Fortifying Time Series: DTW-Certified Robust Anomaly Detection

Shijie Liu[1⋆], **Tansu Alpcan**[1], **Christopher Leckie**[2], **Sarah Erfani**[2]
[1]Department of Electrical and Electronic Engineering
University of Melbourne, Melbourne, Australia
[2]School of Computing and Information Systems
University of Melbourne, Melbourne, Australia
⋆shijie3@unimelb.edu.au

Abstract

Time-series anomaly detection is critical for ensuring safety in high-stakes applications, where robustness is a fundamental requirement rather than a mere performance metric. Addressing the vulnerability of these systems to adversarial manipulation is therefore essential. Existing defenses are largely heuristic or provide certified robustness only under ℓ_p-norm constraints, which are incompatible with time-series data. In particular, ℓ_p-norm fails to capture the intrinsic temporal structure in time series, causing small temporal distortions to significantly alter the ℓ_p-norm measures. Instead, the similarity metric *Dynamic Time Warping* (DTW) is more suitable and widely adopted in the time-series domain, as DTW accounts for temporal alignment and remains robust to temporal variations. To date, however, there has been no certifiable robustness result in this metric that provides guarantees. In this work, we introduce the first *DTW-certified robust defense* in time-series anomaly detection by adapting the randomized smoothing paradigm. We develop this certificate by bridging the ℓ_p-norm to DTW distance through a lower-bound transformation. Extensive experiments across various datasets and models validate the effectiveness and practicality of our theoretical approach. Results demonstrate significantly improved performance, e.g., up to 18.7% in F1-score under DTW-based adversarial attacks compared to traditional certified models.

1 Introduction

In recent years, significant research has advanced the study of adversarial attacks and certified defenses for machine learning systems. Despite the considerable progress in adversarial robustness across various domains [42, 2, 45, 9, 12, 3], robustness in *time-series anomaly detection* remains comparatively underexplored. As a core component of many safety-critical systems—including healthcare [25, 46, 21], finance [41, 22, 64], and mobile networks [56, 70, 35]— anomaly detectors are essential for identifying abnormal behavior in preventing failures or hazards. Robustness in this context is not merely a model performance concern but a core requirement for operational reliability. Recent work has revealed that time-series anomaly detectors are susceptible to adversarial attacks tailored to the characteristics of time-series data [6, 5], underscoring the urgent need for ensuring robustness in this domain.

Adversaries can manipulate detection outcomes by introducing subtle yet strategically crafted perturbations into anomaly time-series data to evade detection [61, 29, 74, 69]. Traditional adversarial threat models typically restrict perturbations to bounded ℓ_p-norms, widely effective in image [1, 31] and text domains [63, 76, 57] due to the alignment with semantic preservation in such data types. However, time-series data exhibit an inherent *temporal structure* that challenges the assumption of

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

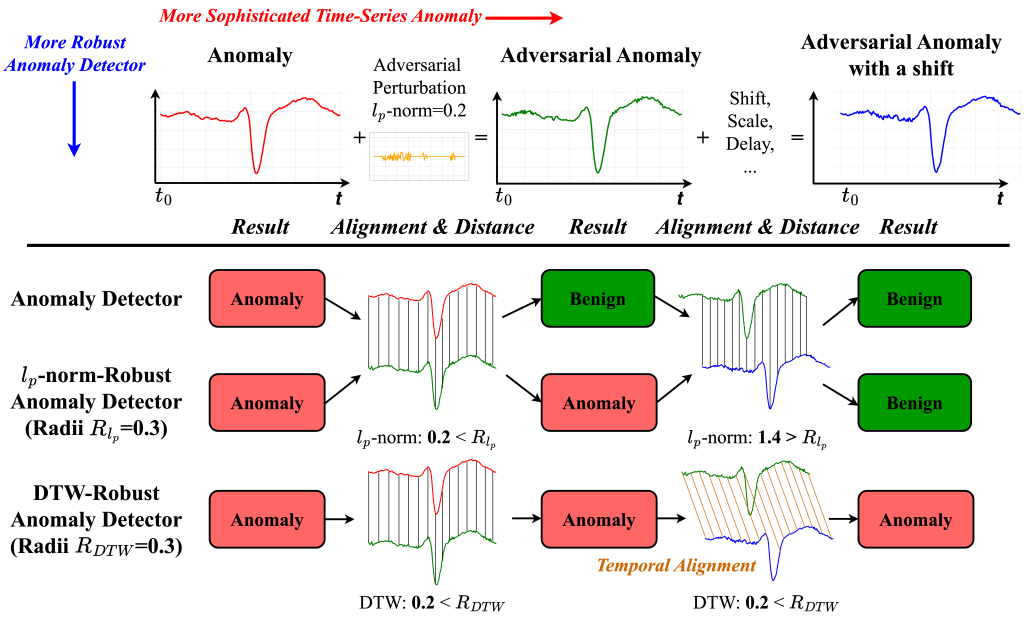


Figure 1: Comparison of standard, ℓ_p-norm-robust, and DTW-robust anomaly detectors under adversarial perturbations. DTW facilitates optimal temporal alignment, offering a more meaningful similarity measure for time-series data, thereby ensuring more comprehensive robustness guarantees against adversarial examples.

ℓ_p-norms. As illustrated in Figure 1, small temporal transformations such as shifts or rescaling can significantly inflate ℓ_p-norm distance (e.g., from 0.2 to 1.4), despite there being no change to the underlying semantics (e.g., green and blue represent the identical time series). This mismatch renders ℓ_p-norms inadequate for measuring meaningful similarity in time-series data, limiting their utility for robustness measurement.

Addressing this limitation, we advocate for the use of *Dynamic Time Warping (DTW)* distance, a commonly used similarity metric specifically designed for time-series data. As shown in the bottom right of Figure 1, DTW accommodates *temporal alignments*, which effectively handles temporal variations such as shifts, stretching, and compression. This alignment flexibility preserves structural similarities better than ℓ_p-norms, consistently demonstrating superior performance for diverse time-series tasks [24, 37, 19]. As a result, a DTW-robust detector exhibits stable distance measurements under temporal variations (e.g., consistently 0.2 as in Figure 1), offering more reliable robustness guarantees compared to ℓ_p-norm. The need for developing defenses under the DTW distance is further emphasized by recent demonstrations of DTW-based adversarial attacks [6, 5], for which no certified defenses currently exist.

In response to adversarial attacks, various defensive strategies have been proposed [38, 11, 20]. Although these empirical defenses provide some resilience, adaptive attackers can often bypass them [10, 65, 75]. *Certified defenses* [32, 18, 55], in contrast, guarantee theoretical robustness against worst-case adversarial scenarios, making them particularly appealing for safety-critical applications. While significant progress has been made in certified robustness under ℓ_p-norm constraints [32, 55, 18, 54, 33, 72], its adaptation to time-series data—under the proper DTW constraints—remains unexplored.

In this paper, we propose the first *DTW-certified robustness* framework in time-series anomaly detection by adapting the randomized smoothing approach [18]. Our approach establishes a novel certification method by bridging ℓ_p-norm guarantees to the DTW distance through a lower-bound transformation. By leveraging the Keogh Lower Bound [28], we are able to derive a closed-form expression of the DTW-certified radius for a smoothed model. The resulting framework is model-agnostic and readily applicable to any pre-trained anomaly detector. Extensive evaluations on real-world datasets and a variety of detection architectures highlight the broad applicability and effectiveness of our method, demonstrating clear advantages over traditional ℓ_p-norm-certified defenses,

e.g., under DTW-based adversarial attacks, our method achieves up to an 18.7% improvement in F1-score.

Our key contributions include:

- We introduce the first theoretical framework that provides certified robustness in DTW distance, addressing an essential yet unexplored gap in time-series anomaly detection.
- We present a generalizable defense mechanism that seamlessly integrates DTW certification with any anomaly detection models, significantly enhancing robustness in practice.
- We provide comprehensive experimental evaluations demonstrating the practical efficacy of our DTW-certified robustness approach across diverse scenarios.

