PatchTrAD: A Patch-based Transformer for time series Anomaly Detection

Vilhes Samy-Melwan [1], Gasso Gilles [1], Z. Alaya Mokhtar [2]

[1] INSA Rouen, Univ Rouen, LITIS UR 4108 [2] Univ de Technologie de Compiègne, LMAC EA 222

samy-melwan.vilhes@insa-rouen.fr

1. Introduction

Context. Time series anomaly detection (TSAD) aims to flag observations that deviate from expected patterns. Models must be **accurate**, **lightweight**, and **fast** at inference to operate in real time on edge devices.

Related works. Deep learning methods for unsupervised TSAD can involve signal **reconstruction** or **prediction**, **latent space modeling**, or the **generation of synthetic anomalies**. Most of the time, when dealing with observation x_t, we consider the w past observations, denoted as $x_{t-w:t}$.

Contributions. We propose **PatchTrAD**, a **lightweight** anomaly detection model that leverages the efficiency of **patch-based Transformers**, the benefits of **channel independence**, and the robustness of **reconstruction-based** approaches for TSAD.

2. PatchTrAD Overview

Channel independence refers to treating each modality as an independent signal within a model, without integrating information across modalities. Empirical studies show it maintains performance while requiring less training data.

Each univariate signal $x^{(m)} \in \mathbb{R}^w$ is divided into **patches** of fixed length P_{len}. A stride S determines the non-overlapping region between patches. We also pad the end of the sequence by repeating its last value S times before patching. The number of patches is:

$$P_{\text{num}} = \left\lfloor \frac{w - P_{\text{len}}}{S} \right\rfloor + 2.$$

Each patch acts as a token (analogous to LLMs), and the target observation always lies in the final patch.

