PAC-Wrap: Semi-Supervised PAC Anomaly Detection

Shuo Li*
University of Pennsylvania
Philadelphia, PA, US
lishuo1@seas.upenn.edu

Xiayan Ji*
University of Pennsylvania
Philadelphia, PA, US
xjiae@seas.upenn.edu

Edgar Dobriban
University of Pennsylvania
Philadelphia, PA, US
dobriban@wharton.upenn.edu

Oleg Sokolsky
University of Pennsylvania
Philadelphia, PA, US
sokolsky@cis.upenn.edu

Insup Lee
University of Pennsylvania
Philadelphia, PA, US
lee@cis.upenn.edu

ABSTRACT

Anomaly detection is essential for preventing hazardous outcomes for safety-critical applications like autonomous driving. Given their safety-criticality, these applications benefit from provable bounds on various errors in anomaly detection. To achieve this goal in the semi-supervised setting, we propose to provide Probably Approximately Correct (PAC) guarantees on the false negative and false positive detection rates for anomaly detection algorithms. Our method (PAC-Wrap) can wrap around virtually any existing semi-supervised and unsupervised anomaly detection method, endowing it with rigorous guarantees. Our experiments with various anomaly detectors and datasets indicate that PAC-Wrap is broadly effective.

CCS CONCEPTS

• **Security and privacy** → **Intrusion/anomaly detection and malware mitigation**; • **Theory of computation** → *Sample complexity and generalization bounds*; • **Computing methodologies** → *Semi-supervised learning settings.*

KEYWORDS

Anomaly Detection, Semi-Supervised Learning, Statistical Machine Learning, PAC Learning

ACM Reference Format:
Shuo Li, Xiayan Ji, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. 2022. PAC-Wrap: Semi-Supervised PAC Anomaly Detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539408

1 INTRODUCTION

Anomaly detection aims to detect points that significantly deviate from the regular pattern of data and may threaten system safety.

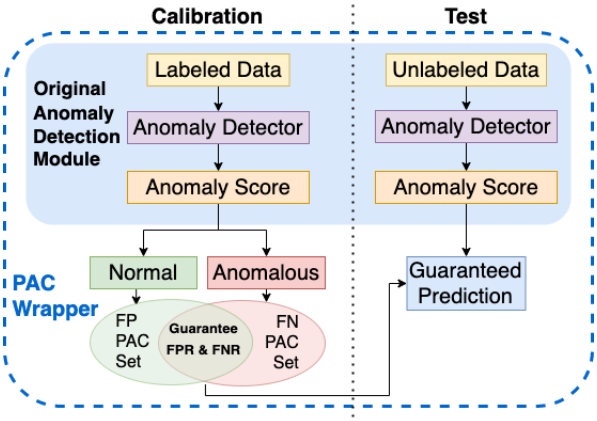


Figure 1: An overview of PAC-Wrap, which wraps around an arbitrary anomaly detector. In the calibration phase, we derive false positive (FP) PAC sets and false negative (FN) PAC sets, which guarantee the false positive (FPR) and negative rates (FNR) respectively. We take an intersection to have both guarantees. After eliminating ambiguity, it is later used at the test phase to detect anomalies with a PAC guarantee.

In recent years, anomaly detectors based on machine learning algorithms have started to outperform classical methods in many tasks [5, 21, 37]. Some of these tasks are safety-critical and require rigorous guarantees on the false negative and false positive rates. However, machine learning-based anomaly detectors usually do not guarantee these rates by default.

Some methods propose using standard conformal prediction [1, 34], an uncertainty quantification technique, for rigorous guarantees. These methods are effective when sufficient data is given, i.e., the dataset is large enough to represent the whole data distribution. Nevertheless, we cannot make this assumption in practical settings, and hence we shall allow for some error margin incurred by the data insufficiency. An alternative approach is to use training-set conditional methods, such as inductive conformal prediction [23], which satisfy a Probably Approximately Correct (PAC) property [26, 30, 32]. As we will argue, this property offers more flexibility than the marginal guarantees for conformal prediction. Furthermore, most anomaly detection methods with rigorous guarantees only control the false positive rate (FPR). The lack of false negative rate (FNR) guarantees could limit the usefulness of a system since classifying anomalies as normal can be a consequential mistake.

*The first two authors, Shuo Li and Xiayan Ji, contributed equally to this paper.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539408

Hence, we propose an algorithm, named *PAC-Wrap*, to add a layer ensuring a PAC guarantee on FPR and FNR to virtually any anomaly detector. In other words, PAC-Wrap acts like a wrapper that helps an anomaly detector attain a rigorous performance guarantee while keeping its internal structure intact. PAC-Wrap takes a user-specified *error* level, denoted as ε, and a user-specified *confidence* level, denoted as δ, to customize the guarantee. We perform this in semi-supervised anomaly detection, where a small amount of labeled data is available [20, 27, 31]. Our algorithm leverages the limited labeled data and provides training-set conditional guarantees, which we argue are more practical than the marginal guarantees provided by standard conformal prediction-based methods. Since we leverage both labeled normal and anomalous data, we can provide PAC guarantees not only on the FPR but also the FNR.

Given any trained anomaly detector that outputs the anomaly score, our wrapper method constructs false positive and false negative PAC prediction sets on the calibration datasets. These two PAC prediction sets provide PAC guarantees on FPR and FNR. Then, we propose to take the intersection of the PAC prediction sets and adopt the classification with rejection option idea [2, 11]. The resulting anomaly detector guarantees FPR and FNR if it is confident about its prediction. On the other hand, if the anomaly score falls into the ambiguity region where it is not sufficiently confident about its prediction, it abstains from predicting. In cases where the rejection option is not allowed, we further propose an algorithm to eliminate the ambiguity regions. Finally, we prove that the prediction sets and the final anomaly detector are probably approximately correct. An overview of our method is in Figure 1.

We conduct experiments to validate the correctness of our theorems on both synthetic and benchmark datasets. Moreover, we also demonstrate that our wrapper can ensure that the performance of the underlying anomaly detector is rigorously guaranteed at a user-specified error and confidence level. Furthermore, to demonstrate the generalizability benefit of the PAC-based guarantee, we compare the performance of our PAC anomaly detection method to standard conformal prediction-based methods [18]. Finally, we explore the relationship between the error level, the confidence level, and the ambiguity region.

In summary, our contributions are as follows:

- We propose to wrap PAC prediction sets around general anomaly detectors. We show rigorous guarantees on the FNR and FPR in semi-supervised anomaly detection.
- We show that the training-set conditional PAC guarantee has both practical and theoretical benefits in generalization and flexibility compared to marginal guarantees provided by the standard conformal prediction.
- We demonstrate empirically in simulations and on challenging benchmark datasets, using a variety of state-of-the-art anomaly detectors, that PAC-Wrap is effective.
- We conduct an ablation study to evaluate the tradeoff between the error level, the confidence level, and the ambiguity region.

2 RELATED WORK

Conformal prediction (CP), also referred to as conformal inference [33], is a general approach to uncertainty quantification. It can provide finite dataset coverage guarantees under exchangeability of the data points and has been widely adopted, e.g., [1, 25], etc. Closely related to our work, [18] also provides guarantees on false negative and false positive rates for classification. However, [18] is based on standard conformal prediction and provides a coverage guarantee that holds marginally over the training set. Similarly, the work [9] proposes a class-wise thresholding scheme for OOD detection algorithms to maintain a comparable true positive rate across classes. Mondrian conformal prediction is a general approach to provide guarantees conditional on a general data clustering (of which class-conditional guarantees are a special case). However, their guarantees studied so far hold marginally over the training set [34]. The guarantees of conformal prediction, which hold marginally over the training set [9, 18, 34] mean that the method works for most collections of training data and one test data point. It implies that the coverage holds for only one test data point. In contrast, the PAC guarantee we use implies that the coverage holds for most future test data points; this is more aligned with the practice setting in which a prediction method is used for many test data points.

Moving beyond standard conformal inference, [6] proposes a block permutation method to account for temporal dependence. EnbPI [36] proposes distribution-free prediction intervals for dynamic time series, extending CP to assume that only the residuals of a fitted model are exchangeable instead of the complete data. Our method differs, as we build upon the PAC framework, which—as discussed above—provides different guarantees. Also, our time series examples are different, as in some cases, our data points are independent time series. Thus the guarantees apply directly, similarly to previous examples such as [17]. In other cases, we take sufficiently separated subsequences of the time series that we expect them to be nearly independent, which holds for certain types of mixing conditions, as in [6].

Inductive Conformal Prediction (ICP) [24] was originally shown to have marginal guarantees, but was later shown to satisfy training-conditional, or PAC guarantees [26, 32]. As discussed in [32], the mathematical structure of these methods is closely related to that of tolerance regions [16, 35]. Inductive conformal anomaly detection [14, 24] builds on ICP to guarantee a bounded false detection rate. In different literature, there are different terminology for the two user-specified inputs. For example, the β − content in [10] is equivalent to $1-\varepsilon$, where ε is the *error parameter* in [15]. The *confidence level* γ in [10] is equivalent to $1-\delta$, where δ is called *confidence parameter* in [15]. In our work, we follow the terminology in [15] and denote ε as the *error parameter*, and δ as the *confidence parameter*. We adopt the core ideas behind this general line of work and focus on adapting it to semi-supervised anomaly detection, where both false positive and false negative rates control are essential.

We focus on semi-supervised anomaly detection (SSAD) tasks, which have been defined in slightly different ways. In these definitions, given a dataset S, we have m unlabeled data points and n labeled data points, where $m \gg n$. SSAD definitions differ in the setup of the training and testing sets. For example, some papers [22, 27] assume that the training set has labeled normal and anomalous data points. This setting is also called *weakly supervised anomaly detection*. On the other hand, some papers [13, 28] assume that the training set only contains normal data points and the test set contains both normal and anomalous data points. We adopt the first definition of SSAD. In both our problem formulation

and experimental evaluation, we assume that the training set has labeled normal and anomalous data points. Note that all the aforementioned semi-supervised algorithms are orthogonal to our work in that we emphasize providing a theoretical guarantee on false negative and false positive rates, whereas they focus on detector performance, like accuracy or F1-score.