2 Related Work

Time-series Anomaly Detection Time-series data, consisting of data points sequentially indexed over time, is prevalent across various domains. Detecting anomalies within time-series data is of significant importance [40, 49, 16, 7, 25, 46, 21], as anomalies often indicate novel, unexpected, or potentially critical events. Recent advances in deep learning have significantly improved detection by enabling models to capture complex temporal and inter-metric dependencies. Modern *deep anomaly detectors* [62, 13, 36] have shown strong performance across a range of time-series tasks. However, they also share the same vulnerabilities to adversarial attacks as other machine learning models.

Adversarial Attacks Adversarial attacks refer to deliberate perturbations introduced into input data to intentionally mislead machine learning models into making incorrect predictions. Typically, these perturbations are minimal in terms of the ℓ_p-norm, ensuring the semantic consistency and being imperceptible to humans. Such vulnerabilities have been widely demonstrated across various deep learning models [1, 31, 63, 76, 57], including anomaly detection tasks [61, 29, 74, 69].

However, the ℓ_p-norm is inadequate for measuring differences in *time-series data*, as it fails to account for the underlying temporal structure. Recent studies [6, 5] have addressed these issues by adopting the DTW distance, a widely recognized measure suitable for time-series analysis, to construct adversarial examples. These studies highlight that DTW-based adversarial attacks are more effective, as the set of permissible perturbations under DTW forms a superset of those constrained by an equivalent ℓ_p-norm. Additionally, they demonstrate that the defensive strategies designed to counter ℓ_p-norm adversarial attacks exhibit limited effectiveness against DTW-based attacks [39], highlighting the need for dedicated defenses under the DTW threat model.

Certified Robustness Prior defenses such as adversarial training [39], defensive distillation [44], and data purification [66] offer empirical robustness, but are often circumvented by adaptive adversaries [10, 65, 75]. In contrast, *certified defenses* have gained significant attention for providing formal, provable guarantees against all possible attacks within a perturbation bound [32, 18, 53]. In the context of time-series anomaly detection, certified defenses against ℓ_p-norm attacks have been initially considered in [23, 8], through direct application of randomized smoothing [18]. However, as discussed, the resulting ℓ_p-norm certificate is inadequate for time-series data and remains vulnerable to DTW-based attacks. Existing defences [6] against DTW-based attacks have been limited to empirical without any certification. This work introduces the first certified defense against DTW-based adversarial attacks.

3 DTW-Certified Defense in Time-Series Anomaly Detection

3.1 Problem Setup

Time-Series Anomaly Detector We define the space of time-series signals as $\mathcal{X} = \mathbb{R}^{L\times C}$, where L represents the signal length and C denotes the number of channels. Following the common framework for time-series anomaly detection [71, 67, 51], we consider a detector $d : \mathbb{R}^{T\times C} \to \mathcal{Y} = \{0, 1\}$ that operates on a sliding window of size $T \leq L$. Given an input sequence $x \in \mathbb{R}^{T\times C}$, the detector computes an *anomaly score* $f(x) \in \mathbb{R}$, which quantifies the likelihood of x being an anomaly, and makes the detection decision via comparing $f(x)$ against the anomaly threshold γ:

$$d(x) = \begin{cases} 1, & f(x) > \gamma \ , \\ 0, & f(x) \leq \gamma \ , \end{cases} \tag{1}$$

where $y = 1$ indicates an anomalous instance, and $y = 0$ denotes a benign instance.

Distance Metrics The difference between two time-series x and x' can be naively measured by the ℓ_p*-norm distance* as

$$\|x - x'\|_p = \left(\sum_{i=1}^{T} |x_i - x'_i|^p \right)^{1/p}, \tag{2}$$

where x_i, x'_i represent the i-th element in x, x'. However, such a measurement fails to capture the temporal structure of time-series data. Therefore, we consider the *Dynamic Time Warping (DTW) distance*, which resolves the issues by finding the optimal temporal alignment that minimizes the total distance between aligned time-series. Formally, the DTW distance of norm order p is defined as:

$$DTW_p(x, x') = \min_{\pi \in \mathcal{A}(x,x')} \left(\sum_{(i,j)\in\pi} |x_i - x'_j|^p \right)^{1/p} \tag{3}$$

where π represents an *alignment path* of length T as a sequence of T index pairs $[(i_1, j_1), \cdots, (i_T, j_T)]$ and $\mathcal{A}(x, x')$ is the set of all admissible paths. An admissible path should satisfy the following conditions: 1) Matched ends, as $\pi_1 = (1, 1)$ and $\pi_T = (T, T)$, and 2) Monotonically increasing and each time series index should appear at least once, as $i_{k-1} \leq i_k \leq i_{k-1} + 1$ and $j_{k-1} \leq j_k \leq j_{k-1} + 1$. We adopt $p = 2$ as the default norm for DTW in the main text for clarity of exposition; however, the proposed approach generalizes readily to arbitrary norm orders p with minimal modification as detailed in Appendix C.

Threat Model We assume a strong adversary with white-box access to the anomaly detector d, meaning the attacker has full knowledge of the detector and unlimited computational power. Given an input x classified by the detector as $y = f(x)$, the attacker seeks an alternative input x' to perform either an *evasion attack*—suppressing the detection of an actual anomaly, or an *availability attack*—inducing a false alarm on benign input, such that $d(x') \neq d(x)$. To preserve the semantics of the original anomaly x, the perturbation in x' must be constrained within a DTW distance e as $DTW(x, x') < e$.

Certified Defense Goal The anomaly detector d is said to provide certified defense at input x of DTW radius e, if there exist no $x' \in \{x' \mid DTW(x, x') < e\}$ such that $d(x') \neq d(x)$ with probability at least $1 - \alpha$.

3.2 Theoretical Analysis of DTW-Certified Robustness

In this section, we first review the core components of our approach: the randomized smoothing framework [18, 15] and the DTW lower bound [28]. We then present Lemma 3.2, which establishes a formal link between ℓ_p-norm distances and the DTW lower bound. Building on this connection, we introduce the main theoretical result Theorem 3.3, which derives a DTW robustness certificate from a smoothed model via the Keogh Lower Bound.

ℓ_p-norm Certificate via Randomized Smoothing Randomized smoothing [17] constructs a *smoothed function* by taking the Gaussian expectation of a base function f (we defer the details in Appendix A). However, in time-series anomaly detection, the base function $f(x)$—which outputs an anomaly score for a time series x—is typically unbounded and may exhibit high variance. As a result, estimating the Gaussian mean can lead to loose and unreliable robustness bounds.

To address this, we adopt the *percentile smoothing* approach [14], which bounds the *p-th percentiles* of the base function outputs instead of the mean. Such a smoothing method is more robust to outliers and variance in the output distribution. We construct the *smoothed anomaly score function* $h_p(x) : \mathcal{X} \to \mathbb{R}$ of the anomaly score function f, as

$$h_p(x) = \sup\{u \in \mathbb{R} \mid \mathbb{P}_{\eta \sim N(0,\sigma^2 I)}[f(x + \eta) \leq u] \leq p\}. \tag{4}$$

The h_p does not admit a closed form, its value can be bounded by Monte Carlo sampling as outlined in Section 3.3. With the percentile smoothed function, the anomaly score $h_p(x')$ of the adversarial input x' can be certifiably bounded by $h(x)$, as

Lemma 3.1. *A percentile smoothed function h_p can be bounded as*

$$h_{\underline{p}}(x) \;\leq\; h_p(x') \;\leq\; h_{\overline{p}}(x) \quad \forall x' \in \{x' \mid \|x - x'\|_2 \leq r\}, \tag{5}$$

where $\underline{p} = \Phi(\Phi^{-1}(p) - \frac{r}{\sigma})$ and $\overline{p} = \Phi(\Phi^{-1}(p) + \frac{r}{\sigma})$, with Φ being the standard Gaussian CDF.

