 Empirical Validation for Assumption (A.1)

To empirical validate the Assumption (A.1): $p_{y_t}(x + \delta + \epsilon|w) < p_{y_t}(x + \delta|w)$, we evaluated the percentage of backdoored test samples satisfying this inequality on CIFAR-10 under varying imbalance ratios, with Gaussian noise (σ=1.0). Tab. 7 showed the assumption is consistent in all settings. Further, we tested backdoor-triggered samples with and without Gaussian noise across various imbalance settings. Tab. 8 clearly demonstrate that as the noise level σ increases, the ASR decreases across all imbalance ratios. This confirms that noise-perturbed poisoned samples are indeed less likely to be classified into the target class, providing strong empirical support for our theoretical assumption.

C.5 Performance of Different Noise Samples

In Fig. 10, we evaluate the performance of RPP across different Gaussian noise samples J. The results demonstrate that RPP achieves over 95% detection performance on datasets with imbalance ratios $\rho = 1, 2, 10$, and 100. Even under a high imbalance ratio of $\rho = 200$, RPP maintains a TPR exceeding 85% while keeping the FPR below 40%. Notably, the results indicate that a small number of noise samples (as few as $J = 3$) is sufficient to achieve strong detection performance. To balance detection effectiveness and computational efficiency, we set $J = 3$ in all subsequent experiments.

	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 1.5$
$\rho = 1$	100.0	84.6	77.4	68.3
$\rho = 2$	99.9	83.7	77.7	70.3
$\rho = 10$	99.7	94.8	89.3	83.9
$\rho = 100$	100.0	98.6	94.4	90.9
$\rho = 200$	100.0	99.1	97.5	93.1

Table 8: The ASR of Badnets backdoor attack was evaluated on the balanced CIFAR10 training dataset ($\rho = 1$) and imbalanced CIFAR10 training datasets ($\rho = 2, 10, 100, 200$) with poisoning ratio (p = 0.3%) under different standard deviations of isotropic Gaussian noise ($\sigma = 0, 0.5, 1.0, 1.5$).

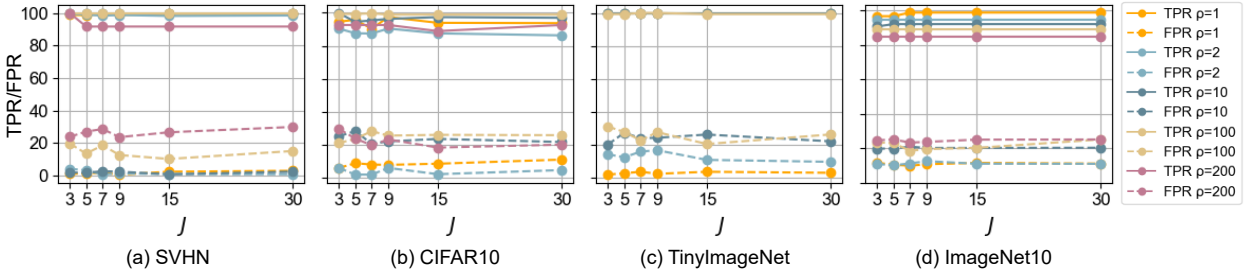

Figure 10: Performance of RPP against Badnets backdoor attack with perturbation magnitude $\|\delta\|_2 \geq 0.8$ across balanced dataset ($\rho = 1$) and varying imbalance ratios ($\mu = 0.9$, $\rho = 2, 10, 100, 200$) under different J with $n = 100$, $\alpha = 0.05$ on SVHN, CIFAR-10, TinyImageNet and ImageNet10 datasets.

C.6 Performance on Different Model Architectures

In Tab. 9, we present the performance of RPP across different model architectures such as VGG16, DenseNet161, and EfficientNetB0 under varying imbalance ratios (ρ). RPP achieves consistently high TPR across all settings, where TPRs often reach or approach 100%. Notably, RPP maintains a strong detection capability even under severe imbalance ($\rho = 100$ or $\rho = 200$), demonstrating its robustness across datasets and architectures. TinyImageNet exhibits a similar trend but tends to yield lower FPRs under balanced conditions and slightly higher TPRs under extreme imbalance compared to CIFAR10.

	CIFAR10						TinyImageNet					
	VGG16		DenseNet161		EfficientNetB0		VGG16		DenseNet161		EfficientNetB0	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
$\rho = 1$	99.9	10.1	94.4	18.3	90.3	13.7	100.0	4.2	100.0	9.6	92.2	7.8
$\rho = 2$	99.9	24.3	92.6	24.6	89.8	12.8	99.9	9.9	86.9	16.7	96.6	13.3
$\rho = 10$	93.5	29.8	87.5	22.0	86.6	25.5	100.0	29.3	91.8	23.7	100.0	31.0
$\rho = 100$	87.1	22.8	86.7	29.7	81.8	23.3	100.0	36.6	82.2	27.8	88.7	28.2
$\rho = 200$	90.0	26.9	88.9	30.6	77.8	23.8	–	–	–	–	–	–

Table 9: Performance of RPP against Badnets backdoor attack with perturbation magnitude $\|\delta\|_2 \geq 0.8$ across balanced dataset ($\rho = 1$) and varying imbalance ratios ($\mu = 0.9$, $\rho = 2, 10, 100, 200$) on CIFAR10 and TinyImageNet datasets under VGG16, DenseNet161, and EfficientNetB0 architectures with $n = 100$, $\alpha = 0.05$, and $\sigma = 1.0$ for CIFAR10 and TinyImageNet.