3 PRELIMINARIES

3.1 PAC Prediction Sets

For independent and identically distributed (i.i.d.) training and test data, training-set conditionally valid (or, PAC) prediction sets [26, 32] are guaranteed to contain the true labels for test inputs with low error level and high confidence level. While the algorithms in [32] and [26] are identical, we follow the latter. To ensure the prediction sets are small, [26] solves an optimization problem to calculate the smallest prediction set (on average) while satisfying the PAC property.

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the finite label space; let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$; let $C : \mathcal{X} \to 2^{\mathcal{Y}}$ denote a prediction set. The probability that C does not cover a test data point $(x, y) \sim \mathcal{D}$ is defined as

$$L_{\mathcal{D}}(C) \coloneqq \mathbb{P}_{(x,y)\sim\mathcal{D}}[y \notin C(x)]. \tag{1}$$

Let $Z \sim \mathcal{D}^n$ be a held-out calibration set of i.i.d. data points from $\mathcal{D}$ with size n, which we can use to tune or calibrate C, as described below. The goal is to find a set of a small size satisfying the PAC property, *i.e.,* given $\varepsilon, \delta \in (0, 1)$,

$$\mathbb{P}_{Z\sim\mathcal{D}^n}[L_{\mathcal{D}}(C) \le \varepsilon] \ge 1 - \delta,$$

where the $\mathbb{P}_{Z\sim\mathcal{D}^n}$ refers to the chances of calibration succeeding. In this case, we say C is (ε, δ)-correct. To calculate such (ε, δ)-correct sets, [26] then proposes the following one-dimensional parametrization of prediction sets:

$$C_\tau(x) = \{y \in \mathcal{Y} \mid f(x, y) \ge \tau\},$$

where $\tau \in \mathbb{R}_{\ge 0}$ and $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\ge 0}$ is any given scoring function (e.g., the label probabilities output by a deep neural network). The parameter value τ is identified by solving the following optimization problem:

$$\hat{\tau} = \underset{\tau\in\mathbb{R}_{\ge 0}}{\arg\max}\ \tau \text{ subj. to } \sum_{(x,y)\in Z} \mathbb{1}\left(y \notin C_\tau(x)\right) \le k^*, \tag{2}$$

where

$$k^* = \underset{k\in\mathbb{N}\cup\{0\}}{\arg\max}\ k \qquad \text{subj. to} \qquad F(k; n, \varepsilon) \le \delta,$$

where $F(k; n, \varepsilon)$ is the cumulative distribution function of the binomial random variable Binomial(n, ε) with n trials and success probability ε. Maximizing τ corresponds to minimizing the prediction set size. This is equivalent to inductive conformal prediction with the non-conformity measure $f(x, y)$, as explained in [32]. Lastly, we have the following theorem:

Theorem 1 ([26, 32]). *$C_{\hat{\tau}}$ is (ε, δ)-correct for $\hat{\tau}$ as in* (2).

Remark. *The optimization problem* (2) *returns the trivial solution $\hat{\tau} = 0$ if the the optimization problem is infeasible.*

3.2 Semi-supervised Anomaly Detection

We assume each labeled data point consists of features and a label, $z_i = (x_i, y_i)$, with $y_i = 1$ indicating an anomaly (positive) and $y_i = 0$ indicating a normal (negative) data point. In a general semi-supervised anomaly detection setting, given an observed labeled data set $\{z_1, \ldots, z_N, z_{N+1}, \ldots, z_{N+K}\}$, we assume that $\{z_{N+1}, \ldots, z_{N+K}\}$ with $K \ll N$ is a small set of anomalies. At the same time, the rest of the data points are normal. We then use the observed data set as the calibration set, which contains both normal and anomalous data points. Finally, after getting the trained anomaly detector from the original semi-supervised training procedure, we calculate the PAC thresholds on the calibration set to identify anomalies.

4 METHOD

Suppose we are given a semi-supervised anomaly detector $d : \mathcal{X} \to \mathbb{R}$ which maps input $x \in \mathcal{X}$ to an anomaly score. We construct PAC prediction sets wrapped around $d(x)$ to control both false positive rate (FPR) and false negative rate (FNR). With our previous definition of positives, FPR is the rate of falsely classifying the normal class as anomalous:

$$FPR = \mathbb{P}(\hat{y} = 1 \mid y = 0).$$

Further, FNR is the rate of erroneously predicting the anomalous class as normal:

$$FNR = \mathbb{P}(\hat{y} = 0 \mid y = 1).$$

The control of the two rates is accomplished by replacing the original prediction error loss (as in (1)) with one that considers either FNR or FPR, which we use to construct a false negative PAC prediction set and a false positive PAC prediction set. We then propose to take the intersection of the two sets to provide a combined guarantee, which inevitably introduces ambiguity regions. Lastly, we propose a strategy to remove such ambiguity by considering the relative position of the two prediction sets.

4.1 Conditional Prediction Sets

In this section, we illustrate in detail our pipeline of loss modification, threshold derivation and the PAC prediction sets construction.

False positive PAC prediction set. Let the false positive PAC prediction set be $C_{\hat{\tau}_{\text{fp}}}$. The loss of $C_{\hat{\tau}_{\text{fp}}}$ is calculated on the normal data distribution $\mathcal{D}_{\text{nm}}$, and it is defined as:

$$L_{\mathcal{D}_{\text{nm}}}(C_{\hat{\tau}_{\text{fp}}}) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{nm}}} \ell_{\text{fp}}^{01}(C_{\hat{\tau}_{\text{fp}}}, x, y), \tag{3}$$

where $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{nm}}}(\cdot)$ means taking the expectation over the normal data distribution, and $\ell_{\text{fp}}^{01}(\cdot) \coloneqq \mathbb{1}(y \notin C_{\hat{\tau}_{\text{fp}}}(x))$. In other words, $\ell_{\text{fp}}^{01}(C_{\hat{\tau}_{\text{fp}}}, x, y)$ indicates whether the correct label 0 is not included in $C_{\hat{\tau}_{\text{fp}}}(x)$.

Let $Z_{\text{nm}} \sim D_{\text{nm}}^n$ be an independent calibration set of i.i.d. data points from D_{nm}. Given a user-specified $(\varepsilon_{\text{fp}}, \delta_{\text{fp}})$, we construct $C_{\hat{\tau}_{\text{fp}}}$ by identifying the optimal $\hat{\tau}$ in equation (2) via binary search using Z_{nm}. We denote the identified $\hat{\tau}$ as $\hat{\tau}_{\text{fp}}$. Then, we construct the $C_{\hat{\tau}_{\text{fp}}}(x)$ for $\hat{y}$ based on $d(x)$ in the following way:

$$C_{\hat{\tau}_{\text{fp}}}(x) \coloneqq \begin{cases} \{1\}, & \text{if } d(x) \ge \hat{\tau}_{\text{fp}} \\ \{0, 1\}, & \text{otherwise} \end{cases}. \tag{4}$$

In other words, we predict the set {1} for x with anomaly scores above $\hat{\tau}_{\text{fp}}$, and {0, 1} otherwise. We have Corollary 1 on the false positive PAC prediction. See Appendix B for a proof.

Corollary 1. *$C_{\hat{\tau}_{fp}}$ is $(\varepsilon_{fp}, \delta_{fp})$-correct for $\hat{\tau}_{fp}$ identified from (2) using loss function (3).*

Given an input x and $C_{\hat{\tau}_{\text{fp}}}$, we can make a class label prediction as:

$$\hat{y}_{\text{fp}} = \mathbb{1}(0 \notin C_{\hat{\tau}_{\text{fp}}}(x)). \tag{5}$$

In other words, we identify the current data point as anomalous if label 0 is not included in $C_{\hat{\tau}_{\text{fp}}}$, and as normal otherwise. Then, we have Theorem 2 on the false positive PAC prediction set. See Appendix C for a proof.

Theorem 2. *$C_{\hat{\tau}_{fp}}$ provides a PAC guarantee on the false positive rate:*

$$\mathbb{P}_{Z_{nm} \sim D_{nm}^{n}} \left[\mathbb{P}_{(x,y) \sim D_{nm}} (\hat{y}_{fp} = 1 \mid y = 0) \leq \varepsilon_{fp} \right] \geq 1 - \delta_{fp}.$$

False negative PAC prediction set. We denote the false negative PAC set by $C_{\hat{\tau}_{\text{fn}}}$. The loss of $C_{\hat{\tau}_{\text{fn}}}$ is calculated on the anomalous data distribution $\mathcal{D}_{\text{ano}}$, and it is defined as:

$$L_{\mathcal{D}_{\text{ano}}}(C_{\hat{\tau}_{\text{fn}}}) = \mathbb{E}_{(x,y) \sim D_{\text{ano}}} \ell_{\text{fn}}^{01}(C_{\hat{\tau}_{\text{fn}}}, x, y), \tag{6}$$

where $\ell_{\text{fn}}^{01}(\cdot) := \mathbb{1}(y \notin C_{\hat{\tau}_{\text{fn}}}(x))$. In other words, $\ell_{\text{fn}}^{01}(C_{\hat{\tau}_{\text{fn}}}, x, y)$ indicates whether the correct label 1 is not included in $C_{\hat{\tau}_{\text{fn}}}(\cdot)$.

