XTadGAN

Generative Adversarial Networks to Detect Extremely Rare Anomalies

Master in Data Science and Engineering
Dissertation

Nuno Vasconcelos
Supervised by Carlos Soares, PhD & Vítor Cerqueira, PhD

Today's agenda

01
Motivation & Research Objectives

02
Novel framework for Time Series Evaluation

03
Detecting Extremely Rare Anomalies

04
Conclusions & Future Work

Today's agenda

01

Motivation & Research Objectives

02

Novel framework for Time Series Evaluation

03

Detecting Extremely Rare Anomalies

04

Conclusions & Future Work

Deciphering the thesis subject

Context and Introduction

XTadGAN

Generative Adversarial Networks to Detect Extremely Rare Anomalies

GANs – Generative Adversarial Networks

- One of the "hottest" and more promising fields of study at the moment
- Proven to be very successful in generative contexts (especially images)
- Not much work done leveraging these two fields of study, despite promising results

Anomaly Detection in Time Series

- One of the most important data structures in real-world applications
- Immense practical applications
- One of the most researched fields of study using traditional approaches

Extremely Rare Anomalies

- Increases the complexity and difficulty of detection
- Turns an already imbalanced problem in an even more challenging scenario
- Better suited for real-world, often critical, applications

GANs for anomaly detection in time series

Using adversarial training to improve on unsupervised anomaly detection techniques

Prior GAN-related work has **rarely involved time series data**, because the complex temporal correlations within this type of data pose significant **challenges to generative modeling** [Geiger et al., 2020]

- **Three works** published between 2019 and 2020 started to **change the landscape** of GANs in the context of time series

MAD-GAN
[Li et al., 2019]

- **Standard GAN** to model the time series data for anomaly detection
- Use the **Discriminator to flag** anomalies
- **Outperforms traditional methods** in terms of accuracy and efficiency

BeatGAN
[Zhou et al., 2019]

- Applied to anomaly detection in **heartbeat signals**
- **Paradigm shift:** introduction of the **reconstruction error** as a metric for detecting and signaling anomalies

  (instead of the output of the Discrminінator network)

TadGAN
[Geiger et al., 2020]

- Introduce two breakthroughs
- Training the model with **cycle consistency loss**
- Explore approaches to **combine reconstruction errors** with **Discriminator outputs**

Exploring Related Works

Current Evaluation Benchmarks

No works have been identified that explore algorithmic **responsiveness to** a spectrum of **anomaly frequencies**

Almost **all academic research** is done using **three main sources** (besides *private* datasets)

Data Source	NASA		YAHOO				NUMENTA				
	MSL	SMAP	A1	A2	A3	A4	Art	AdEx	AWS	Traffic	Tweets
# Series	27	53	67	100	100	100	6	5	17	7	10
Point	0	0	68	33	935	833	0	0	0	0	0
Collective	36	67	110	167	4	2	6	11	30	14	33
Anomalous points	7.766	54.696	1.669	466	943	837	2.418	795	6.312	1.560	15.651
Total points	132.046	562.800	94.866	142.100	168.000	168.000	24.192	7.965	67.644	15.662	158.511
Anomaly %	5.88%	9.72%	1.76%	0.33%	0.56%	0.50%	9.99%	9.98%	9.33%	2.31%	9.87%

- **Strong criticism** over these sources of data as valid benchmarks [Wu and Keogh, 2021]
  - Triviality *(too easy)*
  - Unrealistic Anomaly Density *(too many)*
  - Mislabeled Ground Truth *(too inaccurate)*
  - Run-to-failure Bias *(too biased to the end)*

*« Because of these four flaws, we believe that many published comparisons of anomaly detection algorithms may be **unreliable**, and more importantly, much of the **apparent progress in recent years may be illusionary** »*

Motivation and Research Objectives

Main contributions

Lack of systematization in the process of comparing the performance of different anomaly detection methods, specifically regarding how sensitive they are to **variations in the frequency of anomalies**

1.1 | Develop a robust and reliable **method** for evaluating the performance of anomaly detection models with **increasing levels of anomaly rarity**, filling a gap in current research

1.2 | Create a **'sensitivity index'** to evaluate the performance of different anomaly detection algorithms across a range of anomaly frequencies

2 | Explore a **variation on the TadGAN architecture** for detecting **extremely rare** anomalies in time series data

Today's agenda

01

Motivation & Research Objectives

02

Novel framework for Time Series Evaluation

03

Detecting Extremely Rare Anomalies

04

Conclusions & Future Work

Monte Carlo sampling

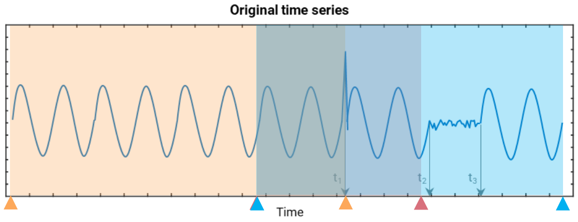
Novel framework for Time Series Evaluation

We want to

(1) **generate an arbitrarily large number** of time series

(2) each representing **different controlled scenarios**

(3) derived from a relatively **small set of original** datasets.