C.7 Impact of Calibration Set Size

The selection of the calibration set size (n) significantly impacts the performance of conformal prediction. Intuitively, a larger n may seem preferable as it generally results in more stable procedures, a notion supported by (Vovk, 2012). Generally speaking, choosing a calibration set of size $n = 1000$ is sufficient for most purposes (Angelopoulos et al., 2023). However, in practical scenarios, it is challenging to obtain an additional and big clean calibration set ($n = 1000$) that is identically distributed (i.i.d.) with the original data set. In Fig. 11, we present the TPR and FPR of RPP against the Badnets attack across three datasets, evaluated with different calibration set sizes (n=100, 200, 400, 600, 800 and 1000) and $\alpha = 0.05$. Our findings demonstrate that RPP's computational and data efficiency can be enhanced by reducing the size of the calibration set, without significantly compromising its detection or certification performance. Specifically, when the calibration set size is decreased from 1000 to 100, the TPR remains nearly unchanged across datasets with varying degrees of imbalance. This robustness of RPP to the calibration set size further validates the clear separation between benign and backdoored samples, underscoring its effectiveness in practical scenarios.

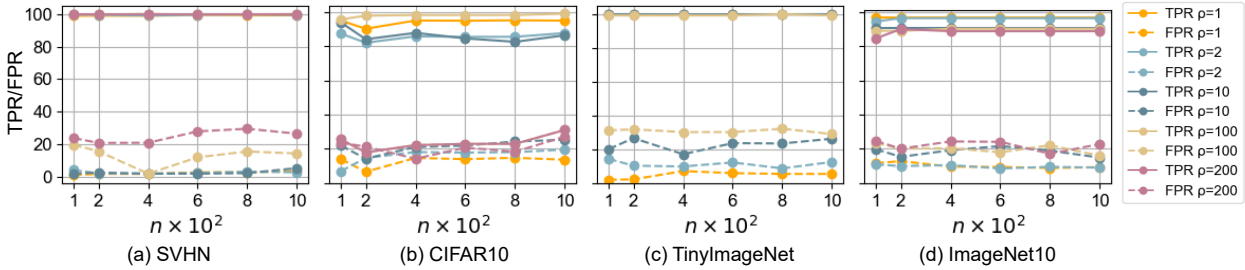

Figure 11: Performance of RPP against Badnets backdoor attacks with perturbation magnitude $\|\delta\|_2 \geq 0.8$, measured by TPR and FPR across balanced dataset ($\rho = 1$), and varying imbalance ratios ($\mu = 0.9$, $\rho = 2, 10, 100, 200$) for a range of n with $\alpha = 0.05$ on SVHN, CIFAR-10, TinyImageNet and ImageNet10 datasets.

	Badnets	Blend	Trojan	SIG	ISSBA	WaNet
TPR	81.1	82.2	74.8	77.0	78.5	87.1
FPR	18.7	17.2	15.6	10.9	24.3	12.8

Table 10: Detection performance (TPR/FPR, %) of RPP against different backdoor attack types with $n = 100$, $\alpha = 0.05$, and $\sigma = 1.0$ on the iNaturalist dataset.

C.8 Performance on MNIST

We use a standard 2-layer convolutional neural network (App. B.3) for MNIST and report TPR/FPR under a balanced setting ($\rho = 1$) and multiple imbalance ratios ($\rho = 2, 10, 100, 200$). Fig. 12 shows as α increases from 0.05 to 0.10, TPR remains near ceiling across all ρ, while FPR rises slightly yet stays modest. Sweeping σ from 0.05 to 0.5 yields TPR close to 100% for most ρ, with only mild degradation at the largest σ under severe imbalance; FPR typically remains below 35%. Increasing n from 100 to 1000 keeps TPR essentially saturated and generally reduces FPR, indicating that larger calibration sets stabilize false-positive control. Finally, varying the number of Gaussian noise samples J maintains TPR around 100% across ρ.

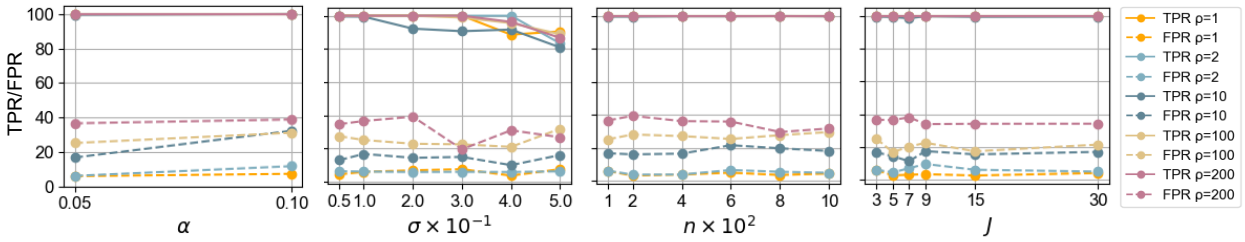

Figure 12: Performance of the RPP detection against Badnets backdoor attacks with perturbation magnitude $\|\delta\|_2 \geq 0.8$, measured by TPR and FPR across balanced dataset ($\rho = 1$) , and varying imbalance ratios ($\mu = 0.9$, $\rho = 2, 10, 100$) on MNIST dataset. (a) For different α values with $\sigma = 0.1$ and $n = 100$. (b) For a range of σ values with $\alpha = 0.05$ and $n = 100$. (c) For various calibration size n with $\alpha = 0.05$ and $\sigma = 0.1$.