Let $Z_{\text{ano}} \sim D_{\text{ano}}^{n}$ be an independent calibration set of i.i.d. data points from D_{ano}. Given a user-specified $(\varepsilon_{\text{fn}}, \delta_{\text{fn}})$, we construct $C_{\hat{\tau}_{\text{fn}}}$ by identifying the optimal $\hat{\tau}$ in equation (2) via binary search using Z_{ano}. We denote the identified $\hat{\tau}$ as $\hat{\tau}_{\text{fn}}$. We then construct the false negative PAC prediction set by

$$C_{\hat{\tau}_{fn}}(x) := \begin{cases} \{0, 1\}, & \text{if } d(x) \geq \hat{\tau}_{\text{fn}} \\ \{0\}, & \text{otherwise} \end{cases}. \tag{7}$$

Then, we have the following Corollary for the false negative PAC prediction set. See Appendix D for a proof.

Corollary 2. *$C_{\hat{\tau}_{fn}}$ is $(\varepsilon_{fn}, \delta_{fn})$-correct for $\hat{\tau}_{fn}$ identified from (2) using loss function (6).*

Moreover, similarly to above, we can define

$$\hat{y}_{\text{fn}} = \mathbb{1}(1 \in C_{\hat{\tau}_{\text{fn}}}(x)). \tag{8}$$

Finally, we have the associated PAC guarantee for the false negative prediction set in Theorem 3. See Appendix E for a proof.

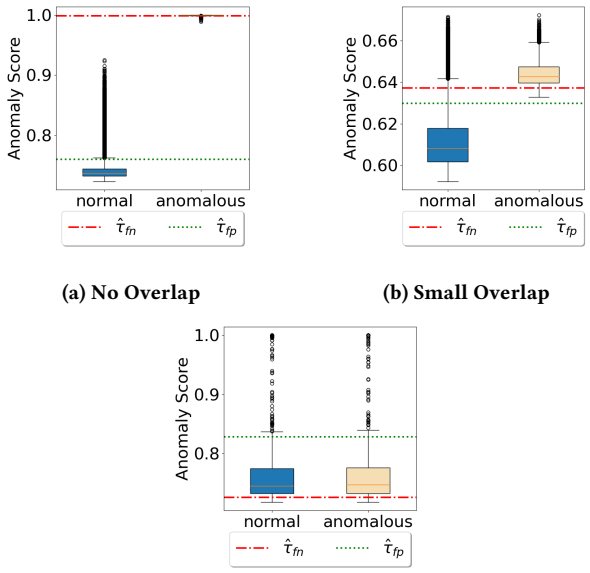
Theorem 3. *$C_{\hat{\tau}_{fn}}$ provides a PAC guarantee on false negative rate:*

$$\mathbb{P}_{Z_{ano} \sim D_{ano}^{n}} \left[\mathbb{P}_{(x,y) \sim D_{ano}} (\hat{y}_{fn} = 0 \mid y = 1) \leq \varepsilon_{fn} \right] \geq 1 - \delta_{fn}.$$

4.2 Anomaly detection with ambiguity region

We aim to use both false positive and false negative PAC prediction sets so that both rates are controlled at the same time. Consequently, we propose to combine false positive and false negative PAC prediction sets via taking their intersection:

$$C_{\text{ad}}(x) := C_{\hat{\tau}_{\text{fn}}}(x) \cap C_{\hat{\tau}_{\text{fp}}}(x).$$



(a) No Overlap **(b) Small Overlap**

(c) Large Overlap

Figure 2: An illustration of ambiguity region cases. $\hat{\tau}_{\text{fn}} \geq \hat{\tau}_{\text{fp}}$ happens when the overlap between the normal and anomalous class is zero or small, as in 2a and 2b; otherwise, $\hat{\tau}_{\text{fp}} \geq \hat{\tau}_{\text{fn}}$ happens, as in 2c.

There are four possible values of the intersection, depending on the relative position of the anomaly score $d(x)$, $\hat{\tau}_{\text{fn}}$ and $\hat{\tau}_{\text{fp}}$, listed in Table 1.

$d(x)$	$< \hat{\tau}_{\text{fn}}$	$\geq \hat{\tau}_{\text{fn}}$
$< \hat{\tau}_{\text{fp}}$	0	{0, 1}
$\geq \hat{\tau}_{\text{fp}}$	$\emptyset$	1

Table 1: The four possible values for $C_{\text{ad}}(x)$.

Given an input x, if its anomaly score $d(x)$ falls into the interval $[\hat{\tau}_{\text{fp}}, \hat{\tau}_{\text{fn}}]$, (or $[\hat{\tau}_{\text{fn}}, \hat{\tau}_{\text{fp}}]$), C_{ad} will contain zero or two labels, which is ambiguous. Therefore, the interval $[\hat{\tau}_{\text{fp}}, \hat{\tau}_{\text{fn}}]$ (or $[\hat{\tau}_{\text{fn}}, \hat{\tau}_{\text{fp}}]$) is defined as the *ambiguity region*, denoted as $\mathcal{U}$. Lemma 1 further explains the setting when $\mathcal{U}$ occurs. See Appendix H for a proof. Intuitively, when there is zero or a small overlap between the normal and anomalous classes, the ambiguity region $\mathcal{U}$ is $[\hat{\tau}_{\text{fp}}, \hat{\tau}_{\text{fn}}]$. This corresponds to the case $C_{\text{ad}} = \emptyset$. A visualization of this case is in Figure 2a and Figure 2b, where there is no overlap (2a) or little overlap (2b) between the normal and anomalous classes.

Lemma 1. *Let k_{fp}^{*} and k_{fn}^{*} be the solutions of (2) when identifying false positive and false negative PAC prediction sets respectively. We have that*

$$\begin{aligned} & \hat{\tau}_{fn} \geq \hat{\tau}_{fp} \\ \iff & \\ & \sum_{(x,y) \in Z_{nm}} \mathbb{1}(d(x) > \hat{\tau}_{fn}) < k_{fp}^{*} \\ \text{and} & \sum_{(x,y) \in Z_{ano}} \mathbb{1}(d(x) < \hat{\tau}_{fp}) < k_{fn}^{*}. \end{aligned} \tag{9}$$

In this case, we predict a class label as

$$\hat{y}_{\text{ad}} \coloneqq \begin{cases} 1, & C_{\text{ad}}(x) = \{1\} \\ 0, & C_{\text{ad}}(x) = \{0\} \\ *, & C_{\text{ad}}(x) = \emptyset \end{cases}, \tag{10}$$

where $*$ means *abstaining from predicting*. This classification with rejection option idea is similar to [2, 11], where a classifier could abstain from classifying an input if the classifier is not sufficiently confident about its prediction.

Further, we define the error rate $\text{ERR}(\mathcal{D})$ over the distribution $\mathcal{D}$ as the probability that the prediction is not equal to the label:

$$\text{ERR}(\mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(y \neq \hat{y}|x).$$

Then, we have the following theorem about the resulting anomaly detector. See Appendix F for a proof.

Theorem 4. *If $C_{\hat{\tau}_{fn}}$ is $(\varepsilon_{fn}, \delta_{fn})$-correct, $C_{\hat{\tau}_{fp}}$ is $(\varepsilon_{fp}, \delta_{fp})$-correct, and $\hat{\tau}_{fn} \geq \hat{\tau}_{fp}$, with probability at least $1 - \delta_{ad}$, where $\delta_{ad} = \delta_{fn} + \delta_{fp}$, the error rate $ERR(\mathcal{D})$ is no greater than ε_{ad}, where $\varepsilon_{ad} = \max(\varepsilon_{fp}, \varepsilon_{fn})$, i.e.,*

$$\mathbb{P}_{Z\sim\mathcal{D}_n}\left[ERR(\mathcal{D}) \leq \varepsilon_{ad})\right] \geq 1 - \delta_{ad}.$$

In contrast, $C_{\text{ad}} = \{0, 1\}$, i.e., $\mathcal{U} = [\hat{\tau}_{\text{fn}}, \hat{\tau}_{\text{fp}}]$, happens when the overlap between normal and anomalous class is large, see Figure 2c. According to Lemma 1, this arises when condition in (9) fails. In this case, the anomaly detector cannot distinguish anomalies from normal data points and therefore cannot satisfy the false positive and false negative constraints at the same time.

Thus, we have to either find a better anomaly detector or relax the constraint for the error or confidence. We propose Algorithm 1 to relax the error constraint. Intuitively, this algorithm first checks whether $\hat{\tau}_{\text{fn}} > \hat{\tau}_{\text{fp}}$ or not. If not, this algorithm increases the ε and recalculates $\hat{\tau}_{\text{fn}}$ and $\hat{\tau}_{\text{fp}}$. After $\hat{\tau}_{\text{fn}} > \hat{\tau}_{\text{fp}}$ is satisfied, this algorithm returns the resulting $\hat{\tau}'_{\text{fn}}$, $\hat{\tau}'_{\text{fp}}$, ε and δ. Here, we use the linear search strategy where we increase the ε by (say) 0.1 at each iteration. Alternatively, we could also double ε at each iteration, which may be faster to find a feasible ε, but the result may be looser. The confidence constraint can be relaxed similarly, but the effect is less salient than that of the error constraint.

Algorithm 1 Relaxing the error constraint

Input: $\hat{\tau}_{\text{fn}}, \hat{\tau}_{\text{fp}}, \varepsilon_{\text{fn}}, \varepsilon_{\text{fp}}, \delta_{\text{fn}}, \delta_{\text{fp}}, \Delta$ (default $\Delta = 0.1$).
Output: $\hat{\tau}'_{\text{fn}}, \hat{\tau}'_{\text{fp}}, \varepsilon, \delta$.
while $\hat{\tau}_{\text{fn}} < \hat{\tau}_{\text{fp}}$ and $\varepsilon_{\text{fn}}, \varepsilon_{\text{fp}} \leq 1$ **do**
  $\varepsilon_{\text{fn}}, \varepsilon_{\text{fp}} = \varepsilon_{\text{fn}} + \Delta, \varepsilon_{\text{fp}} + \Delta$.
  Re-calculate $\hat{\tau}_{\text{fn}}, \hat{\tau}_{\text{fp}}$ using equation (2) with $\delta_{\text{fn}}, \delta_{\text{fp}}$ correspondingly.
end while
$\varepsilon, \delta = \max(\varepsilon_{\text{fn}}, \varepsilon_{\text{fp}}), \delta_{\text{fn}} + \delta_{\text{fp}}$.
$\hat{\tau}'_{\text{fn}}, \hat{\tau}'_{\text{fp}} = \hat{\tau}_{\text{fn}}, \hat{\tau}_{\text{fp}}$.
Return $\hat{\tau}'_{\text{fn}}, \hat{\tau}'_{\text{fp}}, \varepsilon, \delta$.