Input series

Monte Carlo sampling

- Generate a large number of randomly trimmed copies
- Two points are randomly selected from the original series to mark the start and end of the sample

Output samples

// Random locations
// Varying lengths
// Different **properties**

Monte Carlo sampling

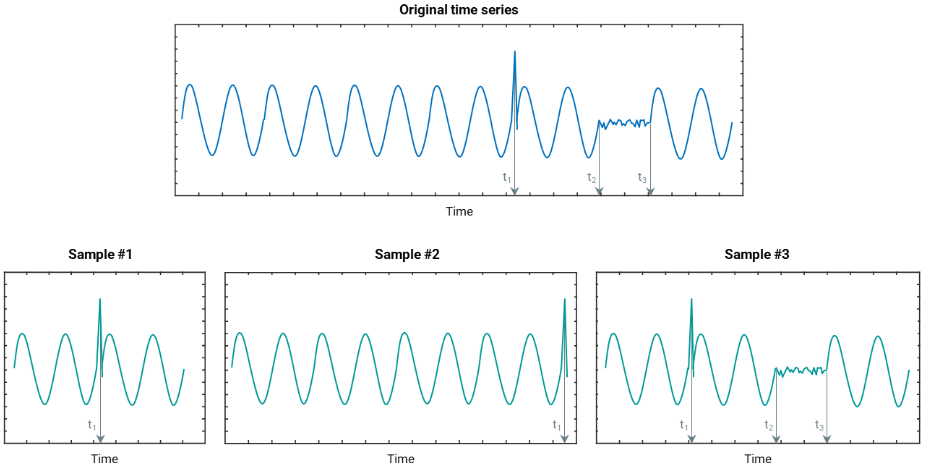
Novel framework for Time Series Evaluation

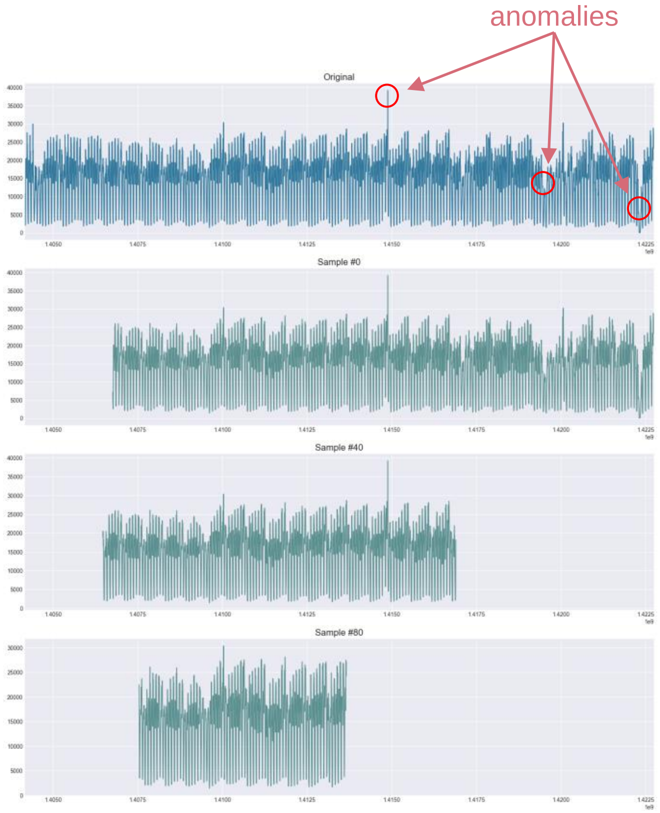
We want to

(1) **generate an arbitrarily large number** of time series

(2) each representing **different controlled scenarios**

(3) derived from a relatively **small set of original** datasets.





Not all samples are created equal

Novel framework for Time Series Evaluation

A wide range of **attributes**, or **dimensions**, is computed to **characterize** each resulting sample.

simple dimensions:

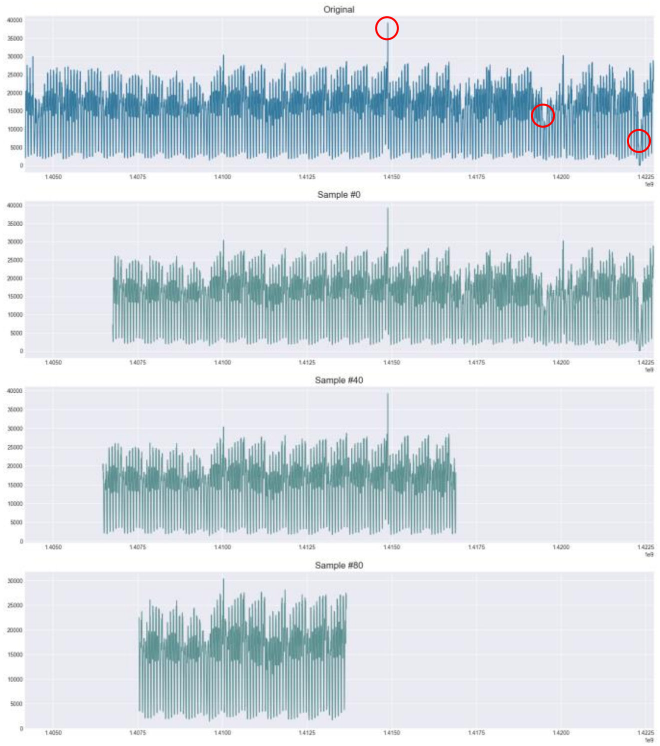
- Series length
- Start / End index *(absolute or percent)*
- Anomaly percentage

or **context-specific** attributes:

- Average distance between anomalies
- Anomaly distance from the mean
- Distance to the first anomaly

These attributes can be **tailored to the specific context** of interest

⋮



Not all samples are created equal

Novel framework for Time Series Evaluation

A wide range of **attributes**, or **dimensions**, is computed to **characterize** each resulting sample.

simple dimensions:

- Series length
- Start / End index *(absolute or percent)*
- Anomaly percentage

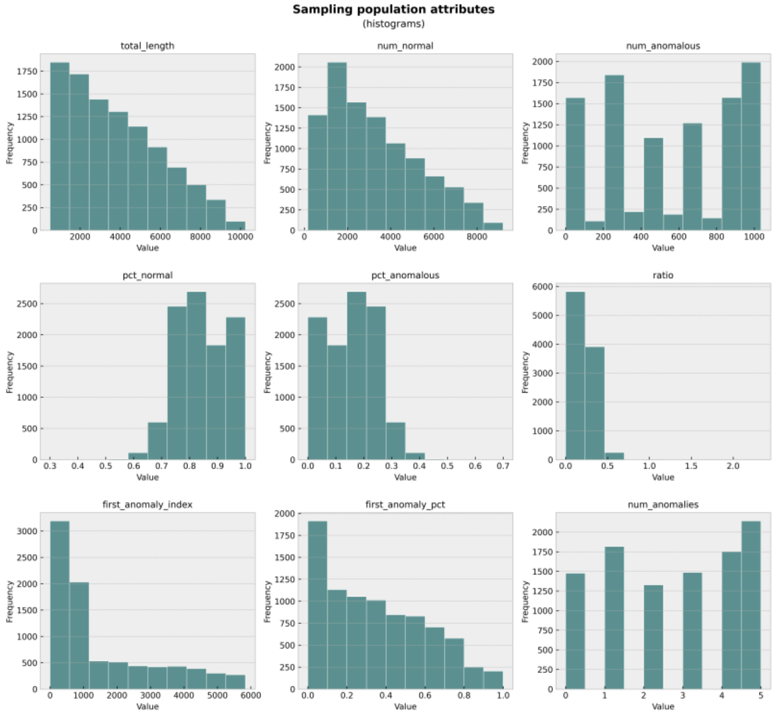
or **context-specific** attributes:

- Average distance between anomalies
- Anomaly distance from the mean
- Distance to the first anomaly

These attributes can be **tailored to the specific context** of interest

⋮

Creates a **multidimensional profile** for each generated sample

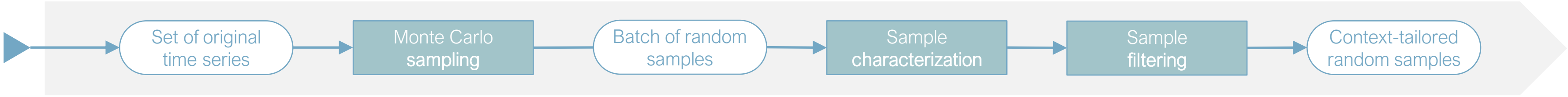


Histogram for each sampling population attribute for 10.000 generated samples using random lengths (cropped)

Complete Monte Carlo pipeline

Novel framework for Time Series Evaluation

We can **build controlled test environments** for evaluating the sensitivity of algorithms.



Systematic analysis of algorithmic sensitivity to series properties

Create **tailored experiments** that **emulate real-world** conditions while maintaining **control over variables**

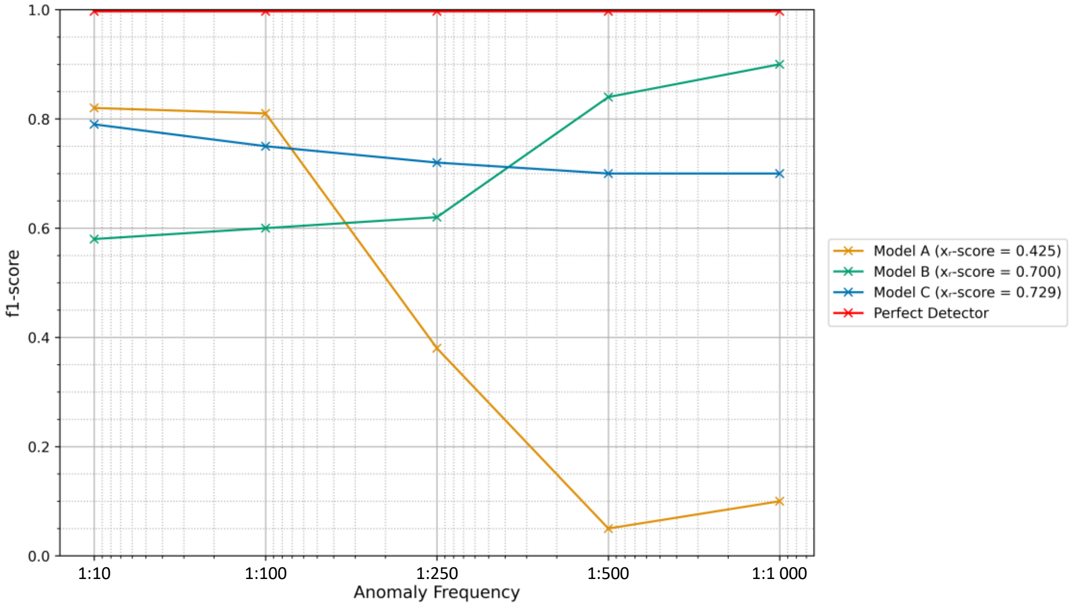
- Investigate **algorithm behavior** under various **conditions** and **parameters**
- Build upon a **limited set of original datasets** and time series (instead of requiring an exhaustive dataset collection effort)

x_r-score: rarity-spectrum score

Visualizing algorithmic performance

An **aggregate measurement** of the **performance across the entire rarity spectrum**

ranging from 0 to 1 (a perfect detector)



Valuable metric to assess the **most suitable model** for a given scenario

- In cases where the expected anomaly frequency is not known a priori:

  // Model C emerges as the safest choice to deploy

and **context-adjustable**

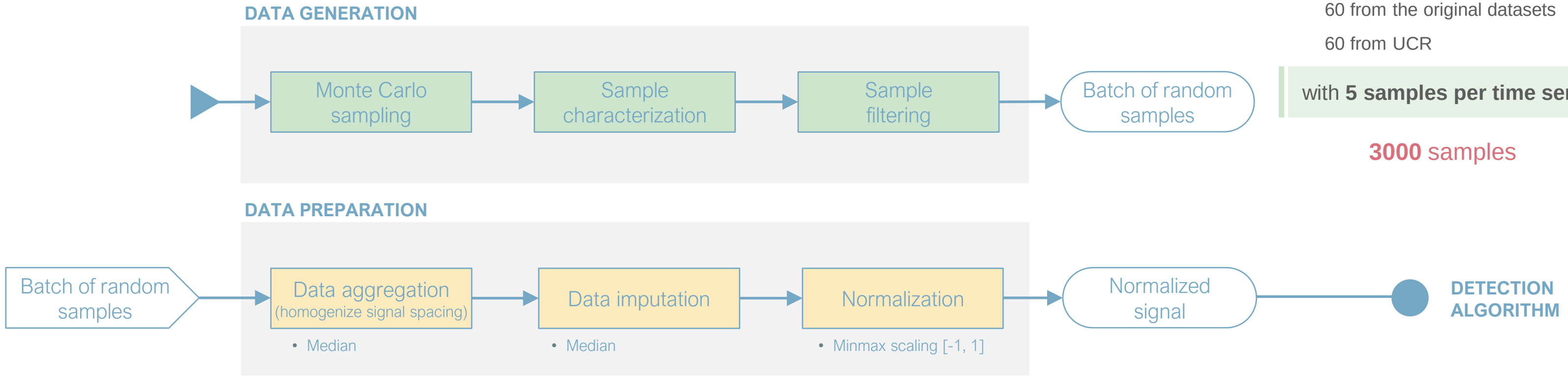
- One can redefine the spectrum for which the metric is calculated. For instance, extremely rare anomalies:

Model	x_r	$x_{r \leq 1:500}$
Model A	0.425	0.145
Model B	0.700	**0.800**
Model C	**0.729**	0.705

// Model B exhibits better performance on extremely rare anomalies

Experimental setup

Detection pipeline



5 levels of anomaly rarity

1:10 · 1:100 · 1:250 · 1:500 · 1:1000

with **120 time series each**

60 from the original datasets

60 from UCR

with **5 samples per time series**

3000 samples

Prediction-based

Classical approach:

ARIMA [Yaacob et al. (2010)]

and **Machine Learning** techniques:

LSTM [Hundman et al. (2018)]

Reconstruction-based

Autoencoders:

LSTM AE [Malhotra et al. (2015)]

LSTM VAE [An and Cho (2015)]

and **GAN-based** methods:

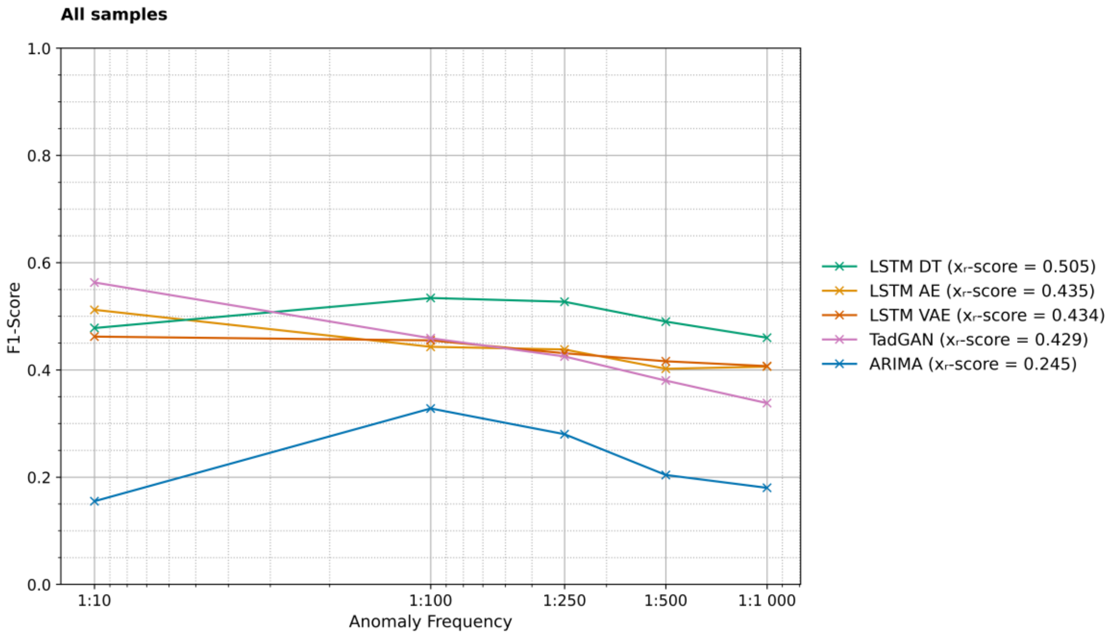
TadGAN [Geiger et al.(2020)]

Baseline Rarity Sensitivity Analysis

Research Results

A **consistent trend** emerges: **performance diminishes** notably as **anomalies become rarer**

(experiment performed on all samples from the Paper and UCR datasets)



LSTM-DT is the **most balanced** approach

- $x_r\ score = 0.505$

TadGAN is the **most affected** by anomaly **rarity**

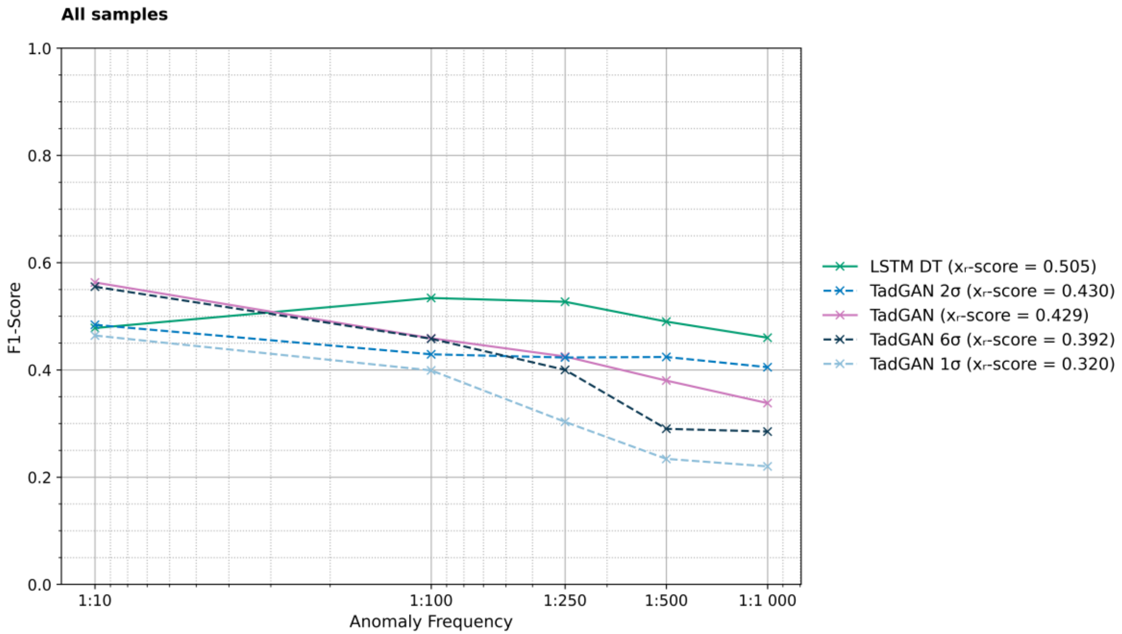
- **Original** training datasets ≈ **6.00%** anomalies
- **Smoothing effect** caused by the empirical sliding-window parameters