C.9 Performance on iNaturalist

To further evaluate the generalization of RPP in realistic long-tailed settings, we conduct experiments on the naturally imbalanced iNaturalist-2018 dataset (Van Horn et al., 2018). Ten classes are randomly selected from the training data while preserving the dataset's inherent long-tailed distribution, forming an iNaturalist-10 subset. We employ a ViT-B/16 model as the classifier and evaluate RPP against multiple backdoor attack types with $n = 100$, $\alpha = 0.05$, and $\sigma = 1.0$. As summarized in Table 10, RPP maintains comparable detection performance on this naturally imbalanced dataset, highlighting its robustness.

C.10 Performance under Imbalanced Calibration Set

In Tab. 11, we present the detection performance of RPP with $\alpha = 0.05$ when the validation set consists of imbalanced clean data that are i.i.d. with the training set. The results suggest that the detection efficacy of RPP is unaffected by whether the calibration set is balanced or imbalanced. We further evaluate RPP under diverse backdoor attacks on CIFAR-10, as shown in Tab. 12. Across different imbalance ratios and attack types, RPP consistently maintains strong detection performance, indicating that its effectiveness is robust to both calibration-set imbalance and attack variation.

	$\rho = 1$		$\rho = 2$		$\rho = 10$		$\rho = 100$		$\rho = 200$	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
MNIST	100.0	5.7	100.0	6.9	82.4	0.5	100.0	1.9	100.0	6.7
SVHN	99.7	11.6	98.8	3.6	100.0	5.9	98.9	2.9	89.7	4.6
CIFAR10	98.0	5.8	93.1	3.0	97.4	7.0	86.1	3.8	87.0	12.4
TinyImageNet	100.0	1.7	100.0	2.9	98.0	2.6	85.7	4.7	–	–
ImageNet10	97.8	7.3	94.5	8.1	90.2	10.9	88.5	12.0	82.6	13.7

Table 11: Performance of the RPP detection against Badnets backdoor attacks with perturbation magnitude $\|\delta\|_2 \geq 0.8$, measured by TPR and FPR across balanced dataset ($\rho = 1$) and varying imbalance ratios ($\mu = 0.9$, $\rho = 2, 10, 100, 200$) on MNIST, SVHN, CIFAR-10, TinyImageNet, and ImageNet10 datasets.

	Badnets		Blend		Trojan		SIG		ISSBA		WaNet		Sleeper Agent		AdaPatch		AdaBlend		Narci.	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
$\rho = 1$	98.0	5.8	98.8	6.1	92.3	9.6	98.5	3.2	94.6	7.7	93.7	6.5	95.0	16.2	92.2	6.0	89.0	10.4	96.1	19.6
$\rho = 2$	93.1	3.0	90.4	7.9	88.4	10.0	95.1	5.6	90.3	9.8	91.0	9.2	94.5	21.7	87.9	4.3	86.2	10.5	96.3	14.8
$\rho = 10$	97.4	7.0	86.0	10.2	96.9	9.0	81.2	4.1	88.5	12.5	87.3	11.6	79.4	19.3	83.6	3.8	83.1	11.0	83.7	20.3
$\rho = 100$	86.1	3.8	80.4	17.1	77.2	3.9	72.9	12.4	56.3	12.5	72.7	23.0	64.1	4.7	66.9	15.2	70.6	14.8	68.5	18.3
$\rho = 200$	87.0	12.4	76.7	20.1	58.8	9.0	51.2	8.3	56.9	9.5	56.8	21.8	70.3	3.6	43.5	6.7	41.6	10.1	57.0	22.8

Table 12: Detection performance of RPP on balanced and imbalanced CIFAR-10 datasets with different imbalance ratios ρ.

C.11 Impact of Preliminary Model Quality

RPP computes its detection score from the output probabilities of a preliminary model. Therefore, RPP is intended for the regime where an effective poisoning attack has been learned by the preliminary model. This is consistent with the data-poisoning backdoor threat model: if the preliminary model yields very low ASR, then the poisoning attack under the given budget has not effectively produced a high-ASR backdoored model.

To study this dependency, we evaluate RPP using early, middle, and final preliminary-model checkpoints trained for 20, 50, and 100 epochs. As shown in Tab. 13, RPP improves as the preliminary model learns the trigger-target association. From early to final checkpoints, AUC/TPR increase from 0.892/0.881 to 0.996/0.985 under $\rho = 1$, and from 0.904/0.868 to 0.992/0.968 under $\rho = 2$.

The preliminary model also needs basic predictive ability to provide meaningful probability-stability signals. Under severe imbalance, such as $\rho = 200$, the early checkpoint gives lower detection performance, while the final checkpoint improves TPR to 0.843 after learning both the clean decision structure and the trigger-target association. RPP remains effective in this setting because clean conformal calibration adapts the detection threshold to the benign RPP-score distribution under different imbalance levels.

C.12 Algorithm for RPP Detection and Certification

In this section, we present Algorithm 1, which details the procedure for identifying backdoored samples using the RPP detection method. The algorithm accepts an input x from $\mathcal{D}(x + \delta)$ to be inspected and a clean calibration dataset, $\mathcal{V}$. By computing the empirical $\tilde{\Delta}P(x|w, \sigma)$, over J noise-injected samples. Then

ρ	Checkpoint	AUC	Minority AUC	ASR	TPR	FPR
1	Early	0.892	0.892	0.829	0.881	0.098
	Middle	0.925	0.925	0.937	0.960	0.063
	Final	0.996	0.996	0.988	0.985	0.045
2	Early	0.904	0.903	0.839	0.868	0.089
	Middle	0.952	0.951	0.958	0.958	0.061
	Final	0.992	0.992	0.999	0.968	0.056
10	Early	0.918	0.907	0.867	0.874	0.082
	Middle	0.963	0.957	0.968	0.959	0.069
	Final	0.963	0.957	0.999	0.962	0.068
100	Early	0.722	0.681	0.830	0.813	0.140
	Middle	0.820	0.805	0.969	0.936	0.132
	Final	0.858	0.823	1.000	0.997	0.137
200	Early	0.630	0.585	0.846	0.722	0.144
	Middle	0.783	0.731	0.942	0.819	0.154
	Final	0.802	0.786	1.000	0.843	0.154