4.3 Anomaly detection with certain prediction

If one is not allowed to abstain from making a prediction, the ambiguity region $\mathcal{U}$ must be removed. In this case, after satisfying $\hat{\tau}'_{\text{fn}} \geq \hat{\tau}'_{\text{fp}}$, we could pick an arbitrary threshold $\tau \in [\hat{\tau}'_{\text{fp}}, \hat{\tau}'_{\text{fn}}]$, e.g., $\tau = (\hat{\tau}'_{\text{fn}} + \hat{\tau}'_{\text{fp}})/2$, and the guarantees will still hold. We state this claim in Theorem 5. See Appendix G for a proof.

Theorem 5. *After using Algorithm 1 and picking an arbitrary threshold $\tau \in [\hat{\tau}'_{fp}, \hat{\tau}'_{fn}]$ for an anomaly detector, its error rate is at most ε, with probability at least $1 - \delta$.*

5 EXPERIMENTAL RESULTS

We apply PAC-Wrap to several anomaly detectors, and on both i.i.d. and time series anomaly detection datasets. The experiments support that PAC-Wrap enables PAC guarantees on the false positive rate (FPR) and false negative rate (FNR). In addition, we compare with standard class-conditional conformal prediction [18]. These experiments empirically support that PAC-Wrap performs well in a variety of scenarios, and compares favorably to standard conformal prediction-based methods.

We address several questions to demonstrate the effectiveness of PAC-Wrap:

- **Q1 (Empirical Validation)**: Are Theorem 2 and 3, empirically supported by results on both synthetic and benchmark datasets?
- **Q2 (Wrapper Effect)**: How does our wrapper affect the underlying anomaly detector's error rates?
- **Q3 (Baseline Comparison)**: How does our work compare to standard class-conditional conformal prediction methods?
- **Q4 (Ablation Study)**: How do different combinations of ε and δ affect the ambiguity region?

5.1 Datasets, Anomamly Detectors, and Metrics

5.1.1 Datasets. We first describe the synthetic and benchmark datasets used in **Q1**. We generate a synthetic dataset by sampling i.i.d. normal and anomalous data points from two clusters, each normally distributed in 6-dimensional space with the same covariance matrix but with different means $\mu_{\text{normal}}, \mu_{\text{anomalous}} \in \mathbb{R}^6$, which are selected so that the two classes are separated by a margin of 5. Let I_p be the p-dimensional identity matrix with $p = 6$, and σ^2 be a uniformly random value drawn over $[1, 100]$. We have:

$$X_{\text{normal}} \sim \mathcal{N}(\mu_{\text{normal}}, \sigma^2 I_p)$$
$$X_{\text{anomalous}} \sim \mathcal{N}(\mu_{\text{anomalous}}, \sigma^2 I_p).$$

To simulate the semi-supervised problem, we generate 100,000 normal data points as the training set, another 2,000 normal and 2,000 anomalous data points as the calibration set, and finally 50,000 normal and 50,000 anomalous data points as the test set. The benchmark dataset *thyroid* is a UCI Machine Learning Repository [8] dataset that contains around 7,200 data points. It treats the *hypothyroid* disease as an anomaly. We randomly sample 80% of the normal data points from the *thyroid* dataset to form the training set. We then take the remaining 20% of the normal data and the anomalous data points to form the calibration and the test set, with the calibration set taking up 30% and the test set taking up 70%.

In **Q2**, we experiment on the benchmark semi-supervised anomaly detection datasets *campaign*, *celeba*, and *census* that are also used in [22]. The *campaign* dataset contains direct marketing campaigns (phone calls) and asks to predict whether a given client will subscribe to a term deposit. Successful campaigning records account for approximately 10% records and are regarded as anomalies. The *celeba* dataset is an image dataset of more than 200K celebrity images. In this task, the anomaly detector detects bald celebrities as anomalies, which account for less than 3% of celebrities. The *census* dataset is extracted from the US census bureau database and aims to detect the high-income people that comprise about 6% of the data as "anomalies". In contrast to the typical supervised classification setting, these datasets are highly imbalanced. In other words, only a small portion of labeled data points are anomalous.

We also conduct experiments on two time series benchmark datasets in **Q2**, the Server Machine Dataset (SMD) [29], and the NASA Telemetry Anomaly Detection (NASA) dataset [12] to see how PAC-Wrap affects the performance of time series anomaly detectors. The detailed result is reported in Appendix I.

In **Q3**, we use the same experimental setup on the MNIST dataset [7] as in [18]: we regard the digits $\{0, 6, 9\}$ as class "0" and digit $\{8\}$ as class "1". The training dataset contains 3044 images, with 541 in class 1. The test dataset contains 872 images, with 166 in class 1. As in [18], we train ℓ_1-penalized logistic regression on two-thirds of the training data points and use the remaining one-third as the calibration data to identify the Conformal/PAC prediction sets. In the calibration dataset, we have 865 images in class 0 and 170 images in class 1.

Finally, we use the same synthetic dataset in **Q4** as in **Q1**.

5.1.2 Anomaly Detectors. We consider the following anomaly detectors:

- **Isolation Forest** (IF) [19] is an unsupervised model based on decision trees.
- **Local Outlier Factor** (LOF) [4] is an unsupervised anomaly detection method which compares an estimated density of a data point to its neighbors.
- **DevNet** [22] is a semi-supervised anomaly detector that uses a few labeled anomalies to separate the anomalies from normal data points.
- **LSTM-based anomaly detector** [3, 12] is commonly used for time series data. For SMD, we wrap around a standard LSTM-encoder-decoder-based anomaly detector [3]. For the NASA data, we use the proposed LSTM-based anomaly detector in [12].

If an anomaly score threshold is not explicitly identified for the above anomaly detectors, we use a threshold that maximizes the F1 score, i.e., the harmonic mean of precision and recall. The F1 score is often used as an efficacy measure in the anomaly detection literature.

5.1.3 Metrics. Let TP, TN, FP, FN be the number of true positives, true negatives, false positives, and false negatives, respectively. We focus on the three most important error rates in anomaly detection: FNR = FN/(FN+TP) and FPR = FP/(FP+TN), ERR = (FN+FP)/(FN+TP+FP+TN). We compare FPR, FNR and ERR to a user-specified error constraint (e.g., $\varepsilon = 0.05$). We repeatedly run the experiments and check if the *error constraint violation rate*, defined as the fraction of times the error rate is above ε, is lower than a user-specified confidence constraint (e.g., $\delta = 0.05$). To compare PAC-Wrap with a conformal prediction-based baseline, we use the definition of ambiguity from [18], estimated as the fraction of data points falling into the ambiguity region in the test dataset:

$$Ambiguity = \frac{\sum_{(x,y)\in Z_{\text{test}}} \mathbb{1}(d(x) \in \mathcal{U})}{|Z_{\text{test}}|}. \tag{11}$$

5.2 Q1. Empirical Validation

We first empirically validate the theoretical guarantees of our false negative and false positive PAC prediction sets. To study how anomaly detector performance affects our guarantees, we experiment with two kinds of anomaly detectors, the Local Outlier Factor (LOF) and the Isolation Forest (IF). Additionally, to study how calibration set size affects our guarantees, we experiment with 50%, 75%, and 100% of the calibration set. In these experiments, we set $\varepsilon_{\text{fn}} = \varepsilon_{\text{fp}} = 0.05$ and $\delta_{\text{fn}} = \delta_{\text{fp}} = 0.05$ as our constraints. Besides, 4000 independent Monte Carlo trials on both synthetic and benchmark datasets are performed. Out of these trials, we compute the empirical error constraint violation rate, which is the fraction of trials where the FPR or FNR is above 0.05.

Note that the PAC guarantee assumes an infinite population, but we only have a finite dataset. To address this problem, we propose the following method. First, we combine the calibration and test datasets to form a known finite population $\mathcal{D}$. We aim to validate the PAC guarantee over the known finite population $\mathcal{D}$, which is convenient since we can enumerate the population. Next, we train the LOF and IF on the training set. For each Monte Carlo trial, we then sample with replacement a new calibration set from the known finite population, of the same size and anomaly ratio as the original calibration set. We construct false positive and false negative PAC prediction sets on each newly sampled calibration set. Finally, we compute the FPR and FNR of the constructed PAC prediction sets over the known finite population $\mathcal{D}$.

We report, in Table 2 (on synthetic data) and Table 3 (on the *thyroid* dataset), a two-sided 95% Clopper–Pearson interval for the error constraint violation rate. If the interval covers 0.05 (or falls below that), the empirical results are consistent with the error and confidence constraints being satisfied. In Table 2 and Table 3, the PAC guarantee is corroborated by the results on the synthetic and benchmark datasets since all the intervals fall below 0.05. The guarantee holds regardless of calibration set size and anomaly detectors. As a result, our results empirically validate Theorems 2 and 3. Another observation is that the constraint violation rates on the benchmark dataset are much lower than 0.05, which means that the constructed PAC prediction sets on the benchmark dataset are conservative. We further discuss this observation in Appendix A.