Systematic evaluation **matters**…
… and **sensitivity analysis** is **key**

Re-calibrating TadGAN for extremely rare anomalies

Research Results

Evaluation influences development: why sensitivity analysis is crucial



A simple **recalibration improves performance**

- Changing the original parameters allows for immediate improvement on rare anomalies

Model	x_r	$x_{r \leq 1:500}$
TadGAN	0.429	0.359
TadGAN 2σ	0.430	**0.414**

Today's agenda

01
Motivation & Research Objectives

02
Novel framework for Time Series Evaluation

03
Detecting Extremely Rare Anomalies

04
Conclusions & Future Work

Detecting Extremely Rare Anomalies

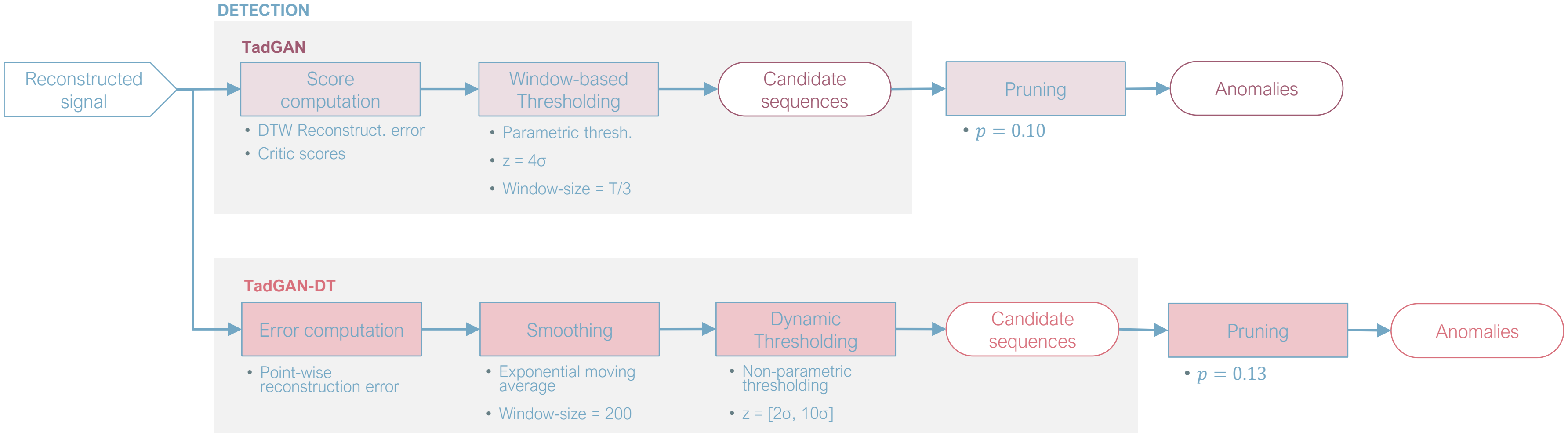
Research Results

TadGAN-DT: revamping the post-processing pipeline by incorporating non-parametric thresholding

$$score(x) = \underbrace{\alpha Z_{RE}(x)}_{\text{RECONSTR. ERROR}} \cdot \underbrace{Z_{C_x}(x)}_{\text{CRITIC SCORE}}$$

Rely solely on the **reconstruction error** to **compute anomaly scores**

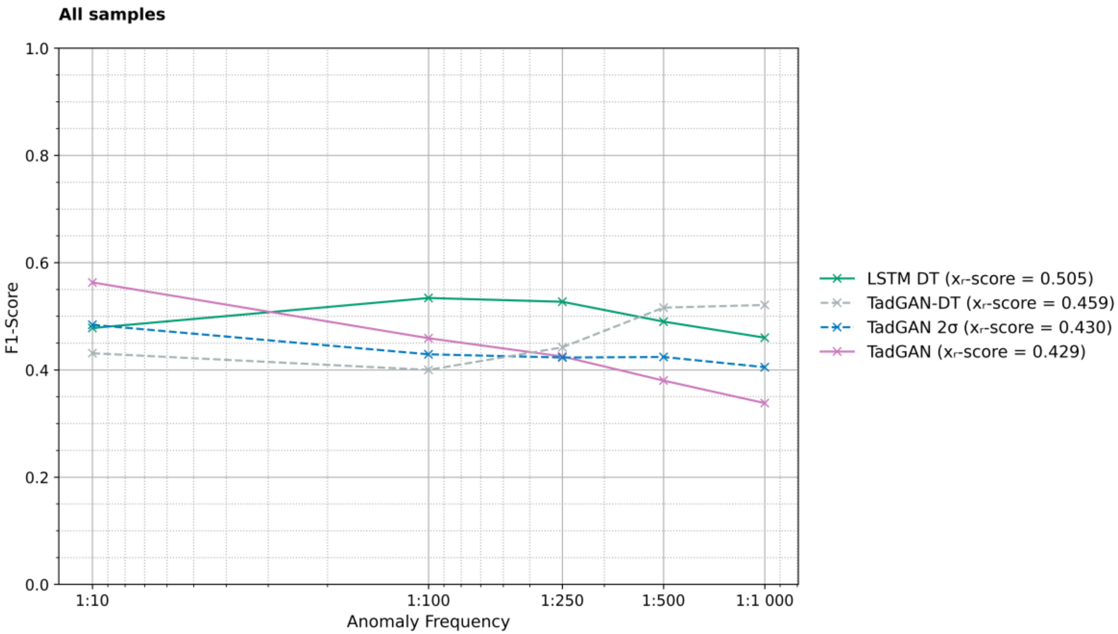
Abandon the assumption that the output **error series** follows a **Gaussian distribution** and use non-parametric Dynamic Thresholding [Hundman et al., 2018]