Lower Bound of DTW The exact computation of DTW is typically expensive and slow, i.e., quadratic time and space complexity. To address this, various lower bounds have been proposed to approximate DTW efficiently. One of the widely used bounds is the Keogh Lower Bound [28, 47] $LB\_Keogh(x, x')$, which is calculated by defining two new time series, upper U and lower L envelopes. For each time step i and channel k, the envelopes are defined as:

$$\begin{aligned} U_{i,k} &= \max(x_{i-w,k} : x_{i+w,k}) \\ L_{i,k} &= \min(x_{i-w,k} : x_{i+w,k}) \end{aligned} \tag{6}$$

where the $w : 1 \leq w \leq T$ is the DTW wrapping window size (Sakoe–Chiba band) [52] that constrains only x_i and x'_j within the window can be aligned. The $LB\_Keogh(x, x')$ is calculated as

$$LB\_Keogh_p(x, x') = \sqrt[p]{\sum_{i=1}^{T}\sum_{k=1}^{N} \begin{cases} (x'_{i,k} - U_{i,k})^p & \text{if } x'_{i,k} > U_{i,k}, \\ (x'_{i,k} - L_{i,k})^p & \text{if } x'_{i,k} < L_{i,k}, \\ 0 & \text{otherwise.} \end{cases}} \tag{7}$$

In summary, the lower bound is calculated as the sum of ℓ_p-norm distances to the envelope of points in x' that are outside the envelope of x.

DTW-Certificate In the following, we present the theoretical foundation for deriving DTW-certified robustness, offering a robustness measure that is better aligned with the temporal nature of time-series data. By leveraging the percentile-smoothed function and the DTW lower bound, we introduce a lemma that establishes a connection between the ℓ_p-norm certificate and a robustness certificate in DTW distance through a lower-bound transformation.

Lemma 3.2. *Suppose the certification of a smoothed function h holds for data x as $a \leq h(x') \leq b, \forall x' \in \{x' \mid \|x' - x\| \leq r\}$. Then, the certification $a \leq h(x') \leq b, \forall x' \in \{x' \mid \mathrm{DTW}(x, x') \leq e\}$ also holds, where $LB(x, x')$ is a strict lower bound of $\mathrm{DTW}(x, x')$ and*

$$e = \inf\{LB(x, x') \mid \|x - x'\| > r\}. \tag{8}$$

Proof. Assume for the sake of contradiction that the chosen x' does not lie in the l_2-ball. Then we have $\|x' - x\| > r$. Since x' is outside the ball, by the definition of e we know that $LB(x, x') \geq e$. This contradicts $LB(x, x') < DTW(x, x') \leq e$ by the definition of the set $\{x' \mid \mathrm{DTW}(x, x') \leq e\}$ and LB is a strict lower bound of DTW. Thus, we conclude that any point x' with $DTW(x, x') \leq e$ must satisfy $\|x' - x\| \leq r$, where the certification holds. □

Building upon Lemma 3.2, we present the theorem that establishes DTW-certified robustness for the smoothed function. Formally, we define the anomaly score function $f : \mathcal{X} \to \mathbb{R}$ of a time-series anomaly detector $d : \mathcal{X} \to \mathcal{Y}$, and construct a percentile smoothed version of f, denoted as $h_p : \mathcal{X} \to \mathbb{R}$, which serves as the new anomaly score function of d. The DTW-certified robustness of d can then be derived through the following theorem.

Theorem 3.3 (Robustness Certification for Time-series Anomaly Detection)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be any deterministic or random function, and $\eta \sim \mathcal{N}(0, \sigma^2 I)$. Let the percentile smoothed function $h_p : \mathcal{X} \to \mathbb{R}$ be defined as in Equation* (4)*. Suppose the anomaly score threshold is γ, and the following is satisfied for a testing input x*

$$\begin{cases} h_{\underline{p}}(x) > \gamma, & \text{if } h_p(x) > \gamma \ , \\ h_{\overline{p}}(x) \leq \gamma, & \text{if } h_p(x) \leq \gamma \ . \end{cases} \tag{9}$$

Then $d(x') = d(x)$ is guaranteed to hold for all $\{x' : DTW(x, x') \leq e\}$, where

$$e = \begin{cases} 0, & \text{if } r \leq R, \\ \sqrt{M^2 + r^2 - R^2} \ - \ M, & \text{if } r > R, \end{cases} \tag{10}$$

with

$$\Delta_i = \max\big(U_i - x_i,\ x_i - L_i\big), \quad R = \sqrt{\sum_{i=1}^{n} \|\Delta_i\|^2}, \quad M = \max_{1 \leq i \leq n} ||\Delta_i||,$$

$$r = \begin{cases} \sigma\big(\Phi^{-1}(p) - \Phi^{-1}(\underline{p})\big), & \text{if } h_p(x) > \gamma \ , \\ \sigma\big(\Phi^{-1}(\overline{p}) - \Phi^{-1}(p)\big), & \text{if } h_p(x) \leq \gamma \ . \end{cases}$$

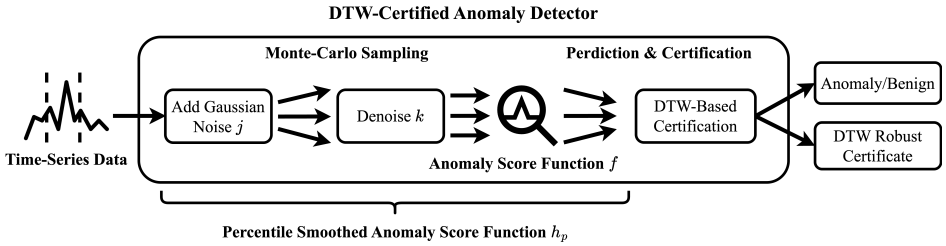


Figure 2: Construct any anomaly detector with anomaly score function f as a DTW-certified detector.

where U and L are envelopes in the Keogh Lower Bound with wrapping window size w as specified in Equation (6)*.*

Proof. We provide a sketch of the proof below due to space constraints, and the complete proof is available in Appendix B.

We begin by showing that $d(x') = d(x)$ holds for all x' satisfying $\|x - x'\| \leq r$ using Lemma 3.1. The radius r can be solved as $r = \sigma\big(\Phi^{-1}(p) - \Phi^{-1}(\underline{p})\big)$ or $r = \sigma\big(\Phi^{-1}(\overline{p}) - \Phi^{-1}(p)\big)$ depending on the classification outcome. Next, we invoke Lemma 3.2 to translate the certificate from ℓ_p-norm r to DTW distance e, defined as $e = \inf\{LB(x, x') \mid \|x - x'\| > r\}$, where $LB(x, x')$ is a strict lower bound of the DTW distance. In the final step, we instantiate $LB(x, x')$ with the Keogh Lower Bound, which satisfies the strictness condition for all $w > 0$ and $x' \neq x$, and derive the corresponding expression for e by exploiting the structural properties of the upper and lower envelopes U and L. □

3.3 DTW-Certified Defense Implementation

Construct smoothed detector Given an anomaly detector with an anomaly score function f, we construct the percentile-smoothed anomaly score function h_p by the definition of Equation (4) following the process as shown in Figure 2. Specifically, the smoothed function is composed as $h_p = j \circ k \circ f$, where the smoothing noise $\eta \sim \mathcal{N}(0, \sigma^2 I)$ injection layer j generates multiple Gaussian-perturbed inputs $x + \eta$ from the original time-series x, and the denoising layer k reduces noise variance to improve score concentration. This denoising process does not compromise the certification guarantee, as randomized smoothing is valid for any downstream pipeline [54]. For each testing input x, the DTW-certified anomaly detector outputs a binary decision based on anomaly score $h_p(x)$ and computes the corresponding certified DTW radius e by the Theorem 3.3. This method does not require modifications to the training process and can be readily applied to pre-trained models.

Bound $h_{\overline{p}}(x)$ and $h_{\underline{p}}(x)$ We utilize Monte-Carlo sampling to estimate and bound the upper and lower percentiles $h_{\overline{p}}(x)$ and $h_{\underline{p}}(x)$, following a similar approach as in [18, 15]. Given n i.i.d. Gaussian noise samples $\{\mu_1, \cdots, \mu_n\}$, we compute anomaly scores $X_i = f(x + \mu_i)$ and sort them to obtain the empirical order statistics $-\infty = K_0 \leq K_1, \cdots \leq K_n \leq K_{n+1} = \infty$. We aim to identify K_{q^u} and K_{q^l} such that $\Pr[K_{q^u} > h_{\overline{p}}(x)] > 1 - \alpha$ and $\Pr[K_{q^l} < h_{\underline{p}}(x)] > 1 - \alpha$ for a confidence level of $1 - \alpha$ (we set $\alpha = 1e{-}3$ in the experiments). The corresponding probabilities are evaluated using the binomial distribution as

$$\Pr[K_{q^u} > h_{\overline{p}}(x)] = \sum_{i=1}^{j=q^u} \binom{n}{i} (\bar{p})^i \, (1 - \bar{p})^{n-i} \ . \tag{11}$$

A similar formula applies for the lower bound. We use binary search to identify the smallest q^u and largest q^l that satisfy the required confidence bounds. In general, increasing the number of samples improves the estimation accuracy of the certified radius, and greater consensus aggregated predictions indicate a stronger certification.