5.3 Q2. Wrapper Effect

In this section, we conduct experiments to check how PAC-Wrap affects the error rates of the underlying anomaly detector. Specifically, we apply Algorithm 1 to remove the ambiguity region and check if the final FPR, FNR and ERR are bounded by the error constraint. For brevity, we omit repeatedly verifying the confidence

Violation	Val Size	IF	LOF
Pr(FPR>0.05)	50%	[0.034, 0.046]	[0.033, 0.045]
	75%	[0.024, 0.034]	[0.020, 0.030]
	100%	[0.031, 0.043]	[0.030, 0.042]
Pr(FNR>0.05)	50%	[0.033, 0.045]	[0.034, 0.047]
	75%	[0.032, 0.044]	[0.035, 0.048]
	100%	[0.035, 0.047]	[0.032, 0.044]

Table 2: The rate of error constraint violation on the synthetic data.

Violation	Val Size	IF	LOF
Pr(FPR > 0.05)	50%	[0.024, 0.035]	[0.022, 0.033]
	75%	[0.007, 0.013]	[0.006, 0.012]
	100%	[0.025, 0.036]	[0.026, 0.037]
Pr(FNR > 0.05)	50%	[0.006, 0.012]	[0.006, 0.012]
	75%	[0.010, 0.018]	[0.014, 0.022]
	100%	[0.012, 0.020]	[0.016, 0.025]

Table 3: The rate of error constraint violation on the benchmark dataset.

constraint, which is already tested in **Q1**. We report the following values:

- **FNR_{or}, FPR_{or}**: the original FNR and FPR of the anomaly detector without our wrapper.
- **FNR_{tt}, FPR_{tt}**: the FNR and FPR of our wrapper using two thresholds $\hat{\tau}_{\text{fn}}, \hat{\tau}_{\text{fp}}$, given the initial error constraints.
- **FNR_{th}, FNP_{th}**: the FNR and FPR of our wrapper using one final threshold τ, given the (possibly relaxed) error constraint.
- **ERR**: the final error rate of our wrapper, which is defined in 5.1.3. It is a weighted combination of FNR_{th} and FPR_{th}.
- ε: the final error level our wrapper can guarantee.

For i.i.d. data, we take DevNet [22] as the baseline anomaly detector. We first train DevNet on the *campaign*, *celeba*, and *census* datasets respectively using default hyperparameters. Second, we wrap the trained model with the constructed false negative and false positive PAC prediction sets. Then, we find that DevNet may not perform well enough to simultaneously satisfy the user-specified errors constraints with the PAC prediction sets. Therefore, we use Algorithm 1 to adaptively relax the error constraint to enable DevNet to fulfill a reasonable guarantee.

In Figure 3, we show how this works on the *celeba* dataset. Specifically, $\hat{\tau}_{\text{fn}}$ and $\hat{\tau}_{\text{fp}}$ are first chosen to satisfy the constraint $\varepsilon_{\text{fn}} = \varepsilon_{\text{fp}} = 0.05, \delta_{\text{fn}} = \delta_{\text{fp}} = 0.05$. Although most anomalies have higher anomaly scores than the normal data points, there is still considerable overlap. As a result, $\hat{\tau}_{\text{fp}}$ is greater than $\hat{\tau}_{\text{fn}}$ (the green dashed line is above the red dashed line in Figure 3), which is an inconclusive case as discussed in 4.2. In other words, the anomaly detector cannot accurately distinguish the normal and the anomalous classes under the current error constraint, and we have to relax the constraint. After relaxing the error constraint by Algorithm 1, we find $\varepsilon = 0.15$ and $\hat{\tau}'_{\text{fn}} \geq \hat{\tau}'_{\text{fp}}$ (solid red line is above solid green line in Figure 3). Then, we can readily remove the ambiguity region by setting $\tau = (\hat{\tau}'_{\text{fn}} + \hat{\tau}'_{\text{fp}})/2$ (dashed blue line in Figure 3) according to Theorem 5. A similar process occurs on the *campaign* and *census* datasets, resulting in the relaxed error constraints of $\varepsilon = 0.35$ and $\varepsilon = 0.25$ respectively.

We report the detailed results for DevNet in Table 4. Table 4 first shows that the FNR_{or}-s and FPR_{or}-s of DevNet violate the error

	FNR_{or}	FPR_{or}	FNR_{tt}	FPR_{tt}	FNR_{th}	FPR_{th}	ERR	ε
campaign	0.000	0.998	0.026	0.043	0.267	0.259	0.266	0.35
celeba	0.029	0.456	0.026	0.042	0.121	0.097	0.120	0.15
census	0.055	0.561	0.048	0.047	0.230	0.202	0.229	0.25

Table 4: Error rate with PAC-Wrap wrapped around DevNet on i.i.d. data. Guarantees on FNR and FPR are met. After removing the ambiguity region, the FPR_{th}, FPR_{th}, and ERR satisfy the error constraints.

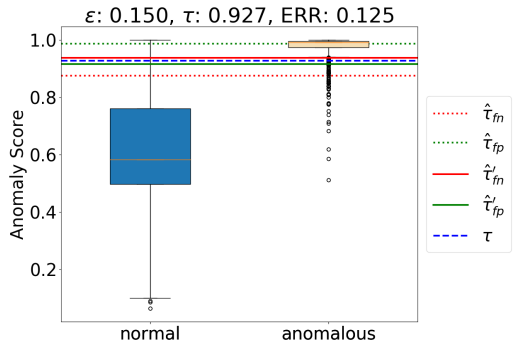
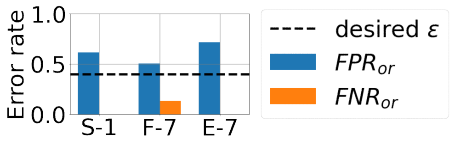


Figure 3: Box Plot and thresholds on the *celeba* dataset using the DevNet anomaly detector. $\hat{\tau}_{\text{fp}} > \hat{\tau}_{\text{fn}}$ holds under the original error constraint ($\varepsilon \leq 0.05$). By using Alg. 1, $\hat{\tau}'_{\text{fn}} \geq \hat{\tau}'_{\text{fp}}$ under the relaxed error constraint ($\varepsilon \leq 0.15$).

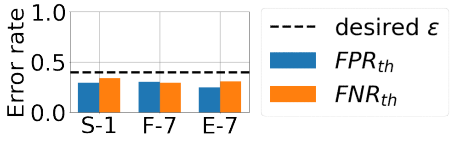
constraint $\varepsilon_{\text{fn}} = \varepsilon_{\text{fp}} = 0.05$ on *campaign*, *celeba*, and *census* datasets. Then, columns FNR_{tt} and FPR_{tt} indicate that PAC-Wrap satisfies the original error constraints. After the constraint relaxation and ambiguity removal, columns FPR_{th}, FPR_{th}, and ERR are lower than the last column ε, indicating that they all satisfy the relaxed error constraints. The underlying anomaly detectors determine the relaxed levels. Without our wrapper, the baselines can usually only control one of the FNR/FPR. Our method provides a principled way to balance the two rates and provides guarantees on their levels.

Time series data are beyond the independence assumptions required for the PAC property, but can be transformed to reduce the dependence across time. We show the detailed result in the in Appendix I that, in certain cases, our wrapper is also effective for time series anomaly detectors. Initially, we set $\varepsilon_{\text{fn}} = \varepsilon_{\text{fp}} = 0.05, \delta_{\text{fn}} = \delta_{\text{fp}} = 0.05$ and find that the sample size is occasionally too small to satisfy the error and confidence constraints. This is because, given a user-specified ε and δ, we have a minimum requirement for the number of data points. According to Theorem 1 in [26], the number of data points n should be at least $\log(1/\delta)/\log(1-\varepsilon)$. For instance, if $\varepsilon = \delta = 0.05$, the minimum required sample sizes for labeled normal and anomalous data points are both 59. When only limited labeled anomalies are available, we can relax the error and confidence level to give PAC guarantees. In the time series experiments where only 30 to 40 labeled anomalies are available, we set $\varepsilon_{\text{fn}} = \varepsilon_{\text{fp}} = 0.10$, $\delta_{\text{fn}} = \delta_{\text{fp}} = 0.10$ to compute the thresholds. After training the LSTM-based anomaly detectors on the training set, we find that they have the same performance issue as DevNet. Hence, we perform a similar constraint relaxation and ambiguity removal procedure.

We show in Figure 4 some representative channels from the NASA dataset. The LSTM-based anomaly detector violates the error constraint, but PAC-Wrap controls both FPR_{th} and FNR_{th} to be



(a) The $\varepsilon = 0.4$ error constraint does not hold for the original anomaly detector.



(b) With our wrapper, the $\varepsilon = 0.4$ error constraint is met.

Figure 4: Results on the NASA data with $\varepsilon = 0.4$. For the NASA anomaly detector, PAC-Wrap helps balance the FNR and FPR.

smaller than ε. For example, for the S-1, F-7, and E-7 channels, FPR_{or} is even greater than the relaxed error constraint $\varepsilon = 0.4$. With a moderate increase in FNR_{th}, our wrapper can ensure both FPR_{th} and FNR_{th} are below $\varepsilon = 0.4$.

5.4 Q3. Baseline Comparison

We compare PAC-Wrap to the method in [18]—denoted as *CPAD*—which also provides guarantees on the FNR and FPR by calculating per-class thresholds. The error and confidence constraints are set as $\varepsilon_{\text{fn}} = \varepsilon_{\text{fp}} = 0.05, \delta_{\text{fn}} = \delta_{\text{fp}} = 0.05$. We use the same known finite population method as in 5.2 for evaluation. Specifically, after constructing the training, calibration and test datasets as in [18], we construct the known finite population by combining the calibration dataset with the test dataset. Next, we perform 300 independent Monte Carlo trials and compute the error constraint violation rates on the known finite population. To identify the conformal/PAC prediction sets, we sample with replacement a new calibration dataset with the same size and anomaly ratio as the original calibration dataset in each trial. After constructing the prediction sets, we evaluate the FNR and FPR on the known finite population.