Detecting Extremely Rare Anomalies

Research Results

TadGAN-DT: non-parametric thresholding improves detection on rare contexts



Model	x_r	$x_{r\leq 1:500}$
LSTM-DT	**0.505**	0.475
TadGAN	0.429	0.359
TadGAN 2σ	0.430	0.414
TadGAN-DT	0.459	**0.518**

Significant **improvement** in **rare anomalies**

Less effective in more **frequent anomalies**

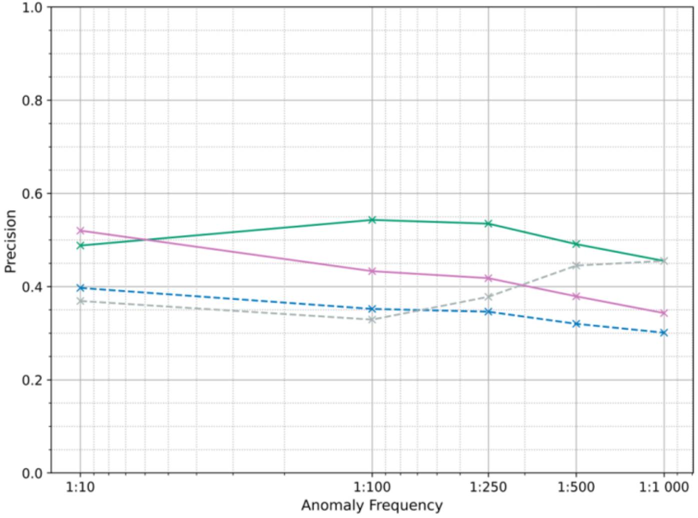
- **More aggressive pruning:** impacts on Recall
- As **rarity increases,** the number of anomalies **tends to approach 1**

Detecting Extremely Rare Anomalies

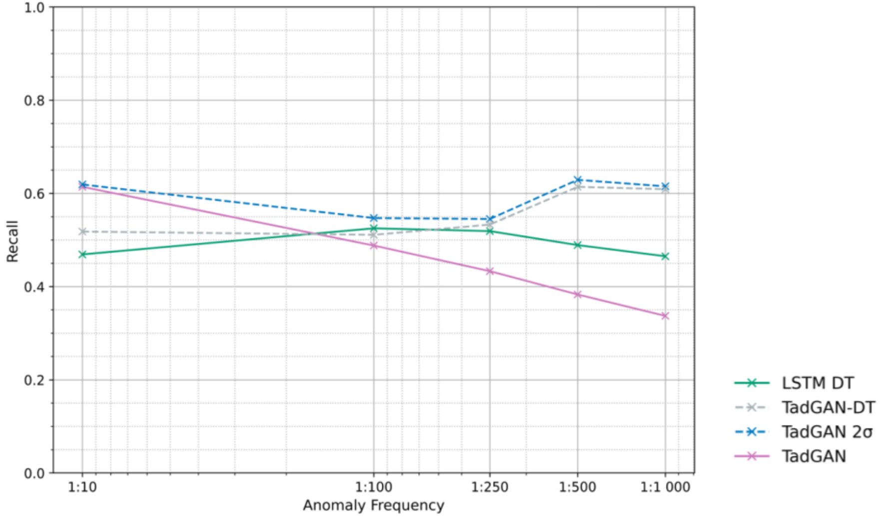
Research Results

TadGAN-DT: a boost in Precision comes at a small cost in Recall on rare contexts

Precision



Recall



Significant **boost in** extremely **rare ranges**

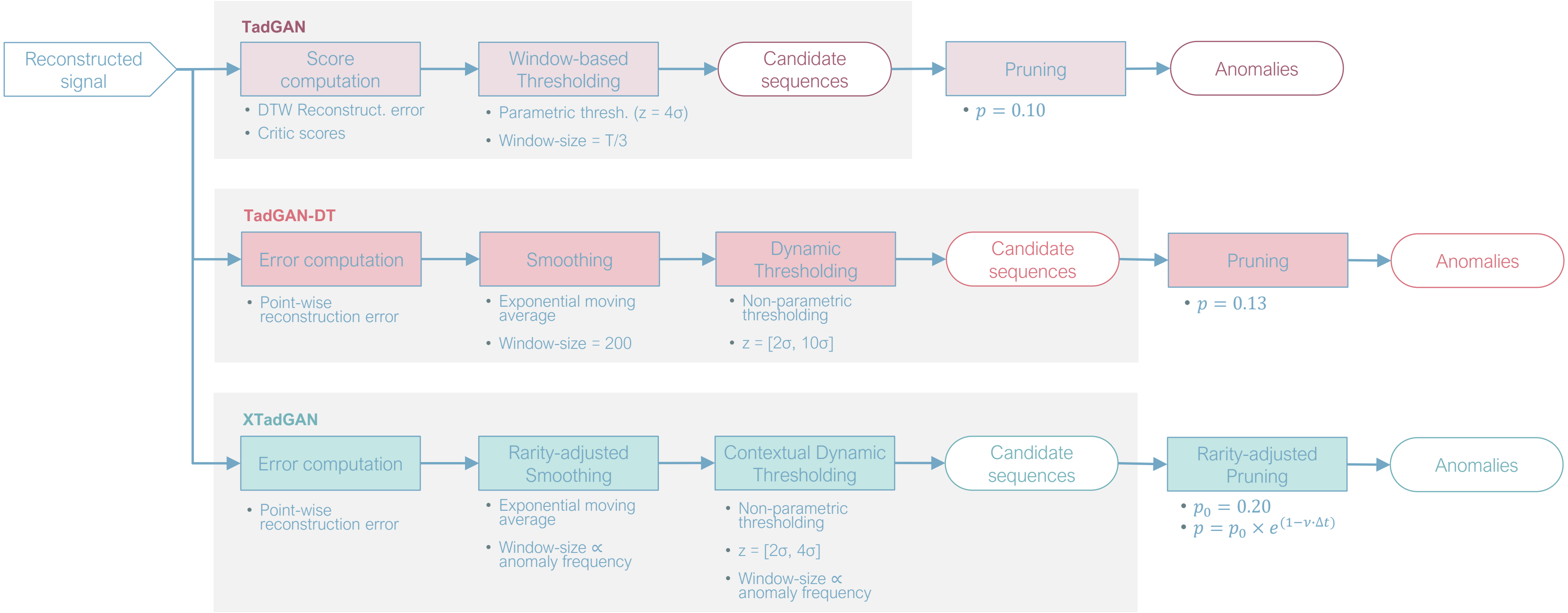
Recall not severely affected in rare contexts