Table 13: Impact of preliminary model quality on RPP detection under different CIFAR-10 imbalance ratios ρ with Badnets attacks. Early, middle, and final denote checkpoints at 20, 50, and 100 epochs, respectively.

compared $\tilde{\Delta}P(x|w,\sigma)$ against a threshold derived from the calibration set $\mathcal{S}$, which consists of RPP values calculated for each instance in $\mathcal{V}$. The threshold is determined by finding the quantile specified by the significance level α and the calibration set size of n. The sample is classified as poisoned if its RPP is less than or equal to this quantile threshold, otherwise, it is deemed clean.

Algorithm 1 Identification of Backdoored Samples

1: **Input:** x: data from $\mathcal{D}(x+\delta)$ to be inspected, δ: backdoor trigger, $\mathcal{V}$: clean calibration data set.
2: **Parameters:** n: size of $\mathcal{V}$, σ: noise standard deviation, J: number of noise injections, α: significance level.
3: **Step 1:** Pretrain a classifier $f(\cdot|w)$ using $\mathcal{D}(x+\delta)$; Compute the empirical RPP $\tilde{\Delta}P(x|w,\sigma)$ over J samples as defined in eq. (2).
4: **Step 2:** Form the calibration set $\mathcal{S} = \{s_1, \ldots, s_n\}$, where $s_i = \tilde{\Delta}P(x_i|w,\sigma)$ for each $x_i \in \mathcal{V}$.
5: **Step 3:** Find the $\frac{\lceil \alpha(n+1) \rceil}{n}$ quantile as

$$\hat{q} = \inf\left\{ q : \frac{|\{i : s_i \leq q\}|}{n} \geq \frac{\lceil \alpha(n+1) \rceil}{n} \right\}$$

6: **Output:** Classify x as poisoned if $\tilde{\Delta}P(x|w,\sigma) \leq \hat{q}$; otherwise, classify x as clean.

In Algorithm 2, we outline a method for certifying the detection of poisoned samples using the RPP approach. Estimating the probability $p(x)$ that the modified data point $x+\delta$ is classified as a target class y_t under the pre-trained classifier $f(\cdot|w)$. We estimate $\zeta(x,\delta)$ using the $\frac{\lceil \alpha(n+1) \rceil}{n}$-quantile of the calibration set $\mathcal{S}$ and let y_{maj} denote the class that appears most frequently among J noise-injected copies of x, where $\epsilon_1, \ldots, \epsilon_J \sim \mathcal{N}(0, \sigma^2 I)$. Specifically, for each $x+\epsilon_j$, we process the sample through the classifier $f(\cdot|w)$, tally the frequency of each class across all J predictions, and identify y_{maj} as the class with the highest occurrence. The count of occurrences for y_{maj} within these J samples, denoted as $n_{\text{maj}} = \text{counts}[y_{\text{maj}}]$, is used to determine the upper confidence bound for this probability, taking into account the noise injections and the significance level α. If the calculated bounds on δ allow for guaranteed detection of the poisoned samples, providing a clear criterion for certifying the presence of backdoors.

C.13 Baselines for Backdoor Sample Detection

We compare our method against 12 empirical backdoor defenses on both balanced ($\rho = 1$) and imbalanced CIFAR-10 datasets ($\rho = 10, 200$), as summarized in Tab. 14. Detection results for RPP are reported under $\alpha = 0.05$, $n = 100$, and $\sigma = 1.0$, across 10 types of backdoor attacks. In all cases, we ensure ASR exceeds 90%. RPP consistently achieves comparable or superior TPRs relative to SOTA defenses across all trigger types. We also briefly review these baseline methods and analyze why they fall short in challenging imbalanced settings.

SS(Tran et al., 2018) uses singular value decomposition to detect anomalies by selecting a subset of data comprising 1.5 times the number of samples with the highest anomaly scores. However, the anomaly detection strategies struggle when the number of poisoned samples is extremely low and the dataset is imbalanced. AC (Chen et al., 2018) captures activation data from the terminal hidden layer of a trained neural network and applies clustering algorithms to pinpoint and expunge anomalous instances from the training dataset. However, the method's efficacy is limited due to the uniform clustering effect and the substantial ratio disparity between poisoned and clean samples, which hinders its performance. ABL(Li et al., 2021a) was initially developed as a robust training defense and subsequently repurposed as a detection method based on output losses. This approach exhibits limited efficacy due to the significantly lower isolation rate of poisoned samples compared to the actual poisoning rate. With only a minimal number of poisoned samples isolated, often as few as two or three, the effectiveness of this method is substantially compromised. CT (Qi et al., 2023b) utilized confusion training for detecting backdoored samples. When the dataset exhibits extreme imbalance and the target label corresponds to the majority class, clean samples are predominantly predicted as belonging to this majority class, resulting in low training losses for these clean samples. ASSET (Pan et al., 2023): In the imbalanced datasets and low poison ratio settings, when training with mini-batches, the number of poisoned samples in each mini-batch is very small, or even nonexistent. In the inner offset loop, different mini-batches may contain different amounts of poisons and even no poisoned samples, so it is hard to determine the number of most suspicious samples within each mini-batch; Additionally, when using the gradient ascent algorithm, the optimization process tends to overlook the characteristics of these few samples. This means that the model may not adequately learn the features of these poisoned samples during the learning process. SCALE-UP (Guo et al., 2023): When the dataset is imbalanced, although the model still exhibits some consistency in classifying backdoor samples, the bias towards the majority class also results in consistent classification predictions for clean samples, which leads to a high rate of misdiagnosis. MSPC (Pal et al., 2024) employs a bi-level optimization technique designed to minimize the cross-entropy loss for clean samples while simultaneously maximizing it for backdoor samples, facilitating the identification of backdoor samples. However, when the dataset is highly imbalanced and the poisoning rate is low, the use of upper-level optimization to increase the loss for backdoor samples yields suboptimal results. STRIP (Gao et al., 2019) relies on entropy-based thresholds, which become unreliable when minority-class benign samples naturally exhibit low entropy, leading to higher false positives. TED (Mo et al., 2024) assumes stable topological evolution for benign samples, but class imbalance distorts layer-wise activation patterns, making poisoned and minority-class samples harder to distinguish. BBCaL (Hu et al., 2024) uses random perturbations to probe behavior changes, but its fixed calibration thresholds are not robust across imbalanced distributions, reducing detection sensitivity. IBD-PSC (Hou et al., 2024) detects backdoor presence based on maximum margin statistics, but this method becomes unstable under severe imbalance due to a similar overfitting phenomenon and biased logit distributions across classes. RE (Xiang et al., 2021) introduces an optimization-based defense that detects backdoors, identifies the target class, and removes poisoned samples by estimating minimal source-to-target perturbations. However, its performance degrades under data imbalance, since minority classes yield unstable perturbation estimates and biased anomaly statistics, making the inferred backdoor signal less distinguishable.