The average FPR and FNR over the 300 Monte Carlo trials are reported in Table 5. The result shows that the average FNR-s and FPR-s of CPAD and PAC-Wrap are basically below 0.05, therefore satisfying the error constraints. This finding is consistent with the class-conditional guarantees of PAC-Wrap and CPAD. To evaluate the satisfaction of the confidence constraint, we report a two-sided 95% Clopper–Pearson interval for the error constraint violation rate in Table 6. The result shows that CPAD's violation rates are approaching 50%, while that of PAC-Wrap are at the desired level (below 0.05). That is because CPAD's guarantee holds marginally over the training dataset, which differs from the conditional guarantee of PAC-Wrap. It is possible that an insufficiently representative calibration set is drawn, and PAC-Wrap accounts for the scenario via introducing the confidence parameter δ. In contrast, standard conformal prediction-based methods like CPAD do not consider the data representativeness and cannot provide a training-set conditional guarantee with high confidence. While satisfying the error constraint with much higher probability, PAC-Wrap induces slightly higher ambiguity than that of CPAD, as shown in Table 5. However,

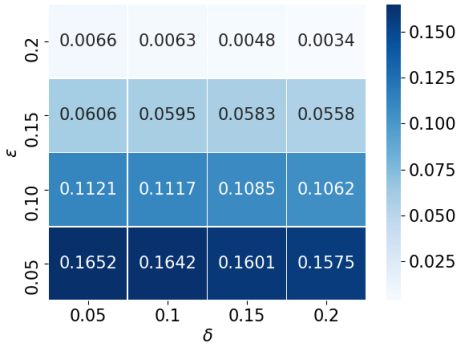


Figure 5: The average ambiguity as a function of ε and δ. As ε and δ grow, the ambiguity shrinks.

the increment in ambiguity is mostly tolerable, especially in safety-critical applications where the violation of the error constraint might lead to a catastrophe.

Dataset	FPR	FNR	Ambiguity
Desired	0.050	0.050	0
CPAD	0.049	0.051	0.222
PAC-Wrap	0.038	0.020	0.345

Table 5: Average FNR and FPR for CPAD and PAC-Wrap. On Average, both CPAD and PAC-Wrap satisfy the error constraint.

Method	Pr(FPR > 0.05)	Pr(FNR > 0.05)
CPAD	[0.344, 0.458]	[0.498, 0.614]
PAC-Wrap	[0.001, 0.024]	[0.000, 0.018]

Table 6: Comparison of the error constraint violation rate of the CPAD and PAC-Wrap. CPAD violates the error constraint for nearly 50% of the time and hence fails the confidence constraint. PAC-Wrap satisfies the 0.05 confidence constraint.

5.5 Q4. Ablation Study

In this experiment, we want to see how the ambiguity (defined in Equation (11)) changes with respect to the error parameter ε and confidence parameter δ. Specifically, we set $\varepsilon_{\text{fn}} = \varepsilon_{\text{fp}} = \varepsilon, \delta_{\text{fn}} = \delta_{\text{fp}} = \delta$. We then vary ε and δ, and construct false positive and false negative PAC prediction sets on the synthetic dataset. For every combination of error parameter and confidence parameter, we do 100 Monte Carlo trials and compute the average ambiguity. As shown in Figure 5, the ambiguity monotonically decreases with respect to ε and δ. It suggests that there is an empirical trade-off between the constraints and ambiguity. We can relax constraints to decrease the ambiguity or vice versa. Moreover, ε has a larger effect on the ambiguity than δ.

6 CONCLUSION AND DISCUSSION

We have developed a general framework called PAC-Wrap for guarantees in semi-supervised anomaly detection. Given many normal data points and a small number of anomalous data points, we use PAC-Wrap to control the false negative rate (FNR) and false positive rate (FPR). We conduct experiments on synthetic and benchmark datasets with various anomaly detectors to showcase the effectiveness of PAC-Wrap. Our method can readily wrap around virtually any existing anomaly detection algorithm, making our

framework an off-the-shelf tool to provide rigorous PAC guarantees for these algorithms. Our method can be applied to safety-critical applications such as autonomous vehicles, surveillance video, and tumor diagnosis. By leveraging a limited number of labeled datapoints, PAC-Wrap can guarantee the FPR and FNR of anomaly detectors, which is highly important.

PAC-Wrap can be directly extended to a multi-class framework to provide conditional guarantees for each class for an immediate next step. One limitation of PAC-Wrap is that if the normal and anomalous distributions in the testing stage are significantly different from those in the calibration stage, the false negative and false positive guarantees might not hold, since PAC-Wrap cannot automatically adapt to the distribution shift. To see how distribution shifts affect the guarantees, we show an additional experiment in Appendix J. In future work, it is important to enable PAC-Wrap to adapt to distribution shift during the testing stage.

7 ACKNOWLEDGMENTS

This work was supported in part by DARPA/AFRL FA8750-18-C-0090, ARO W911NF-20-1-0080, ONR N00014-20-1-2744, NSF-1915398, NSF-2125561, NSF-1934960, NSF-2046874 and SRC Task 2894.001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Air Force Research Laboratory (AFRL), the Army Research Office (ARO), the Defense Advanced Research Projects Agency (DARPA), the Office of Naval Research (ONR) or the Department of Defense, or the United States Government.

REFERENCES

[1] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. 2014. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.
[2] Peter L Bartlett and Marten H Wegkamp. 2008. Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research* 9, 8 (2008).
[3] Aadyot Bhatnagar, Paul Kassianik, Chenghao Liu, Tian Lan, Wenzhuo Yang, Rowan Cassius, Doyen Sahoo, Devansh Arpit, Sri Subramanian, Gerald Woo, et al. 2021. Merlion: A machine learning library for time series. *arXiv preprint arXiv:2109.09265* (2021).
[4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 93–104.
[5] Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep Learning for Anomaly Detection: A Survey. https://doi.org/10.48550/ARXIV.1901.03407
[6] Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. 2018. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*. PMLR, 732–749.
[7] Li Deng. 2012. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142.
[8] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[9] Matteo Guarrera, Baihong Jin, Tung-Wei Lin, Maria Zuluaga, Yuxin Chen, and Alberto Sangiovanni-Vincentelli. 2021. Class-wise Thresholding for Detecting Out-of-Distribution Data. arXiv:2110.15292 [cs.LG]
[10] I. Guttman. 1970. *Statistical Tolerance Regions: Classical and Bayesian*. Hafner Publishing Company. https://books.google.com/books?id=3Q7vAAAAMAAJ
[11] Radu Herbei and Marten H Wegkamp. 2006. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* (2006), 709–721.
[12] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 387–395.
[13] Jehn-Ruey Jiang and Jian-Bin Kao. 2021. Semi-supervised time series anomaly detection based on statistics and deep learning. *Applied Sciences* 11, 15 (2021), 6698.
[14] Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. 2022. iDECODe: In-distribution Equivariance for Conformal Out-of-distribution Detection. arXiv:2201.02331 [cs.LG]
[15] Michael J. Kearns and Umesh V. Vazirani. 1994. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, USA.
[16] Kalimuthu Krishnamoorthy and Thomas Mathew. 2009. *Statistical tolerance regions: theory, applications, and computation*. Vol. 744. John Wiley & Sons.
[17] Rikard Laxhammar and Göran Falkman. 2012. Online Detection of Anomalous Sub-trajectories: A Sliding Window Approach Based on Conformal Anomaly Detection and Local Outlier Factor. In *8th International Conference on Artificial Intelligence Applications and Innovations (AIAI) (Artificial Intelligence Applications and Innovations)*. Springer, Halkidiki, Greece, 192–202. https://doi.org/10.1007/978-3-642-33412-2_20 Part 4: First Conformal Prediction and Its Applications Workshop (COPA 2012).
[18] Jing Lei. 2014. Classification with confidence. *Biometrika* 101, 4 (2014), 755–769.
[19] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 1 (2012), 1–39.
[20] Manpreet Singh Minhas and John Zelek. 2020. Semi-supervised Anomaly Detection using AutoEncoders. arXiv:2001.03674 [eess.IV]
[21] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* 54, 2, Article 38 (mar 2021), 38 pages. https://doi.org/10.1145/3439950
[22] Guansong Pang, Chunhua Shen, and Anton van den Hengel. 2019. Deep Anomaly Detection with Deviation Networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).
[23] Harris Papadopoulos. 2008. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka.
[24] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. 2002. Inductive confidence machines for regression. In *European Conference on Machine Learning*. Springer, 345–356.
[25] Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. 2011. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research* 40 (2011), 815–840.
[26] Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. 2020. PAC Confidence Sets for Deep Neural Networks via Calibrated Prediction. arXiv:2001.00106 [cs.LG]
[27] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. 2019. Deep Semi-Supervised Anomaly Detection. *CoRR* abs/1906.02694 (2019). arXiv:1906.02694 http://arxiv.org/abs/1906.02694
[28] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.* 13, 7 (July 2001), 1443–1471. https://doi.org/10.1162/089976601750264965
[29] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2828–2837.
[30] Leslie G Valiant. 1984. A theory of the learnable. *Commun. ACM* 27, 11 (1984), 1134–1142.
[31] Vincent Vercruyssen, Wannes Meert, Gust Verbruggen, Koen Maes, Ruben Bäumer, and Jesse Davis. 2018. Semi-Supervised Anomaly Detection with an Application to Water Analytics. In *2018 IEEE International Conference on Data Mining (ICDM)*. 527–536. https://doi.org/10.1109/ICDM.2018.00068
[32] Vladimir Vovk. 2012. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*. PMLR, 475–490.
[33] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.
[34] Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. 2003. Mondrian confidence machine. *Technical Report* (2003).
[35] Samuel S Wilks. 1941. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics* 12, 1 (1941), 91–96.
[36] Chen Xu and Yao Xie. 2021. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*. PMLR, 11559–11569.
[37] Yahan Yang, Ramneet Kaur, Souradeep Dutta, and Insup Lee. 2022. Interpretable Detection of Distribution Shifts in Learning Enabled Cyber-Physical Systems. *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS) Interpretable* (2022), 225–235. https://doi.org/10.1109/ICCPS54341.2022.00027

A CONSERVATIVENESS

The conservativeness over *thyroid* dataset could be explained by the fact that the calibration set size (432) is significantly smaller than that of synthetic dataset (4000). Since a smaller calibration set is less representative of the true distribution, Equation (2) will

construct a possibly over-conservative prediction set to satisfy the confidence constraint, which leads to the violation rate being much lower than the confidence constraint. Moreover, a small calibration set, which is unrepresentative of the true distribution, could also contribute to a high violation rate. The fact that the calibration set is small could have the opposite effects on the violation rates. On the *thyroid* dataset, the effect of Equation (2) is dominant. As a result, the constructed PAC prediction sets are relatively conservative.