4 Experiments

In our experiments, we evaluate the general applicability of the DTW-certified defense across a range of anomaly detection models and time-series datasets. We demonstrate improved robustness

compared to ℓ_p-norm certified defenses and provide ablation studies to analyze the trade-off between detection performance and certified robustness.

Settings Our empirical evaluation of the DTW-certified defense spans seven widely used benchmark datasets, including SMAP [48], MSL [27], SML [60], NIPS-TS-SWAN, NIPS-TS-CREDITCARD, NIPS-TS-WATER [30], UCR-1 ane UCR-2 [68], encompassing both univariate and multivariate time-series data. Detailed descriptions and dataset statistics are provided in Appendix D. To ensure broad applicability, we evaluate our approach using three state-of-the-art anomaly detection models: COUTA [71], TimesNet [67], and DeepSVDDTS [51]. The effectiveness is further validated through comparison with ℓ_p-norm certified defense [18] under DTW-based adversarial attack [6].

We use the following default hyperparameters across all experiments unless otherwise specified: sequence length $T = 50$, DTW wrapping window size $w = 4$, number of noisy samples $n = 1,000$, smoothing noise level $\sigma = 0.5$ in $\mathcal{N}(0, \sigma^2 I)$, and percentile $p = 0.5$ in the percentile-smoothed function h_p. Additional ablation studies on the hyperparameters are available in Appendix F.

All experiments are implemented using PyTorch and executed on a Linux server equipped with Intel(R) Xeon(R) Gold 6326 CPUs and NVIDIA A100 GPUs with 80 GB of memory.

Evaluation Metrics For evaluating the *detection performance*, we report the point-adjusted **F1-score** and Area Under the Receiver Operating Characteristic Curve (**ROC AUC**) following the common practice in the domain of time-series anomaly detection [4, 50, 58, 59, 34, 71, 51, 67].

To evaluate *certified robustness*, we report the **mean**, **maximum**, and **standard deviation (std.)** of the certified radii computed for all test instances. Additionally, we report the **certified proportion (prop.)**, defined as the fraction of test inputs with a non-zero certified radius e.

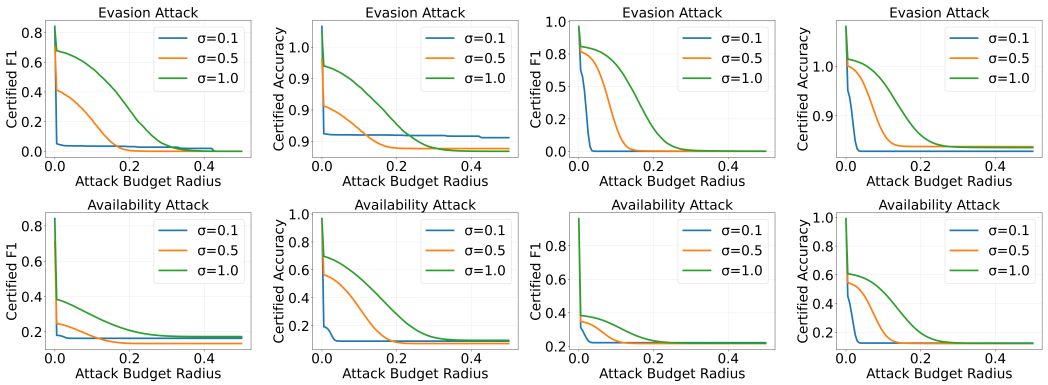
Following the notion of *certified accuracy* from the certified robustness literature [32, 18, 53], which is defined as the proportion of instances for which the model guarantees correct predictions within a specified *attack budget* as radius t. We extend this evaluation to the confusion matrix components by considering worst-case adversarial scenarios. Specifically, for evasion attacks, we define Certified True Positives (TP) as the count of true positive instances for which the model is provably robust within a DTW radius of t as $\sum_{i=1}^{N} \mathbb{I}\{\forall x' \,:\, DTW(x_i, x') \leq t :\, f(x') = 1, y_i = 1\}$. Similarly, for availability attacks, we define the Certified True Negatives (TN) as the number of benign instances that remain correctly classified under all perturbations within the DTW radius t as $\sum_{i=1}^{N} \mathbb{I}\{\forall x' \,:\, DTW(x_i, x') \leq t :\, f(x') = 0, y_i = 0\}$. We construct the corresponding certified confusion matrix as detailed in Appendix E, and derive the certified metrics, **certified accuracy** and **certified F1-score** that represent the guaranteed performance under bounded attack.

4.1 Results

The DTW-certified defense is broadly applicable, though the certified robustness performance varies across datasets and models. We evaluate the applicability of our DTW-certified defense across benchmark datasets and anomaly detection models. As shown in Table 1, our approach generally achieves strong certified robustness with minimal trade-offs in detection performance. For instance, on the NIPS-TS-WATER using DeepSVDDTS, our method certifies 99.46% of test inputs with an average certified robust DTW-radius of 0.189, without any degradation in F1-score and ROC AUC. Additionally, DeepSVDDTS often achieves the strongest performance, which we attribute to its superior handling of noisy data. However, we observe weaker robustness on certain datasets, such as SMAP and NIPS-TS-SWAN, due to their high channel dimensionality and greater data variance, which reduces the tightness of the lower-bound estimation and thus limits certifiable robustness.

Figure 3 presents the certified F1-score and certified accuracy of the COUTA model on the MSL and SMAP datasets under evasion and availability attacks. The x-axis denotes the attack budget radius t, while the y-axis shows the corresponding certified metrics. These curves represent lower bounds on model performance under the worst-case adversarial perturbations constrained by $DTW(x, x') \leq t$, as guaranteed by our DTW-certified defense. With appropriately chosen hyperparameters (e.g., $\sigma = 1.0$), the defense exhibits strong certified robustness. For instance, on the SMAP dataset, it maintains a certified F1-score of approximately 0.5 under an evasion attack with a budget of $t = 0.2$.

Improved performance over ℓ_p-norm certified defense. Table 2 evaluates the effectiveness of our DTW-certified defense under strong DTW-based adversarial attacks [6]. The adversary is granted a



(a) Certified metrics evaluated on the MSL dataset. (b) Certified metrics evaluated on the SMAP dataset.

Figure 3: Certified accuracy and certified F1-score as functions of the DTW perturbation threshold $t \in [0.0, 0.5]$ under evasion or availability attack. Results are reported for the COUTA model on the MSL (a) and SMAP (b) datasets across varying values of the hyperparameter σ.