C.14 Certified Defenses under Imbalanced Dataset

RAB (Weber et al., 2023), the inaugural robust training methodology, certifies its robustness against backdoor attacks via randomized smoothing. DPA (Levine & Feizi, 2020) offers a certifiable defense against general and label-flipping poisoning attacks by partitioning the training set into k segments. These methodologies provide either probabilistic or deterministic assurances that a model will produce the intended outputs under

Algorithm 2 RPP Certification

1: **Input:** x: data point to be inspected, $\mathcal{V}$: clean calibration data set, $f(\cdot|w)$: pre-trained classifier, δ: backdoor trigger, Φ: a standard Gaussian CDF.
2: **Parameters:** n: size of $\mathcal{V}$, σ: noise standard deviation, J: number of noise injections, α: significance level.
3: $p(x) \leftarrow \mathbb{P}(f(x+\delta|w) = y_t)$
4: Get $\zeta(x,\delta) \leftarrow \frac{\lceil \alpha(n+1) \rceil}{n}$ quantile from the calibration set $\mathcal{S}$
5: $y_{\text{maj}} \leftarrow$ top class in $f(x+\epsilon_j|w)$
6: $n_{\text{maj}} \leftarrow \text{counts}[y_{\text{maj}}]$
7: $\overline{p_t} \leftarrow \textsc{UpperConfBound}(n_{\text{maj}}, J, 1-\alpha)$
8: **Output:** If $\sigma\left(\Phi^{-1}(p(x) - \zeta(x,\delta)) - \Phi^{-1}(\overline{p_t})\right) \leq \|\delta\|_2 \leq \sigma\left(\Phi^{-1}(p(x)) - \Phi^{-1}(\overline{p_t})\right)$, the poisoned samples are guaranteed to be detected; otherwise, the poisoned samples may be detected but without guarantee.