B PROOF OF COROLLARY 1

We replace the original prediction set $L_{\mathcal{D}}(C)$ with $L_{\mathcal{D}_{\text{nm}}}(C_{\hat{\tau}_{\text{fp}}})$, setting $\varepsilon = \varepsilon_{\text{fp}}, \delta = \delta_{\text{fp}}$, and construct the false positive PAC prediction set via solving (2). By Theorem 1, we have $\mathbb{P}_{Z\sim\mathcal{D}^n_{\text{nm}}}[L_{\mathcal{D}_{\text{nm}}}(C_{\hat{\tau}_{\text{fp}}}) \leq \varepsilon_{\text{fp}}] \geq 1-\delta_{\text{fp}}$. Therefore, $C_{\hat{\tau}_{\text{fp}}}$ is $(\varepsilon_{\text{fp}}, \delta_{\text{fp}})$-correct. □

C PROOF OF THEOREM 2

We have

$$\begin{aligned}&\Pr(\hat{y}=1 \mid y=0) = \mathbb{E}_{x|y=0}[\mathbb{1}(\hat{y}=1)]\\ &= \mathbb{E}_{x|y=0}[\mathbb{1}(0 \notin C_{\hat{\tau}_{\text{fp}}}(x))] = \mathbb{E}_{x|y=0}[\mathbb{1}(y \notin C_{\hat{\tau}_{\text{fp}}}(x)]\\ &= \mathbb{E}_{x|y=0}[\ell^{01}_{\text{fp}}(x)] = L_{D_{\text{nm}}}(C_{\hat{\tau}_{\text{fp}}}).\end{aligned}$$

By Corollary 1, we have

$$\mathbb{P}_{Z\sim D^n_{\text{nm}}}[L_{D_{\text{nm}}}(C_{\hat{\tau}_{\text{fp}}}) \leq \varepsilon_{\text{fp}}] \geq 1-\delta_{\text{fp}}.$$

Since $\Pr(\hat{y}=1 \mid y=0) = L_{D_{\text{nm}}}(C_{\hat{\tau}_{\text{fp}}})$, we find

$$\mathbb{P}_{Z\sim D^n_{\text{nm}}}[\mathbb{P}(\hat{y}=1 \mid y=0) \leq \varepsilon_{\text{fp}}] \geq 1-\delta_{\text{fp}}.\square$$

D PROOF OF COROLLARY 2

We replace the original prediction set $L_{\mathcal{D}}(C)$ with $L_{\mathcal{D}_{\text{ano}}}(C_{\hat{\tau}_{\text{fn}}})$, setting $\varepsilon = \varepsilon_{\text{fn}}, \delta = \delta_{\text{fn}}$, and construct the false positive PAC prediction set via solving (2). By Theorem 1, we have $\mathbb{P}_{Z\sim\mathcal{D}^n_{\text{ano}}}[L_{\mathcal{D}_{\text{ano}}}(C_{\hat{\tau}_{\text{fn}}}) \leq \varepsilon_{\text{fn}}] \geq 1-\delta_{\text{fn}}$. Therefore, $C_{\hat{\tau}_{\text{fn}}}$ is $(\varepsilon_{\text{fn}}, \delta_{\text{fn}})$-correct. □

E PROOF OF THEOREM 3

We have

$$\begin{aligned}&\Pr(\hat{y}=0 \mid y=1) = \mathbb{E}_{x|y=1}[\mathbb{1}(\hat{y}=0)]\\ &= \mathbb{E}_{x|y=1}[\mathbb{1}(1 \notin C_{\hat{\tau}_{\text{fn}}}(x))] = \mathbb{E}_{x|y=1}[\mathbb{1}(y \notin C_{\hat{\tau}_{\text{fn}}}(x)]\\ &= \mathbb{E}_{x|y=1}[\ell^{01}_{\text{fn}}(x)] = L_{D_{\text{ano}}}(C_{\hat{\tau}_{\text{fn}}}).\end{aligned}$$

By Corollary 2, we have

$$\mathbb{P}_{Z\sim D^n_{\text{ano}}}[L_{D_{\text{ano}}}(C_{\hat{\tau}_{\text{fn}}}) \leq \varepsilon_{\text{fn}}] \geq 1-\delta_{\text{fn}}.$$

Since $\Pr(\hat{y}=0 \mid y=1) = L_{D_{ano}}(C_{\hat{\tau}_{\text{fn}}})$, we find

$$\mathbb{P}_{Z\sim D^n_{\text{ano}}}[\mathbb{P}(\hat{y}=0 \mid y=1) \leq \varepsilon_{\text{fn}}] \geq 1-\delta_{\text{fn}}.\square$$

F PROOF OF THEOREM 4

When $d(x) \geq \hat{\tau}_{\text{fn}}$ and $d(x) \geq \hat{\tau}_{\text{fp}}$, by Equation (10), $\hat{y}=1$. In this case, the error rate ε_{ad} equals to the FPR. (The anomaly detector's prediction is correct when $y=1$.) By Theorem 2, we have

$$\mathbb{P}_{Z\sim D^n_{\text{nm}}}\left[\mathbb{P}(\hat{y}=1 \mid y=0) \leq \varepsilon_{\text{fp}}\right] \geq 1-\delta_{\text{fp}}.$$

In other words, the error rate when $d(x) \geq \hat{\tau}_{\text{fn}}$ and $d(x) \geq \hat{\tau}_{\text{fp}}$ satisfies

$$\mathbb{P}_{Z_{\text{nm}}\sim D^n_{\text{nm}}}\left[\varepsilon_{\text{ad}} \leq \varepsilon_{\text{fp}}\right] \geq 1-\delta_{\text{fp}}.$$

Similarly, when $d(x) \leq \hat{\tau}_{\text{fn}}$ and $d(x) \leq \hat{\tau}_{\text{fp}}$, by Equation 10, $\hat{y}=0$. In this case, the error rate ε_{ad} equals to the FNR. (The anomaly detector's prediction is correct when $y=0$.) By Theorem 3, we have

$$\mathbb{P}_{Z\sim D^n_{\text{ano}}}[\mathbb{P}(\hat{y}=0 \mid y=1) \leq \varepsilon_{\text{fn}}] \geq 1-\delta_{\text{fn}}.$$

Thus, the error rate when $f(x) \geq \hat{\tau}_{\text{fn}}$ and $f(x) \geq \hat{\tau}_{\text{fp}}$ satisfies

$$\mathbb{P}_{Z_{\text{ano}}\sim D^n_{\text{ano}}}[\varepsilon_{\text{ad}} \leq \varepsilon_{\text{fn}}] \geq 1-\delta_{\text{fn}}.$$

Therefore, if we make a certain prediction by Equation (10), we can bound the error rate as

$$\begin{aligned}\varepsilon_{\text{ad}} &= \varepsilon_{\text{fp}} \cdot \Pr(y=0) + \varepsilon_{\text{fn}} \cdot \Pr(y=1)\\ &\leq \max(\varepsilon_{\text{fp}}, \varepsilon_{\text{fn}}) \cdot \Pr(y=0) + \max(\varepsilon_{\text{fp}}, \varepsilon_{\text{fn}}) \cdot \Pr(y=1)\\ &= \max(\varepsilon_{\text{fp}}, \varepsilon_{\text{fn}}).\end{aligned}$$

The inequality holds with probability at least $1-(\delta_{\text{fp}}+\delta_{\text{fn}})$ due to the union bound. Thus, the claim follows. □

G PROOF OF THEOREM 5

Since Algorithm 1 returns $\hat{\tau}'_{\text{fp}}, \hat{\tau}'_{\text{fn}}$ and ε only when $\hat{\tau}'_{\text{fn}} \geq \hat{\tau}'_{\text{fp}}$, we first prove that the error rate of the anomaly detector is bounded by ε. Following the proof for Theorem 4, we have the guarantee that the FNR and FPR of the anomaly detector is bounded by the updated ε_{fn} and ε_{fp} in Algorithm 1, using $\hat{\tau}'_{\text{fn}}$ and $\hat{\tau}'_{\text{fp}}$ respectively. If we use a threshold $\tau' < \hat{\tau}'_{\text{fn}}$, the corresponding FNR, denoted as ε', obeys

$$\varepsilon' \leq \varepsilon_{\text{fn}} \leq \varepsilon. \tag{12}$$

This is because using a lower threshold corresponds to a lower quantile of the lower tail part for $\hat{y}=1$ distribution, and we have a smaller chance of making false negative prediction, i.e., classifying an anomaly as a normal point.

Similarly, if $\tau' > \hat{\tau}'_{\text{fp}}$, for the FPR, denoted as ε', we will have:

$$\varepsilon' \leq \varepsilon_{\text{fp}} \leq \varepsilon. \tag{13}$$

The same logic follows here; a higher threshold corresponds to a higher quantile of the upper tail of the distribution $x|\hat{y}=0$. Hence we have a smaller chance of making false positive prediction.