Converges with LSTM-DT as frequency decreases

Detecting Extremely Rare Anomalies

Research Results

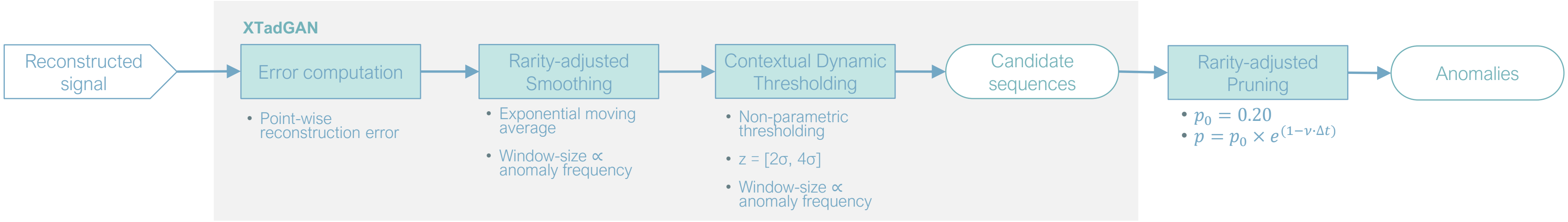
XTadGAN: use the expected anomaly frequency as a *meta-parameter* to condition detection and pruning

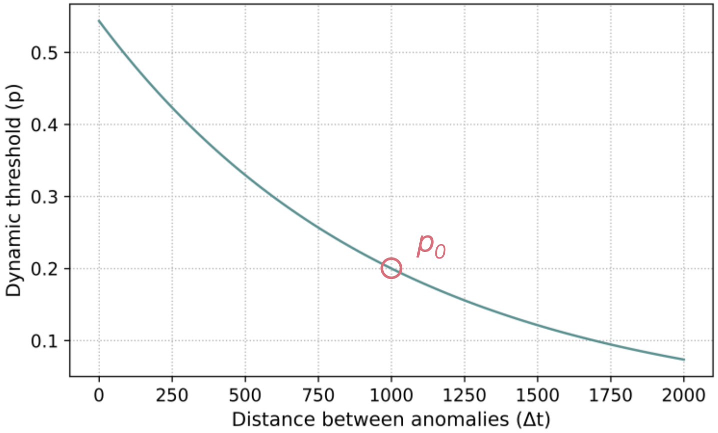


Detecting Extremely Rare Anomalies

Research Results

XTadGAN: use the expected anomaly frequency as a *meta-parameter* to condition detection and pruning





$$p = p_0 \times e^{(1-\nu\cdot\Delta t)}$$

where, p : *threshold used to classify next sequences as normal*

p_0 : *base value for the parameter* p

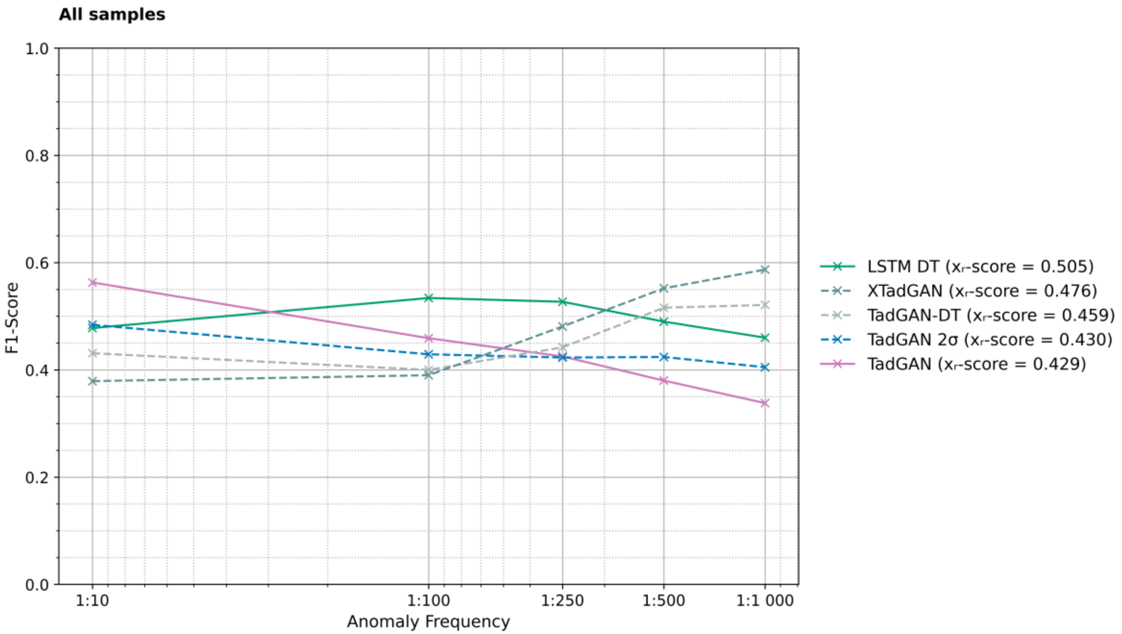
ν : *expected anomaly frequency*

Δt : *distance between candidate anomalies*

Detecting Extremely Rare Anomalies

Research Results

XTadGAN: rarity-based thresholding greatly improves detection of extremely rare anomalies