		Badnets		Blend		Trojan		SIG		ISSBA		WaNet		Sleeper Agent		AdaPatch		AdaBlend		Narci.	
	Defenses	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
	SS	93.1	5.6	91.8	4.9	60.2	10.7	61.0	20.1	45.5	4.2	59.6	8.8	61.3	5.1	85.9	6.7	83.6	8.2	12.1	8.5
	AC	97.8	2.8	96.4	5.7	73.6	16.0	78.9	10.4	51.2	7.0	23.1	10.6	45.6	17.9	91.7	6.6	88.0	9.3	9.3	26.7
	RE	92.1	10.3	90.8	8.9	88.2	12.4	64.2	9.5	80.1	11.9	78.5	15.0	66.4	9.0	82.3	10.4	80.5	7.5	14.3	5.8
	ABL	86.4	3.9	90.2	10.3	85.0	4.8	85.1	8.3	52.8	3.5	78.6	8.2	20.3	1.7	55.8	3.6	50.2	3.9	1.8	2.8
	CT	93.9	8.7	95.2	6.3	88.9	5.0	92.0	19.6	95.2	11.0	92.6	9.5	82.1	11.2	84.0	4.3	86.7	6.5	22.1	17.5
	ASSET	89.9	17.8	83.1	16.0	90.4	18.3	86.7	20.0	91.2	8.8	93.0	11.7	52.9	15.4	55.7	21.3	80.3	10.6	92.1	4.3
$\rho = 1$	SCALE-UP	100.0	19.3	77.1	20.1	67.4	18.8	**98.8**	32.3	92.5	21.2	84.2	19.9	45.1	8.5	80.7	30.0	79.9	22.4	67.3	19.5
	MSPC	92.2	22.7	94.8	12.4	66.7	16.0	91.4	22.3	89.5	15.5	94.9	8.4	79.1	21.2	80.3	29.6	84.4	18.5	82.4	23.6
	BBCaL	98.9	10.1	96.8	12.4	80.3	9.5	98.3	13.1	**99.7**	9.6	90.2	14.2	91.0	16.6	75.1	15.7	72.7	13.6	86.6	9.4
	STRIP	98.5	10.3	84.4	15.6	33.1	11.0	94.3	22.1	48.6	19.7	13.3	29.9	22.6	16.1	88.8	19.6	87.6	20.5	0.0	0.8
	TED	**100.0**	2.9	94.3	6.4	80.6	9.3	91.1	13.6	93.6	12.5	92.6	8.8	90.3	11.4	88.4	10.3	86.2	13.5	91.7	10.5
	IBD-PSC	99.4	4.5	97.9	5.1	**93.0**	8.6	87.2	10.4	98.6	8.7	**95.6**	7.3	69.0	9.5	92.0	10.1	86.7	9.6	93.1	24.7
	Ours	98.5	5.9	**98.8**	6.1	92.3	9.6	98.5	3.2	94.6	7.7	93.7	6.5	**95.0**	16.2	**92.2**	6.0	**89.0**	10.4	**96.1**	19.6
	SS	35.2	7.3	37.7	8.2	35.3	4.1	60.7	0.2	15.1	10.0	38.7	6.6	40.9	31.2	22.5	3.7	13.3	6.6	0.0	0.6
	AC	50.8	3.8	34.1	4.8	17.9	3.0	60.0	36.1	35.5	16.8	50.2	12.4	20.1	16.2	42.5	43.6	20.8	4.9	0.0	0.0
	RE	41.6	18.8	40.3	15.5	43.7	13.0	29.6	15.4	39.6	17.4	35.0	12.6	25.9	11.4	40.7	14.5	37.6	13.0	13.9	17.6
	ABL	0.0	0.2	0.0	0.6	0.0	0.1	0.0	0.0	9.2	2.7	47.5	7.3	0.0	2.6	0.6	0.2	0.0	0.3	0.0	0.8
	CT	68.9	0.9	60.4	2.9	60.4	8.9	59.8	0.1	75.9	8.9	80.6	13.9	26.1	0.0	0.0	0.0	32.5	22.4	0.0	8.9
	ASSET	60.0	6.9	51.2	18.4	39.3	41.5	10.6	19.7	78.6	18.8	69.5	14.1	71.4	15.0	38.3	31.1	40.6	31.3	51.7	13.9
$\rho = 10$	SCALE-UP	95.9	84.3	67.3	35.1	67.2	41.5	71.1	34.0	86.2	22.6	78.0	19.6	47.3	21.5	68.3	32.5	62.9	37.7	39.0	12.0
	MSPC	89.6	16.5	**89.5**	11.3	55.7	21.6	**83.9**	10.0	80.1	20.3	76.9	9.4	63.7	0.9	51.2	24.3	41.0	9.5	49.6	29.4
	BBCaL	83.2	17.9	88.7	22.0	67.3	19.8	79.7	14.5	78.6	21.3	72.2	20.7	69.4	25.3	60.1	22.4	59.9	25.6	66.4	15.0
	STRIP	63.4	19.7	50.0	26.8	18.3	17.7	60.6	28.5	28.4	21.6	6.5	28.9	26.8	16.3	64.8	26.8	44.3	28.8	0.0	0.5
	TED	90.4	12.2	82.5	14.3	54.3	17.6	72.0	16.9	79.9	21.0	76.5	16.4	70.1	19.4	66.8	21.1	61.2	20.3	76.9	18.6
	IBD-PSC	91.2	14.1	87.4	12.7	83.6	17.5	70.3	19.6	85.9	19.4	77.4	20.2	61.4	20.4	74.8	16.6	75.5	18.9	80.3	28.6
	Ours	**96.7**	3.2	85.2	10.4	**96.4**	9.1	80.0	4.2	**88.1**	13.6	**87.5**	11.7	**78.7**	19.5	**82.4**	3.3	**82.6**	12.5	**83.7**	20.1
	SS	0.1	0.0	4.9	0.7	14.7	4.9	0.0	5.0	4.4	8.1	7.1	6.0	0.0	0.1	7.5	1.4	9.5	3.6	0.0	0.3
	AC	0.0	0.0	0.0	0.5	0.0	1.0	0.0	9.4	0.6	2.4	0.0	5.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RE	16.4	7.0	12.8	7.4	8.6	7.8	10.6	0.0	6.9	12.2	10.5	11.5	0.0	10.3	0.0	0.7	2.1	1.9	0.0	6.2
	ABL	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.5	0.0	0.3	0.0	0.3	0.0	0.0	0.0	2.9
	CT	0.0	3.2	0.0	3.6	0.0	0.0	0.0	0.0	0.0	1.6	0.0	6.7	0.0	1.6	0.0	0.0	0.0	0.0	0.0	15.7
	ASSET	0.0	3.3	7.3	2.8	0.0	0.0	0.0	0.0	9.3	2.2	14.8	5.5	6.7	1.8	3.2	0.4	5.5	3.1	2.0	1.3
$\rho = 200$	SCALE-UP	67.3	54.1	51.4	48.3	0.0	0.0	0.0	0.0	33.8	19.9	48.6	30.1	0.2	42.1	6.3	0.6	29.7	30.1	10.1	21.8
	MSPC	10.6	2.8	20.2	6.1	15.3	3.1	19.2	2.9	18.8	5.6	32.6	17.3	10.6	5.8	8.2	1.7	9.4	3.2	21.7	32.8
	BBCaL	41.5	33.6	42.2	35.1	30.6	26.7	37.0	35.8	30.1	31.9	29.3	36.2	49.2	27.7	22.8	28.6	23.0	29.1	39.3	18.9
	STRIP	18.6	20.1	20.4	31.4	0.0	12.7	22.6	33.8	9.5	21.7	0.3	8.4	1.2	10.5	21.0	30.6	10.3	21.8	0.0	1.2
	TED	46.7	12.6	44.9	19.7	12.7	18.3	50.4	22.1	46.8	14.5	29.7	22.7	29.7	22.7	30.2	22.6	27.5	28.9	50.2	20.4
	IBD-PSC	51.7	17.8	50.0	24.7	55.8	26.3	50.0	24.7	41.9	23.8	40.5	22.0	30.4	24.1	32.7	22.8	34.0	25.5	44.8	30.0
	Ours	**83.3**	14.3	**75.5**	19.8	**58.2**	9.1	**50.5**	9.1	**57.1**	10.6	**56.8**	23.3	**69.7**	3.8	**42.8**	6.2	**41.5**	10.2	**56.7**	23.2