Dataset	Model	Base Model		DTW-Certified Defense Model					
		Detection Performance		Detection Performance		Certified Robustness (Radii Statistics)			
		F1-score	ROC AUC	F1-score	ROC AUC	Radii Mean	Radii Max	Radii Std.	Certified Prop.
SMAP	COUTA	0.794	0.958	0.961	0.998	0.037	0.816	0.043	51.06%
	TimesNet	0.783	0.929	0.910	0.992	0.039	1.001	0.044	51.03%
	DeepSVDDTS	0.694	**0.861**	0.719	**0.964**	**0.052**	0.626	0.055	**53.55%**
SMD	COUTA	0.675	0.956	0.624	0.966	0.083	0.386	0.060	77.99%
	TimesNet	0.827	0.995	0.755	0.964	0.032	0.267	0.036	56.69%
	DeepSVDDTS	0.725	**0.958**	0.733	**0.957**	**0.255**	2.318	0.150	**93.94%**
MSL	COUTA	0.911	0.993	0.706	0.966	0.057	0.534	0.062	55.92%
	TimesNet	0.744	0.956	0.910	0.993	0.056	0.262	0.056	62.18%
	DeepSVDDTS	0.825	**0.973**	0.816	**0.979**	**0.123**	2.347	0.139	**73.89%**
NIPS-TS-SWAN	COUTA	0.780	0.788	0.738	0.710	0.022	0.574	0.064	14.52%
	TimesNet	0.770	0.906	0.773	0.866	0.013	0.451	0.080	10.56%
	DeepSVDDTS	0.740	**0.827**	0.744	**0.830**	**0.227**	2.242	0.196	**71.05%**
NIPS-TS-CREDITCARD	COUTA	0.192	0.900	0.003	0.300	**0.232**	0.450	0.046	**99.86%**
	TimesNet	0.422	**0.942**	0.450	**0.846**	0.028	0.262	0.035	51.75%
	DeepSVDDTS	0.132	0.772	0.119	0.719	0.105	0.817	0.056	91.55%
NIPS-TS-WATER	COUTA	0.515	0.537	0.598	0.989	0.070	0.359	0.034	95.18%
	TimesNet	0.778	0.997	0.550	0.974	0.120	0.279	0.032	99.07%
	DeepSVDDTS	0.512	**0.764**	0.513	**0.907**	**0.189**	0.540	0.039	**99.46%**
UCR-1	COUTA	0.672	0.986	0.949	0.999	0.177	0.413	0.124	69.43%
	TimesNet	0.886	0.996	0.845	0.995	0.011	0.207	0.024	25.30%
	DeepSVDDTS	0.813	**0.994**	0.984	**1.000**	**0.351**	0.758	0.220	**74.89%**
UCR-2	COUTA	0.886	0.998	0.842	0.939	0.022	0.190	0.033	38.84%
	TimesNet	0.984	**1.000**	0.982	**1.000**	**0.036**	0.256	0.042	**52.15%**
	DeepSVDDTS	0.118	0.900	0.306	0.970	0.033	0.218	0.043	46.02%

Table 1: Detection performance and certified robustness of the DTW-certified defense across various datasets and models with hyperparameter $\sigma = 0.5$. The results show minimal degradation in detection performance while consistently achieving meaningful DTW-certified robustness.

generous attack budget of $e_{att} = 1.0$, which exceeds the average certified radius of 0.5 measured by both ℓ_p-norm and proposed DTW-certified defenses across datasets for model COUTA. As shown in Table 2, the DTW-based adversarial attack is highly effective against undefended models, causing significant drops in F1-score and ROC AUC (e.g., a 60.6% drop in F1-score on SMD and 89.7% on UCR-1). While the ℓ_p-norm certified defense offers partial resilience, it fails to provide consistent protection, especially on datasets with strong temporal distortions under attacks (e.g., SMD and NIPS-TS-WATER). In contrast, our DTW-certified defense consistently outperforms both baselines under attack, yielding substantially higher F1-scores and AUCs. For example, on MSL and UCR-1, our method improves the F1-score under attack by 11.2% and 18.7%, respectively, compared to the ℓ_p-norm certified defense. These results affirm that robustness guarantees aligned with DTW—rather than ℓ_p-norm—are essential for effective defense in time-series anomaly detection.

Trade-off between detection performance and certified robustness. We investigate the trade-off between detection performance and certified robustness by varying the hyperparameter σ, which

Dataset	Unattacked		Under DTW-based Adversarial Attack with Attack Budget DTW $e_{att} = 1.0$					
	Base Model		Undefended Base Model		ℓ_p-norm Certified Defense		DTW-Certified Defense	
	F1-score	ROC AUC	F1-score	ROC AUC	F1-score	ROC AUC	F1-score	ROC AUC
MSL	0.896	0.992	0.694	0.938	0.672	0.943	**0.784**	**0.966**
SMD	0.575	0.938	0.253	0.698	0.392	0.741	**0.464**	**0.838**
NIPS-TS-WATER	0.516	0.689	0.240	0.665	0.423	0.797	**0.525**	**0.927**
UCR-1	0.682	0.987	0.084	0.846	0.761	0.893	**0.948**	**0.998**

Table 2: Detection performance under DTW-based adversarial attacks, evaluated using the COUTA model across multiple datasets. Comparisons are made among the undefended base model, the ℓ_p-norm certified defense, and the proposed DTW-certified defense.

Dataset	σ	DTW-Certified Defense					
		Detection Performance		Certified Robustness (Radii Statistics)			
		F1-score	ROC AUC	Radii Mean	Radii Max	Radii Std.	Certified Prop.
SMAP	0.1	0.956	0.997	0.007	0.320	0.010	41.13%
	0.5	0.961	0.998	0.037	0.816	0.043	51.06%
	1.0	**0.961**	**0.998**	0.080	1.051	0.082	58.02%
	2.0	0.913	0.958	**0.202**	1.571	0.165	**72.24%**
SMD	0.1	0.504	0.881	0.025	0.190	0.028	52.42%
	0.5	0.624	0.966	0.083	0.386	0.060	77.99%
	1.0	**0.673**	**0.977**	0.121	0.511	0.084	83.28%
	2.0	0.315	0.822	**0.456**	1.969	0.285	**94.66%**
MSL	0.1	**0.841**	0.982	0.003	0.426	0.020	11.13%
	0.5	0.706	0.966	0.057	0.534	0.062	55.92%
	1.0	0.830	**0.984**	0.108	0.457	0.098	68.27%
	2.0	0.739	0.914	**0.294**	1.457	0.196	**87.72%**
NIPS-TS-WATER	0.1	0.515	0.775	0.192	0.473	0.038	**99.42%**
	0.5	**0.598**	**0.989**	0.070	0.359	0.034	95.18%
	1.0	0.555	0.986	0.161	0.430	0.061	98.79%
	2.0	0.462	0.983	**0.261**	0.711	0.114	98.54%
UCR-1	0.1	0.871	0.996	0.023	0.112	0.027	50.63%
	0.5	**0.949**	**0.999**	0.177	0.413	0.124	69.43%
	1.0	0.919	0.984	0.309	0.692	0.196	75.36%
	2.0	0.821	0.973	**0.541**	1.209	0.300	**82.67%**

Table 3: Detection performance and certified robustness results evaluated under varying $\sigma = \{0.1, 0.5, 1.0, 2.0\}$ using COUTA. Higher σ generally yield improved certified robustness (Mean, Max, Prop.), but could at the expense of reduced detection performance (F1-socre, ROC AUC).

controls the magnitude of Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ used in the smoothing process. Notably, under a moderate setting ($\sigma = 0.5$), many configurations exhibit improved detection performance compared to the base model, as shown in datasets SMAP and UCR-1 in Table 1. This observation is consistent with prior work [18, 73, 43, 26], where smoothing is shown to enhance generalization by stabilizing decision boundaries. As illustrated in Table 3, increasing σ generally improves certified robustness, as reflected in larger average certified radii and a higher proportion of certified inputs. However, overly large values (e.g., $\sigma = 2.0$) often degrade detection performance, including both F1-score and ROC AUC. Therefore, the choice of σ should be tuned carefully, considering both the model architecture and dataset characteristics.

5 Conclusion and Limitations

We present the first certified defense for time-series anomaly detection under the Dynamic Time Warping (DTW) distance—a metric well-suited for capturing temporal structure in time-series data. By adapting randomized smoothing and leveraging the Keogh lower bound, we derive a DTW-certified radius that provides formal robustness guarantees. This method is model-agnostic across diverse datasets and architectures. Empirical results demonstrate that it consistently delivers strong DTW-certified robustness while maintaining strong detection performance.

The Monte Carlo sampling process introduces testing-time overhead, which future work may address by exploring more efficient sampling strategies or adaptive noise injection methods. Additionally, tightening the DTW relaxation by incorporating more precise lower bounds could lead to stronger robustness guarantees. Finally, extending the proposed framework to broader time-series tasks, such as classification, presents a promising direction for future research.

Acknowledgements

We thank Dr. Tarun Soni and Kerry Brown for their helpful discussions and valuable feedback. Sarah Monazam Erfani is in part supported by the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) DE220100680.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[2] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021.

[3] Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3395–3404, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Taha Belkhouja and Janardhan Rao Doppa. Adversarial Framework With Certified Robustness for Time-Series Domain via Statistical Features. *Journal of Artificial Intelligence Research*, 73:1435–1471, apr 2022. arXiv:2207.04307 [cs].