Model	x_r	$x_{r\leq 1:500}$
LSTM-DT	**0.505**	0.475
TadGAN	0.429	0.359
TadGAN 2σ	0.430	0.414
TadGAN-DT	0.459	0.518
XTadGAN	0.476	**0.570**

Highest scoring algorithm in **rare anomalies**

- The **rarity-based threshold** is **doing its job**

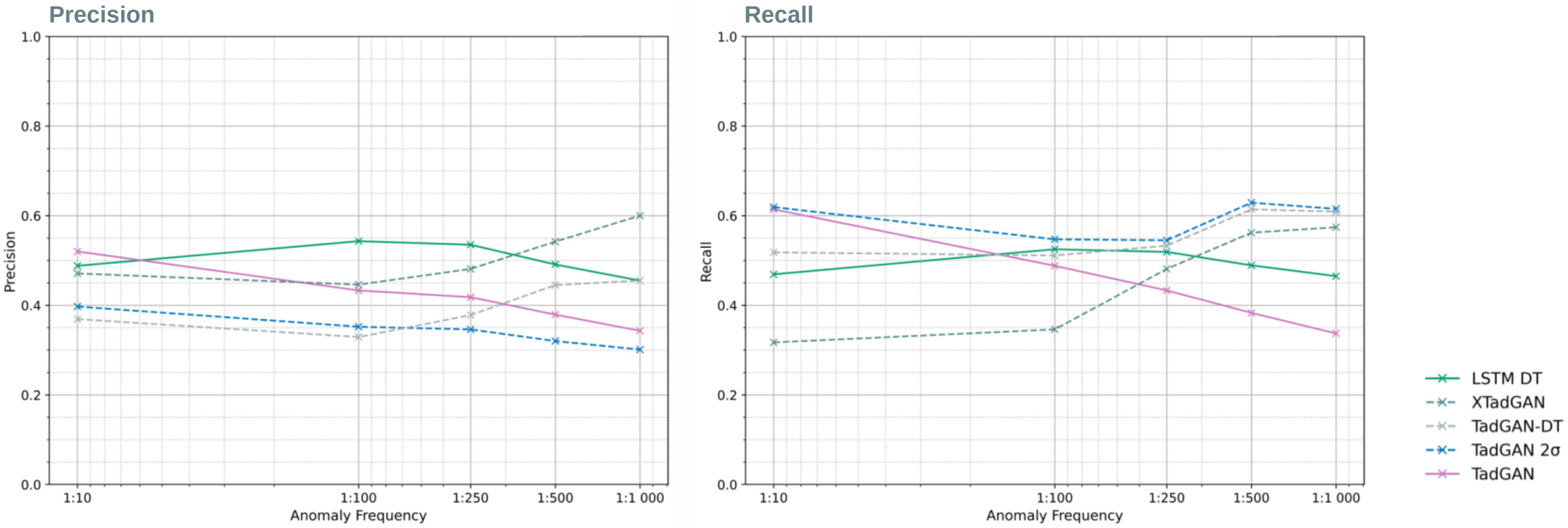
Declined performance in **frequent anomalies**

- **More aggressive pruning:** elevated FN rate

Detecting Extremely Rare Anomalies

Research Results

XTadGAN: RB pruning heightens Precision in rare anomalies, but is not as effective in higher frequencies



Precision is greatly improved as rarity increases

Recall drops, especially in more frequent anomalies

The **drawback** of **rarity-based pruning**

Today's agenda

01

Motivation & Research Objectives

02

Novel framework for Time Series Evaluation

03

Detecting Extremely Rare Anomalies

04

Conclusions & Future Work

Conclusion: New Evaluation Framework

Key Takeaways

Reviewing our **research objectives**

1.1 Develop a robust and reliable **method** for evaluating the performance of anomaly detection models with **increasing levels of anomaly rarity**, filling a gap in current research

1.2 Create a **’sensitivity index’** to evaluate the performance of different anomaly detection algorithms across a range of anomaly frequencies

Developed a comprehensive **framework for evaluating anomaly detection models**

- **Newly proposed Monte Carlo sampling** method

  Allows the creation of standardized controlled experiments to evaluate algorithmic sensitivity to series attributes

- Introduced a **sensitivity score (x_r-score)** for quantitative comparisons

Established a baseline rarity sensitivity analysis

(between state-of-the-art algorithms)

Conclusion: New GAN-based Architectures

Key Takeaways

Reviewing our **research objectives**

2 Explore a **variation on the TadGAN architecture** for detecting **extremely rare** anomalies in time series data

Introduced two novel GAN-based architectures for rare anomaly detection:

- **TadGAN-DT**

  Integrates non-parametric dynamic thresholding and pruning techniques

- **XTadGAN**

  Leverages meta-information about expected anomaly frequencies to enhance rare anomaly detection

XTadGAN outperforms other methods in **rare anomaly detection**

Future Work

Proposed research avenues

Our research opens the door to **several promising avenues** for **future exploration**

Exploring **multivariate time series data**, which was not covered in the current research

Addressing the **slow training times and high computational demands** of adversarial models in real-world applications

Expanding the **sensitivity analysis framework**

- **Increasing the number of samples** across a **wider range of anomaly levels** to bolster the robustness and comprehensiveness of our results
- Exploring algorithm behavior in different scenarios: investigating the impact of **varying the number of anomalies in fixed-length samples**

Quantifying the impact of anomaly rarity on model performance: how **changes in anomaly frequency** affect model outcomes

- Subjecting models trained on specific rarity values to **samples with different anomaly frequencies**
- Quantify **how sensitive** a particular model is **to abrupt shifts in real-world conditions** – uncovering the **"shadow price"** of rarity

Thank you