Table 14: Comparison with SoTA defenses on balanced ($\rho = 1$) and imbalanced CIFAR-10 datasets ($\mu = 0.9$, $\rho = 10, 200$).

specific adversarial conditions or when subjected to a certain number of label flips. As shown in Tab. 15, the efficacy of these certified defenses tends to decline as the dataset becomes more imbalanced. For RAB, we configure $\|\delta\|_2 = 0.1$ with a smoothing parameter $\sigma = 1.0$ and a 2% poisoning ratio against Badnets attacks. For DPA, we established 50 partitions with 9 sample labels flipped.

	$\rho = 1$	$\rho = 2$	$\rho = 10$	$\rho = 100$	$\rho = 200$
RAB	41.5	40.2	38.8	21.0	15.7
DPA	56.2	46.1	26.6	22.7	14.8

Table 15: Certified accuracy (%) of RAB and DPA against Badnets attack on the balanced CIFAR-10 training dataset ($\rho = 1$) and imbalanced CIFAR-10 training datasets ($\mu = 0.9$, $\rho = 2, 10, 100, 200$).

C.15 Impact of Target Label Class

We evaluated the performance of our method when the backdoor target label was assigned to either a minority class (class 9) or an intermediate-frequency class (class 4). In all experiments, we ensured that the ASR remained above 90% to maintain a consistent attack strength. As shown in Tab. 16, the proposed RPP method consistently achieved high TPR across various imbalance ratios ρ, while maintaining low FPR.

	Minority class (9)		Intermediate class (4)	
	TPR	FPR	TPR	FPR
$\rho = 2$	99.4	7.3	93.9	3.1
$\rho = 10$	93.4	10.1	96.2	3.3
$\rho = 100$	99.9	2.9	99.3	1.5
$\rho = 200$	90.0	6.0	79.6	4.6

Table 16: Performance of RPP against the BadNets backdoor attack on imbalanced CIFAR-10 training datasets ($\rho = 2, 10, 100, 200$) with a 0.3% poisoning ratio, where the target class is either a minority class (class 9) or an intermediate-frequency class (class 4).

C.16 Performance under Logits-Adjusted Training

To address the performance degradation caused by class imbalance, we adopted logits adjustment (LA) (Menon et al., 2020) as a long-tailed training strategy. This technique explicitly calibrates the logits of each class based on its frequency, thereby mitigating the dominance of majority classes and enhancing the learning of minority-class features. While it slightly reduces ASR under severe imbalance, our RPP method still achieves high TPR (83–99%) across all settings, as shown in Tab. 17. The result demonstrates the robustness of our method under long-tailed training, and its ability to remain effective even when the backdoor attack becomes slightly weakened by class rebalancing techniques.

	ASR (Without LA)	ASR (With LA)	TPR	FPR
$\rho = 2$	100.0	99.9	99.2	22.2
$\rho = 10$	99.8	99.4	96.2	20.3
$\rho = 100$	100.0	97.4	97.8	26.1
$\rho = 200$	99.7	94.3	83.0	10.4

Table 17: Performance of RPP against BadNets attack under logits adjustment on imbalanced CIFAR-10 datasets ($\rho = 2, 10, 100, 200$) with a poisoning ratio of 0.3.

C.17 Performance against Different Trigger Sizes

In Tab. 18, we present the ASR and AUC metrics for various trigger sizes, and we evaluate the efficacy of RPP when the ASR exceeds 50%. Our findings indicate that for $\|\delta\|_2 = 3$, in imbalanced datasets with $\rho = 100, 200$, the ASR surpasses 50%, and our RPP achieves a TPR of 80% with an FPR of less than 30%. When $\|\delta\|_2 = 6$, for both the balanced dataset ($\rho = 1$) and imbalanced datasets ($\rho = 2, 10$), the ASR is above 50%, and RPP attains TPRs of 70.6%, 89.9%, and 87.4%, respectively, with an FPR below 15%. These results align well with our certification.

C.18 Validation of the FPR Upper Bound Guarantee

We validate the probabilistic upper bound for the FPR in Thm. 5.4 by conducting experiments on the CIFAR-10 dataset. Specifically, we first set aside $1,000$ clean images from CIFAR-10 as a validation pool. From this pool, we randomly sample $n \in \{100, 200, 400, 600\}$ images as the calibration dataset and the remaining images are used to evaluate the FPR. We consider two significance level $\alpha = 0.05$ and $\alpha = 0.1$. For