[6] Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Dynamic time warping based adversarial framework for time-series domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7353–7366, 2022.

[7] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, 54(3):1–33, 2021.

[8] Yuanpu Cao, Lu Lin, and Jinghui Chen. Adversarially robust industrial anomaly detection through diffusion model. *arXiv preprint arXiv:2408.04839*, 2024.

[9] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[10] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[11] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[12] Pin-Yu Chen and Cho-Jui Hsieh. *Adversarial robustness for machine learning*. Academic Press, 2022.

[13] Wenchao Chen, Long Tian, Bo Chen, Liang Dai, Zhibin Duan, and Mingyuan Zhou. Deep variational graph convolutional recurrent network for multivariate time series anomaly detection. In *International conference on machine learning*, pages 3621–3633. PMLR, 2022.

[14] Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33:1275–1286, 2020.

[15] Ping-yeh Chiang, Michael J. Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as Regression: Certified Object Detection by Median Smoothing, feb 2022. arXiv:2007.03730 [cs].

[16] Kukjin Choi, Jihun Yi, Changhwa Park, and Sungroh Yoon. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access*, 9:120043–120065, 2021.

[17] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019.

[18] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. *arXiv:1902.02918 [cs, stat]*, jun 2019. arXiv: 1902.02918.

[19] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[20] Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33:8270–8283, 2020.

[21] Hossein Estiri, Zachary H Strasser, Jeffy G Klann, Pourandokht Naseri, Kavishwar B Wagholikar, and Shawn N Murphy. Predicting covid-19 mortality with electronic medical records. *NPJ digital medicine*, 4(1):15, 2021.

[22] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of operational research*, 270(2):654–669, 2018.

[23] Nicola Franco, Daniel Korth, Jeanette Miriam Lorenz, Karsten Roscher, and Stephan Guennemann. Diffusion denoised smoothing for certified and adversarial robust out-of-distribution detection. *arXiv preprint arXiv:2303.14961*, 2023.

[24] Tomasz Górecki and Maciej Łuczak. Non-isometric transforms in time series classification using dtw. *Knowledge-based systems*, 61:98–108, 2014.

[25] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

[26] Miklós Z Horváth, Mark Niklas Müller, Marc Fischer, and Martin Vechev. Boosting randomized smoothing with variance reduced classifiers. *arXiv preprint arXiv:2106.06946*, 2021.

[27] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 387–395, New York, NY, USA, 2018. Association for Computing Machinery.

[28] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7:358–386, 2005.

[29] Sultan Uddin Khan, Mohammed Mynuddin, and Mahmoud Nabil. Adaptedge: Targeted universal adversarial attacks on time series data in smart grids. *IEEE Transactions on Smart Grid*, 2024.

[30] Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1)*, 2021.

[31] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.

[32] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified Robustness to Adversarial Examples With Differential Privacy. *arXiv:1802.03471 [cs, stat]*, may 2019. arXiv: 1802.03471.

[33] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified Adversarial Robustness With Additive Noise. *arXiv:1809.03113 [cs, stat]*, nov 2019. arXiv: 1809.03113.

[34] Hui Li, Yunpeng Cui, Shuo Wang, Juan Liu, Jinyuan Qin, and Yilin Yang. Multivariate Financial Time-Series Prediction With Certified Robustness. *IEEE Access*, 8:109133–109143, 2020. Conference Name: IEEE Access.

[35] Shuyang Li, Gianluca Francini, and Enrico Magli. Temporal dynamics clustering for analyzing cell behavior in mobile networks. *Computer Networks*, 223:109578, 2023.

[36] Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3220–3230, 2021.

[37] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29:565–592, 2015.

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]*, sep 2019. arXiv: 1706.06083.

[40] Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2018.

[41] David MQ Nelson, Adriano CM Pereira, and Renato A De Oliveira. Stock market's price movement prediction with lstm neural networks. In *2017 International joint conference on neural networks (IJCNN)*, pages 1419–1426. IEEE, 2017.

[42] Guillermo Ortiz-Jiménez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. *Proceedings of the IEEE*, 109(5):635–659, 2021.

[43] Ambar Pal and Jeremias Sulam. Understanding noise-augmented training for randomized smoothing. *arXiv preprint arXiv:2305.04746*, 2023.

[44] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[45] Zhuang Qian, Kaizhu Huang, Qiu-Feng Wang, and Xu-Yao Zhang. A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies. *Pattern Recognition*, 131:108889, 2022.

[46] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[47] Toni M Rath and R Manmatha. Lower-bounding of dynamic time warping distances for multivariate time series. *University of Massachusetts Amherst Technical Report MM*, 40:1–4, 2002.

[48] Rolf Reichle, Gabrielle De Lannoy, Randal Koster, Wade Crow, John Kimball, Qing Liu, and Michel Bechtold. Smap l4 global 3-hourly 9 km ease-grid surface and root zone soil moisture geophysical data, version 8, 2025.

[49] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3009–3017, 2019.

[50] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 3009–3017, New York, NY, USA, 2019. Association for Computing Machinery.

[51] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR, 10–15 Jul 2018.

[52] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 2003.

[53] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

[54] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Black-box smoothing: A provable defense for pretrained classifiers. *arXiv preprint arXiv:2003.01908*, 2(2), 2020.

[55] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sébastien Bubeck. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. *Advances in neural information processing systems*, page 31, 2019.

[56] Ibraheem Shayea, Abdulraqeb Alhammadi, Ayman A El-Saleh, Wan Haslina Hassan, Hafizal Mohamad, and Mustafa Ergen. Time series forecasting model of future spectrum demands for mobile broadband networks in malaysia, turkey, and oman. *Alexandria Engineering Journal*, 61(10):8051–8067, 2022.

[57] Ryan Sheatsley, Nicolas Papernot, Michael Weisman, Gunjan Verma, and Patrick McDaniel. Adversarial examples in constrained domains. *arXiv preprint arXiv:2011.01183*, 2020.

[58] Lifeng Shen, Zhuocong Li, and James Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13016–13026. Curran Associates, Inc., 2020.

[59] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 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*, KDD '19, page 2828–2837, New York, NY, USA, 2019. Association for Computing Machinery.

[60] Yixin Su, Yongxin Zhao, Chao Niu, Rong Liu, Weijie Sun, and Jian Pei. 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*, pages 2828–2837. ACM, 2019.

[61] Shahroz Tariq, Binh M Le, and Simon S Woo. Towards an awareness of time series anomaly detection models' adversarial vulnerability. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3534–3544, 2022.

[62] Chunzhi Wang, Shaowen Xing, Rong Gao, Lingyu Yan, Naixue Xiong, and Ruoxi Wang. Disentangled dynamic deviation transformer networks for multivariate time series anomaly detection. *Sensors*, 23(3):1104, 2023.

[63] Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*, 2019.

[64] Yongfeng Wang and Guofeng Yan. Survey on the application of deep learning in algorithmic trading. *Data Science in Finance and Economics*, 1(4):345–361, 2021.

[65] Zhibo Wang, Mengkai Song, Siyan Zheng, Zhifei Zhang, Yang Song, and Qian Wang. Invisible adversarial attack against deep neural networks: An adaptive penalization approach. *IEEE Transactions on Dependable and Secure Computing*, 18(3):1474–1488, 2019.

[66] Baoyuan Wu, Shaokui Wei, Mingli Zhu, Meixi Zheng, Zihao Zhu, Mingda Zhang, Hongrui Chen, Danni Yuan, Li Liu, and Qingshan Liu. Defenses in adversarial machine learning: A survey. *arXiv preprint arXiv:2312.08890*, 2023.

[67] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

[68] Renjie Wu and Eamonn J Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2421–2429, 2021.

[69] Tao Wu, Xuechun Wang, Shaojie Qiao, Xingping Xian, Yanbing Liu, and Liang Zhang. Small perturbations are enough: Adversarial attacks on time series prediction. *Information Sciences*, 587:794–812, 2022.

[70] Fengli Xu, Yuyun Lin, Jiaxin Huang, Di Wu, Hongzhi Shi, Jeungeun Song, and Yong Li. Big data driven mobile traffic understanding and forecasting: A time series approach. *IEEE Transactions on Services Computing*, 9(5):796–805, 2016.