Since $\hat{\tau}_{\text{fp}} \leq \tau \leq \hat{\tau}_{\text{fn}}$, let $\tau' = \tau$. Based on Theorem 4, equation (12) and equation (13), the error rate of the anomaly detector ε_{ad} is bounded by ε:

$$\varepsilon_{\text{ad}} = \max(\varepsilon_{\text{fp}}, \varepsilon_{\text{fn}}) \leq \max(\varepsilon, \varepsilon) = \varepsilon.$$

Since we use $\delta_{\text{fn}}, \delta_{\text{fp}}$ to re-calculate $\hat{\tau}'_{\text{fp}}, \hat{\tau}'_{\text{fn}}$, the resulting δ can be taken as $\delta_{\text{fn}} + \delta_{\text{fp}}$ according to Theorem 4. Therefore, the claim follows. □

H PROOF OF LEMMA 1

We first prove that if $\hat{\tau}_{\text{fn}} \geq \hat{\tau}_{\text{fp}}$, then $\sum_{(x,y)\in Z_{\text{nm}}} \mathbb{1}(d(x) > \hat{\tau}_{\text{fn}}) \leq k^*_{\text{fp}}$ and $\sum_{(x,y)\in Z_{\text{ano}}} \mathbb{1}(d(x) < \hat{\tau}_{\text{fp}}) \leq k^*_{\text{fn}}$. To see this, we construct a false positive PAC prediction set using (2) and make a prediction using (7). Therefore, we have

$$\sum_{(x,y)\in Z_{\text{nm}}} \mathbb{1}(d(x) > \hat{\tau}_{\text{fp}}) \leq k^*_{\text{fp}}.$$

	FNR_{or}	FPR_{or}	FNR_{tt}	FPR_{tt}	FNR_{th}	FPR_{th}	ERR	ε
S-1	0.000	0.615	0.059	0.088	0.340	0.293	0.337	0.40
F-7	0.132	0.503	0.060	0.071	0.292	0.304	0.293	0.40
E-7	0.000	0.714	0.076	0.081	0.306	0.246	0.304	0.40
T-1	0.001	0.653	0.103	0.099	0.367	0.448	0.382	0.50
T-2	0.011	0.738	0.063	0.084	0.384	0.428	0.393	0.50
P-3	0.013	0.724	0.053	0.065	0.363	0.379	0.366	0.50

Table 7: Error rate with PAC-Wrap applied to the LSTM-based anomaly detector on the NASA data. First column is the corresponding channels. FNR_{tt} and FPR_{tt} satisfy the $\varepsilon = 0.1$ guarantees. After removing the ambiguity region, the FNR_{th}, FPR_{th}, and ERR satisfy the relaxed error constraints.

Since $\hat{\tau}_{fn} > \hat{\tau}_{fp}$, we find

$$\sum_{(x,y)\in Z_{nm}} \mathbb{1}(d(x) > \hat{\tau}_{fn}) \leq \sum_{(x,y)\in Z_{nm}} \mathbb{1}(d(x) > \hat{\tau}_{fp}) \leq k^*_{fp}.$$

Similarly, for the false negative PAC prediction set, we have

$$\sum_{(x,y)\in Z_{ano}} \mathbb{1}(d(x) < \hat{\tau}_{fp}) \leq \sum_{(x,y)\in Z_{ano}} \mathbb{1}(d(x) < \hat{\tau}_{fn}) \leq k^*_{fn}.$$

Next, we prove that if $\sum_{(x,y)\in Z_{nm}} \mathbb{1}(d(x) > \hat{\tau}_{fn}) < k^*_{fp}$ and $\sum_{(x,y)\in Z_{ano}} \mathbb{1}(d(x) < \hat{\tau}_{fp}) < k^*_{fn}$, then $\hat{\tau}_{fn} \geq \hat{\tau}_{fp}$. We argue by contradiction. Suppose that $\sum_{(x,y)\in Z_{nm}} \mathbb{1}(d(x) > \hat{\tau}_{fn}) < k^*_{fp}$, and $\sum_{(x,y)\in Z_{ano}} \mathbb{1}(d(x) < \hat{\tau}_{fp}) < k^*_{fn}$, but $\hat{\tau}_{fp} < \hat{\tau}_{fn}$. Then, for the false negative PAC prediction set, we should choose $\hat{\tau}_{fp}$ instead of $\hat{\tau}_{fn}$, since the identified $\hat{\tau}_{fn}$ should be the largest threshold satisfying $\hat{\varepsilon}_{fn}$ and $\hat{\tau}_{fp} > \hat{\tau}_{fn}$. This contradicts that $\hat{\tau}_{fn}$ is the chosen threshold. As a result, our assumption does not hold, and we have $\hat{\tau}_{fn} \geq \hat{\tau}_{fp}$.

In summary, $\hat{\tau}_{fn} \geq \hat{\tau}_{fp}$ if and only if $\sum_{(x,y)\in Z_{nm}} \mathbb{1}(d(x) > \hat{\tau}_{fn}) < k^*_{fp}$ and $\sum_{(x,y)\in Z_{ano}} \mathbb{1}(d(x) < \hat{\tau}_{fp}) < k^*_{fn}$. □

I TIME-SERIES EXPERIMENTS

We also experiment with two challenging time series anomaly detection datasets, the Server Machine Dataset [29], and NASA Telemetry Anomaly Detection [12], to illustrate the effectiveness of PAC-Wrap on sequential data. The NASA dataset consists of spacecraft telemetry data like radiation, temperature, and power from the Soil Moisture Active Passive satellite (SMAP), and the Curiosity Rover on Mars (MSL). In addition, it contains 193500 records for training and 501346 records for testing, of which around 10% are anomalies. SMD is a dataset collected from a large Internet company over five weeks, with 38 features such as CPU load, network usage, and memory usage. It contains a training set of 708405 records and a test set of 708420 records, among them 4.16% are anomalies. We split the original test set into a calibration set (20%) and a final test set (80%) for both SMD and NASA.

The detailed result for the NASA data is reported in Table 7. In the T-1, T-2, and P-3 channels, both FPR_{th} and FNR_{th} are guaranteed to be smaller than the relaxed error constraint, and the final error rate ERR is also below the required error constraint $\varepsilon = 0.5$. Wrapped around the original NASA anomaly detector, PAC-Wrap can reduce the gap between the FNR and FPR and thus has a more balanced performance.

Since there are more datapoints in the SMD dataset than in the NASA one, to approach the independence condition formally required by our guarantees, we consider the windows of the first 15 and 30 contiguous timesteps as data points for every 60, 120, and 240 timesteps. As shown in Table 8, $\varepsilon = 0.6, 0.8$ is the relaxed error constraint given the anomaly score distribution. For the baseline anomaly detector, FPR_{or}-s sometimes fail the $\varepsilon = 0.6$ guarantee for the 15-timestep settings. For all the 30-timestep settings, the original anomaly detectors violate the $\varepsilon = 0.8$ guarantee on either FPR_{or} or FNR_{or}. However, using PAC-Wrap as a wrapper, we ensure that both FNR_{th} and FPR_{th} fall below 0.6 and 0.8. The final error rates (ERR) are smaller than the maximum of FNR_{th} and FPR_{th}, which also empirically supports Theorem 4.

	FNR_{or}	FPR_{or}	FNR_{tt}	FPR_{tt}	FNR_{th}	FPR_{th}	ERR	ε
15-60	0.036	0.937	0.074	0.087	0.487	0.474	0.487	0.60
15-120	0.368	0.563	0.052	0.000	0.598	0.359	0.587	0.60
15-240	0.011	0.918	0.086	0.088	0.538	0.399	0.532	0.60
30-60	0.077	1.000	0.077	0.000	0.780	0.201	0.748	0.80
30-120	0.818	0.127	0.057	0.000	0.781	0.174	0.746	0.80
30-240	0.379	0.467	0.058	0.000	0.738	0.158	0.706	0.80

Table 8: Error rate with PAC-Wrap wrapped around the LSTM-based anomaly detector on the SMD data. First column is the corresponding combinations. FNR_{tt} and FPR_{tt} satisfy the $\varepsilon = 0.1$ guarantees. After removing the ambiguity region, the FNR_{th}, FPR_{th}, and ERR satisfy the relaxed error constraints.

J DISTRIBUTION SHIFT

To see how shifts in the anomaly distribution affect our guarantees, we generate data from three distributions in the following way:

$$X_{normal} \sim \mathcal{N}(\mu_{normal}, \sigma^2 I_p)$$
$$X_{anomalous} \sim \mathcal{N}(\mu_{anomalous}, \sigma^2 I_p)$$
$$X_{mixture} \sim \mathcal{N}(\gamma \cdot \mu_{normal} + (1-\gamma) \cdot \mu_{anomalous}, \sigma^2 I_p),$$

where $\gamma \in [0, 1]$ is a mixing ratio. We set $\mu_{normal} = [0, 0, 0, 0, 0]^\top$, $\mu_{anomalous} = [3, 3, 3, 3, 3]^\top$, and $\sigma = 2.0$. Then, we construct the training set by sampling 98,000 data points from $\mathcal{N}(\mu_{normal}, \sigma^2 I_p)$; we construct the calibration set by sampling 1,000 anomalies from $\mathcal{N}(\mu_{anomalous}, \sigma^2 I_p)$; we construct the testing set by sampling 1,000 data points from $\mathcal{N}(\gamma \cdot \mu_{normal} + (1-\gamma) \cdot \mu_{anomalous}, \sigma^2 I_p)$ with different γ-s. We set γ to $\{0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2\}$ and $\varepsilon = \delta = 0.05$. We run PAC-Wrap on the training, calibration, and testing sets. As shown in Figure 6, guarantees nearly hold when γ equals to 0.02, 0.04, and 0.06. However, when the mixing rate is too large, the guarantees might fail to hold.

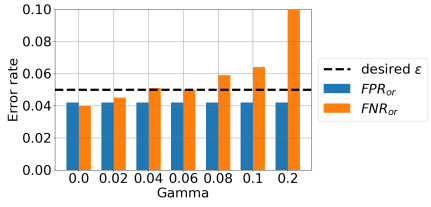


Figure 6: FPR and FNR after mixing different anomaly distributions