	$\|\delta\|_2 = 0.8$		$\|\delta\|_2 = 2$		$\|\delta\|_2 = 3$				$\|\delta\|_2 = 4$				$\|\delta\|_2 = 6$				$\|\delta\|_2 = 8$			
	No Defense		No Defense		No Defense		RPP		No Defense		RPP		No Defense		RPP		No Defense		RPP	
	ASR	AUC	ASR	AUC	ASR	AUC	TPR	FPR	ASR	AUC	TPR	FPR	ASR	AUC	TPR	FPR	ASR	AUC	TPR	FPR
$\rho = 1$	14.3	100.0	33.7	100.0	35.6	100.0	–	–	39.2	100.0	–	–	52.3	100.0	73.1	4.1	90.5	100.0	97.2	8.0
$\rho = 2$	14.8	100.0	37.8	100.0	33.5	100.0	–	–	47.9	100.0	–	–	69.6	100.0	88.3	6.9	90.3	100.0	96.0	6.1
$\rho = 10$	15.4	100.0	30.5	100.0	38.3	100.0	–	–	49.4	100.0	–	–	84.8	100.0	90.1	9.6	99.8	100.0	97.9	7.7
$\rho = 100$	23.0	99.3	39.6	99.4	52.1	98.6	73.8	10.1	100.0	99.4	87.5	10.7	100.0	99.6	95.2	10.1	100.0	99.5	100.0	10.9
$\rho = 200$	36.0	92.7	49.2	92.1	68.0	96.2	89.5	14.8	99.5	97.5	96.2	15.3	100.0	97.1	100.0	20.1	100.0	97.8	100.0	16.4

Table 18: Performance of RPP under various trigger sizes with a 0.3% poisoning ratio on the balanced MNIST ($\rho = 1$) and imbalanced MNIST datasets ($\mu = 0.9$, $\rho = 2, 10, 100, 200$) for $n = 100$ and $\alpha = 0.05$.

each (n, α) pair, we repeat the experiments 300 times and report the 0.95-quantile FPR in Tab. 19. The results show that the FPR in the validation set is bounded by the theoretically derived upper bound.

	$\alpha = 0.05$				$\alpha = 0.1$			
	$n = 100$	$n = 200$	$n = 400$	$n = 600$	$n = 100$	$n = 200$	$n = 400$	$n = 600$
$\rho = 1$	4.0	5.3	5.1	4.4	10.7	9.9	9.1	8.8
$\rho = 2$	4.9	4.1	4.6	4.0	9.3	9.9	10.1	9.0
$\rho = 10$	5.5	4.9	4.6	4.2	10.7	9.8	9.5	9.1
$\rho = 100$	5.5	4.7	4.8	5.0	9.9	9.2	10.2	10.0
$\rho = 200$	5.1	4.9	4.8	4.7	10.4	10.0	9.8	10.1
Upper bound	**6.0**	**5.5**	**5.3**	**5.2**	**11.0**	**10.5**	**10.3**	**10.2**

Table 19: The FPR (%) under varying imbalance ratio (ρ), isotropic Gaussian noise standard deviation (α), and calibration set size (n) on CIFAR10 dataset.

D Differences Between RPP and CBD

We delineate the key differences between RPP and CBD (Xiang et al., 2023) from both theoretical and methodological perspectives.

Theoretical Perspective. Although both RPP and CBD leverage randomized smoothing and draw on the Neyman–Pearson lemma as formal justification (as inspired by (Cohen et al., 2019)), the objects being certified and the proof strategies differ substantially:

- Theorem 4.1 in CBD finds the lower bound of $s(w)$ (i.e. ℓ_∞ norm of Local Dominant Probability), while we bound below the RPP (expectation of the difference probability vectors). These certified quantities reflect fundamentally different detection criteria.
- CBD applies part (1) of Lemma 4 from (Cohen et al., 2019)) to derive an *upper bound* on the response to backdoor triggers. RPP, by contrast, uitilize part (2) of Lemma 4 from (Cohen et al., 2019)) to construct a *lower bound* for reliably detecting poisoned samples. These choices lead to distinct forms of certification, with differing robustness implications.

Methodological Setting (Threat Model). CBD is designed for **post-training model inspection**: it aims to determine whether a given *model* has been compromised by a backdoor, without requiring access to the training data. In contrast, RPP addresses a more proactive threat model—**pre-training poisoned sample identification**. It operates directly on the training data, identifying and filtering out backdoor instances *prior* to any model training, enabling safe downstream learning even under stealthy poisoning attacks.

E Computational Complexity

The preliminary model is trained only once for the inspected dataset. On a single NVIDIA A100-SXM4-80GB device, the one-time training cost varies with the dataset size under different imbalance ratios. For $\rho = \{1, 2, 10, 100, 200\}$, the training time is 1.27 min, 55.38 s, 31.23 s, 18.97 s, and 17.13 s on MNIST with a two-layer CNN; 8.85, 6.41, 3.62, 2.20, and 1.98 min on SVHN with PreActResNet18; 9.64, 6.98, 3.94, 2.39, and 2.16 min on CIFAR-10 with PreActResNet18; 3.65, 2.64, 1.43 h, 47.46 min, and 41.53 min on TinyImageNet with ResNet34; and 2.37, 1.72 h, 58.07 min, 35.26 min, and 31.84 min on ImageNet10 with ViT-B/16, respectively.

On a single NVIDIA A100-SXM4-80GB device, the detection time per sample is 0.006 s for MNIST, 0.0154 s for SVHN, 0.0162 s for CIFAR-10, 0.0261 s for TinyImageNet, and 0.03571 s for ImageNet10. On an NVIDIA GeForce RTX 4090 device, the corresponding detection times are 0.021, 0.0272, 0.0281, 0.0404 and 0.0521 s, respectively.

F Limitations & Future Work

Our theoretical analysis relies on the assumption of i.i.d. Gaussian perturbations. Extending this analysis to encompass non-Gaussian or correlated perturbations could offer a more comprehensive understanding of robustness under realistic conditions and serves as a promising direction for future work.

G Use of Large Language Models (LLMs)

We used LLMs for editorial assistance (clarity, grammar, wording, and minor reorganizations). LLMs were not used to generate ideas, design experiments, or manipulate results, or draft related-work claims. The authors take full responsibility for all content.