[71] Hongzuo Xu, Yijie Wang, Songlei Jian, Qing Liao, Yongjun Wang, and Guansong Pang. Calibrated one-class classification for unsupervised time series anomaly detection. *arXiv preprint arXiv:2207.12201*, 2022.

[72] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized Smoothing of All Shapes and Sizes. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10693–10705. Pmlr, nov 2020. Issn: 2640-3498.

[73] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020.

[74] Wenbo Yang, Jidong Yuan, Xiaokang Wang, and Peixiang Zhao. Tsadv: Black-box adversarial attack on time series with local perturbations. *Engineering Applications of Artificial Intelligence*, 114:105218, 2022.

[75] Chengyuan Yao, Pavol Bielik, Petar Tsankov, and Martin Vechev. Automated discovery of adaptive attacks on adversarial defenses. *Advances in Neural Information Processing Systems*, 34:26858–26870, 2021.

[76] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's contributions and underlying assumptions are clearly stated at the end of the abstract and in the concluding paragraph of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work are discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theoretical results in this paper are provided with full set of assumptions and a complete proof as in Section 3.2 and Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details to reproduce the algorithm are provided in Section 3.3, the experiments settings are provided in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and environment file are provided in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: he training and testing details are provided in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The statistical significance of the experiments are discussed in Section 4 Evaluation Metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational resources used for the experiments are specified in Section 4 Settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research was conducted in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are discussed Appendix G.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release new data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The data and models used in the paper are properly cited as detailed in Section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Randomized Smoothing Details

Randomized smoothing [17] was originally proposed and widely applied in classification tasks by constructing the *smoothed function* $g(x)$ that takes the Gaussian means of the base function f as

$$g(x) = \mathbb{E}_{\eta \sim N(0,\sigma^2 I)}[f(x+\eta)] \ . \tag{12}$$

Lemma A.1. *Given a bounding output range as $f : \mathcal{X} \to [l, u]$, the upper and lower bounds on the output of the Gaussian smoothed function $g(x')$ can be shown as [14]*

$$l + (u-l)\cdot\Phi\Big(\frac{k(x) - \|x - x'\|_2}{\sigma}\Big) \ \le \ g\big(x'\big) \ \le \ l + (u-l)\cdot\Phi\Big(\frac{k(x) + \|x - x'\|_2}{\sigma}\Big) \ , \tag{13}$$

where $k(x) = \sigma \cdot \Phi^{-1}(\frac{g(x)-l}{u-l})$ and Φ denote the cumulative distribution function (CDF) of the standard Gaussian distribution.

In the context of classification, $g(x)$ is interpreted as the bounded probability score in range $[0, 1]$ for each label, e.g., the softmax score, and a certificate can be obtained by bounding the gap between the highest and second-highest scores.

B Proof of Theorem 3.3

Here we provided the complete proof of the Theorem 3.3.

Proof. First, we prove that $d(x') = d(x)$ holds for all x' satisfying $\|x - x'\| \le r$. In the case of $d(x) = 1$, i.e., $h_p(x) > \gamma$, we prove that all x' satisfies $h_p(x') > \gamma$. Given the inequality $h_{\underline{p}}(x) < h_p(x')$ for all $\|x - x'\|_2 < r$, as established in Lemma 3.1, it follows that if $h_{\underline{p}}(x) > \gamma$, then $\gamma < h_{\underline{p}}(x) < h_p(x')$, which ensures that $h_p(x') > \gamma$. In the case of $d(x) = 0$, the proof follows by a similar argument. The radius r can be solved by the definitions of $\underline{p} = \Phi(\Phi^{-1}(p) - \frac{r}{\sigma})$ and $\overline{p} = \Phi(\Phi^{-1}(p) + \frac{r}{\sigma})$ as given in Lemma 3.1.

By Lemma 3.2, such certification can be transited to $\{x' \mid DTW(x, x') \le e\}$ by solving the $e = \inf\{LB(x, x') \mid \|x - x'\| > r\}$ with a proper choice of the lower bound $LB(x, x')$.

We consider the Keogh Lower Bound [28] $LB\_Keogh(x, x')$, which is a strict lower bound of DTW for any $w > 0$ and $x' \neq x$. The $LB\_Keogh(x, x')$ is calculated as the sum of deviation of x' outside the envelope of x. For each time step i we define the slack (allowable deviation) without incurring any penalty in $LB\_Keogh(x, x')$ by $\Delta_i = \max\big(U_i - x_i,\ x_i - L_i\big)$ and the sum of all time steps as $R = \sqrt{\sum_{i=1}^{n} \Delta_i^2}$. Thus, if x' could "hide" all the norm deviation r within the slacks for all time steps as $r \le R$, the $LB\_Keogh(x, x') = 0$. Hence, in that case, $e = 0$.

Then consider the case when $r > R$, which means any x' with $\|x - x'\| > r$ must have at least one coordinate outside the envelope. Since we are solving for the infimum, the smallest possible $LB\_Keogh(x, x')$ is when $\|x - x'\| = r$ and the x' use up the available slack Δ_i in every coordinate except one i^* where the slack Δ_{i^*} is largest. Then, such a worst-case x' is defined as

$$x'_i = \begin{cases} x_i + \Delta_i, & \text{if } i \neq i^*, \\ x_{i^*} + \Delta_{i^*} + d, & \text{if } i = i^*. \end{cases} \tag{14}$$

with d as the part outside the envelops and $\|x - x'\| = r$. In that case

$$r^2 = \|x - x'\|^2 = \sum_{i \neq i^*} \|\Delta_i\|^2 + \|\Delta_{i^*} + d\|^2 = \left(\sum_{i=1}^{n} \Delta_i^2\right) + 2\, d^T\, \Delta_{i^*} + \|d\|^2 \tag{15}$$

Note that the $LB\_Keogh(x, x')$ is calculated as the sum of deviations outside the envelope. Thus, the infimum e when $r > R$ can be obtained by solving $\|d\|$. To yield the extreme value of $\|d\|$, d and Δ_{i^*} should be collinear and can be written as $d = \lambda \Delta_{i^*}$. With the substitution, the equation becomes

$$(\lambda^2 + 2\lambda)M^2 + R^2 = r^2 \ . \tag{16}$$

Solve for the λ, we have

$$\lambda = -1 \pm \sqrt{1 + \frac{r^2 - R^2}{M^2}} \ . \tag{17}$$

Therefore, the infimum value of $\|d\|$ is

$$||d|| = \sqrt{M^2 + r^2 - R^2} \ - \ M = e \ . \tag{18}$$

□

C Extension to ℓ_p Norm

Generalization of Randomized Smoothing to Arbitrary Norms Our certified robustness analysis in the main text is built upon randomized smoothing under the ℓ_2 norm using Gaussian noise. The framework naturally extends to arbitrary ℓ_p norms by replacing the isotropic Gaussian distribution with a noise distribution that is radially symmetric with respect to the chosen norm [72]. Let $\|\cdot\|_p$ denote the base norm and $\|\cdot\|_q$ its dual norm, where $\frac{1}{p} + \frac{1}{q} = 1$. We consider a noise vector η drawn from a distribution that is *spherically symmetric* with respect to $\|\cdot\|_p$, such that the density of η depends only on $\|\eta\|_p/s$, where s is a scale parameter. Typical choices include: ℓ_2 with Gaussian noise $\eta \sim \mathcal{N}(0, \sigma^2 I)$; ℓ_1 with Laplace noise $\eta_i \sim \text{Laplace}(0, b)$ i.i.d.; ℓ_∞ with Uniform noise $\eta_i \sim \text{Unif}[-\tau, \tau]$ i.i.d. General ℓ_p can use generalized Gaussian noise with density proportional to $\exp(-\|\eta\|_p^\alpha / \lambda^\alpha)$ for $\alpha = p$.

Quantile Stability under ℓ_p Perturbations Let f denote the base anomaly scoring function and $h_p(x)$ the p-th percentile of $f(x + \eta)$ under the smoothing noise distribution. For any $r \geq 0$, define F as the cumulative distribution function of $\langle u, \eta \rangle$ for any unit vector u with $\|u\|_q = 1$. Then, the following holds for all $\|x' - x\|_p \leq r$:

$$h_{F(F^{-1}(p) - r/s)}(x') \leq h_p(x) \leq h_{F(F^{-1}(p) + r/s)}(x'). \tag{19}$$

Equation (19) generalizes the Gaussian case by replacing the standard normal CDF Φ with the 1-D marginal F of the chosen noise distribution, and the Gaussian scale σ with the corresponding scale parameter s. This yields a certified radius in ℓ_p norm space.

DTW Certification via ℓ_p Lower Bounds The DTW-based certification derived in Lemma 3.2 remains valid once ℓ_2 is replaced by ℓ_p. Specifically, let $\text{LB}_{\text{Keogh},p}(x, x')$ denote a *strict* lower bound of $\text{DTW}_p(x, x')$. For any perturbation $\|x' - x\|_p \leq r$, the certified DTW radius is

$$e_p = \inf\big\{ \text{LB}_{\text{Keogh},p}(x, x') : \ \|x' - x\|_p > r \big\}. \tag{20}$$

A practical closed-form lower bound can be obtained via

$$e_p \ \geq \ \big(r^p - R_{\text{in},p}^p\big)_+^{1/p}, \tag{21}$$

where $R_{\text{in},p}$ is the largest ℓ_p ball centered at x fully contained within the envelope $[L, U]$ used in the $\text{LB}_{\text{Keogh},p}$ construction. This provides a conservative yet efficient computation of certified DTW radius for arbitrary norms.

This extension preserves the overall structure of the certification pipeline:

1. Replace Gaussian noise with a norm-symmetric distribution;
2. Replace the Gaussian CDF Φ by the 1-D marginal CDF F of that noise;
3. Compute the ℓ_p certified radius r via (19);
4. Translate the ℓ_p certificate into DTW certificate e_p using (20).

This demonstrates that the proposed percentile-based randomized smoothing framework is inherently norm-agnostic, supporting robustness certification under any ℓ_p metric and its induced DTW variants.

D Dataset Details

- The Soil Moisture Active Passive (SMAP) dataset [48] contains soil moisture and telemetry measurements collected by NASA's Mars rover.

Datasets	Channels	Training Timesteps	Testing Timesteps	Testing Anomalies Ratio %
SMAP	25	135,183	427,617	13.13%
MSL	55	58,317	73,729	10.72%
SMD	25	708,405	708,420	4.16%
NIPS-TS-SWAN	38	60,000	60,000	32.60%
NIPS-TS-CREDITCARD	29	284,807	284,807	0.17%
NIPS-TS-WATER	9	69,260	69,260	1.05%
UCR-1	1	35,000	44,795	1.38%
UCR-2	1	35,000	45,000	0.67%

Table 4: Statistics of the benchmark datasets for time-series anomaly detection.

- The Mars Science Laboratory (MSL) dataset [27] includes comprehensive sensor and actuator data directly obtained from the Mars rover.
- The Server Machine Dataset (SMD)[60] offers stacked resource utilization data from 28 machines within a compute cluster, collected over a five-week duration.
- The NIPS-TS benchmark suite[30] and the UCR collection [68], which provide standardized datasets widely employed in time-series anomaly detection.

E Certified Confusion Matrix

	Predicted Positive	Predicted Negative
Positive	Certified TP$(t) = \sum_{i=1}^{N} \mathbb{I}\{\forall x' \ : \ DTW(x_i, x') \leq t : \ f(x') = 1, y_i = 1\}$	$\sum_{i=1}^{N} \mathbb{I}\{y_i = 1\}$ − Certified TP(t)
Negative	FP	TN

Table 5: Certified Confusion Matrix for evasion attacks.

	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	$\sum_{i=1}^{N} \mathbb{I}\{y_i = 0\}$ − Certified TN(t)	Certified TN$(t) = \sum_{i=1}^{N} \mathbb{I}\{\forall x' \ : \ DTW(x_i, x') \leq t : \ f(x') = 0, y_i = 0\}$

Table 6: Certified Confusion Matrix for availability attacks.

Certified accuracy is a metric widely used in certified robust machine learning, measuring the fraction of examples for which a model can provably maintain correct predictions under specific perturbations. For a certified radius e, it is defined as

$$\text{Certified Accuracy}(e) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{\forall x' \text{ with } DTW(x_i, x') \leq e : \ f(x') = y_i\} \tag{22}$$

Following the definition of certified accuracy, we construct the certified confusion matrix as described in Section 4. Given the certified confusion matrix, the *certified accuracy* is computed as the proportion of instances for which the model guarantees correct predictions within a perturbation threshold. It is defined as:

$$\text{Certified Accuracy} = \frac{\text{Certified TP} + \text{Certified TN}}{N}, \tag{23}$$

where N is the total number of test instances.

Similarly, the *certified F1-score*, which balances precision and recall under certification constraints, is calculated as:

$$\text{Certified F1} = \frac{2 \cdot \text{Certified Precision} \cdot \text{Certified Recall}}{\text{Certified Precision} + \text{Certified Recall}}, \tag{24}$$

where

$$\text{Certified Precision} = \frac{\text{Certified TP}}{\text{Certified TP} + \text{FP}}, \tag{25}$$

$$\text{Certified Recall} = \frac{\text{Certified TP}}{\text{Certified TP} + \text{FN}}. \tag{26}$$

Here, FP and FN refer to false positives and false negatives, respectively, counted as the remaining instances not included in the certified true predictions. These metrics provide a conservative evaluation of model robustness under worst-case adversarial perturbations.

F Additional Experiment Results

Seq. Length T	Window Size w	Standard		DTW-Certified Defense					
		F1-score	ROC AUC	F1-score	ROC AUC	Radii Mean	Radii Max	Radii Std.	Certified Prop.
10	2	0.571	0.856	0.671	0.943	0.088	0.337	0.053	94.40%
	4					0.085	0.333	0.053	92.59%
	10					0.083	0.326	0.054	91.01%
50	2	0.675	0.956	0.624	0.966	0.090	0.401	0.058	82.75%
	4					0.083	0.386	0.060	77.99%
	10					0.074	0.374	0.061	71.05%
100	2	0.656	0.929	0.447	0.878	0.131	0.699	0.104	79.69%
	4					0.119	0.678	0.107	71.54%
	10					0.107	0.635	0.107	63.30%
200	2	0.681	0.963	0.440	0.880	0.057	0.392	0.076	48.50%
	4					0.050	0.391	0.074	41.05%
	10					0.041	0.389	0.070	33.33%

Table 7: Empirical and certified robustness results for the SMD dataset using the COUTA model with $\sigma = 0.5$, evaluated under varying sequence length T and DTW wrapping window size w.

Table 7 presents an ablation study on the impact of sequence length T and DTW wrapping window size w on both detection performance and certified robustness. The results indicate a trade-off between these parameters and robustness guarantees. Increasing the sequence length generally enhances detection performance (F1-score and ROC AUC) by incorporating more temporal context for anomaly detection. However, this comes at the cost of reduced certified radius, as the higher dimensionality magnifies the impact of injected noise. Similarly, increasing the wrapping window w allows greater temporal flexibility in DTW alignment but leads to looser Keogh lower bounds and higher slack (as defined by the value R in Theorem 3.3), thereby weakening the robustness guarantee.

G Border Impact

Time-series anomaly detection plays a crucial role in many safety-critical domains, including healthcare monitoring, financial fraud detection, industrial control systems, and mobile communication networks. In such applications, robustness to adversarial manipulation is not only a matter of performance but also of safety, reliability, and trust. This work contributes to the broader goal of deploying machine learning systems that are resilient to worst-case perturbations in time-series data, particularly those involving temporal distortions.

Our proposed DTW-certified defense offers a principled approach to formally quantifying and improving the robustness of anomaly detection systems under realistic threat models. By aligning the certification metric with the temporal structure of time-series data, we aim to enable more reliable AI systems in high-stakes environments. However, we acknowledge that any advancement in robustness may also encourage the development of stronger adversarial strategies. As such, we encourage responsible deployment and continuous evaluation of these defenses in real-world conditions.

This work is primarily beneficial to organizations seeking reliable time-series analytics in critical domains. It does not disproportionately disadvantage any particular group. Nonetheless, as with any security-related research, care should be taken to ensure that the methodology is not misused to benchmark or strengthen attack strategies without accompanying safeguards.