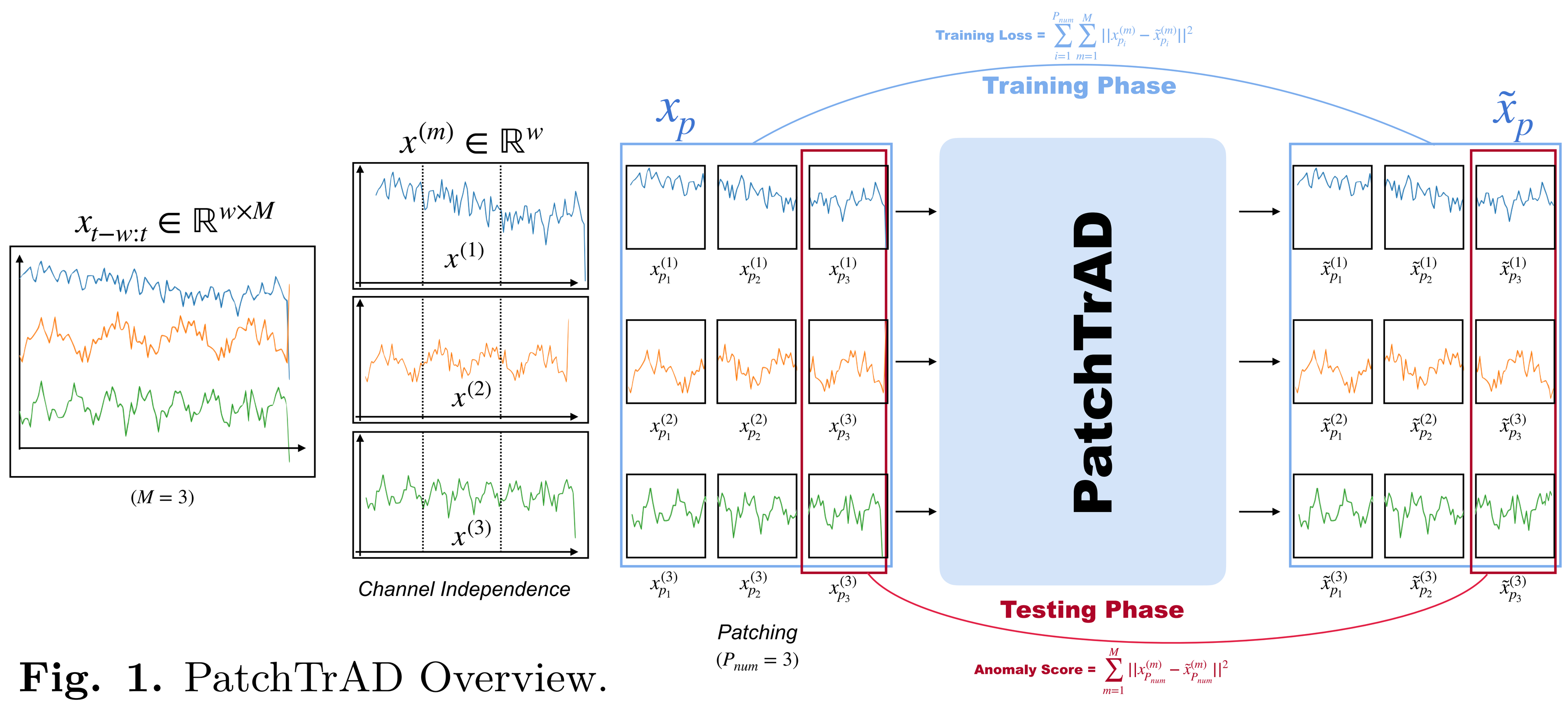
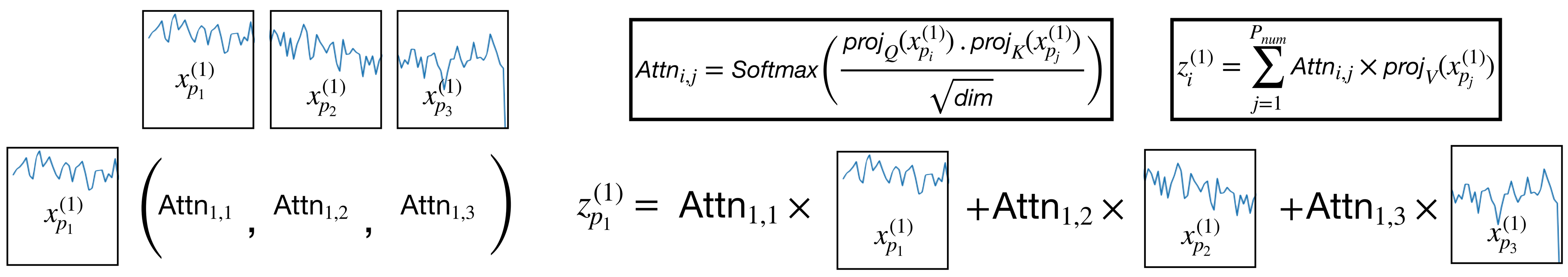


Fig. 1. PatchTrAD Overview.

3. Transformer attention mechanism of PatchTrAD

PatchTrAD consists of a Transformer Encoder, where the time dimension is represented by P_{num}.



$\text{Attn}_{i,j}$ represents how much information the model extracts from patch j to construct the representation of patch i. Each patch is projected into a D-dimensional space and we consider positional encoding to model temporal dependencies. Thus, $z_{p_1}^{(1)} \in \mathbb{R}^D$.

4. Training and Inference

Projection heads map embedded patches $z_{p_i}^{(m)}$ to reconstructed patches $\tilde{x}_{p_i}^{(m)} \in \mathbb{R}^{P_{\text{len}}}$. PatchTrAD is trained to reconstruct all input patches by minimizing the following loss::

$$\mathcal{L}_{\text{train}} = \sum_{i=1}^{P_{\text{num}}} \sum_{m=1}^{M} \|x_{p_i}^{(m)} - \tilde{x}_{p_i}^{(m)}\|^2.$$

As x_t lies in the last patch, the **anomaly score** is:

$$\mathcal{A}(x_t) = \sum_{m=1}^{M} \|x_{P_{\text{num}}}^{(m)} - \tilde{x}_{P_{\text{num}}}^{(m)}\|^2,$$

where $\mathcal{A}(x_t)$ is compared to a **predefined threshold** τ. The decision rule is:

$$x_t = \begin{cases} \text{abnormal} & \text{if } \mathcal{A}(x_t) \geq \tau \\ \text{normal} & \text{otherwise} \end{cases}$$

5. Results

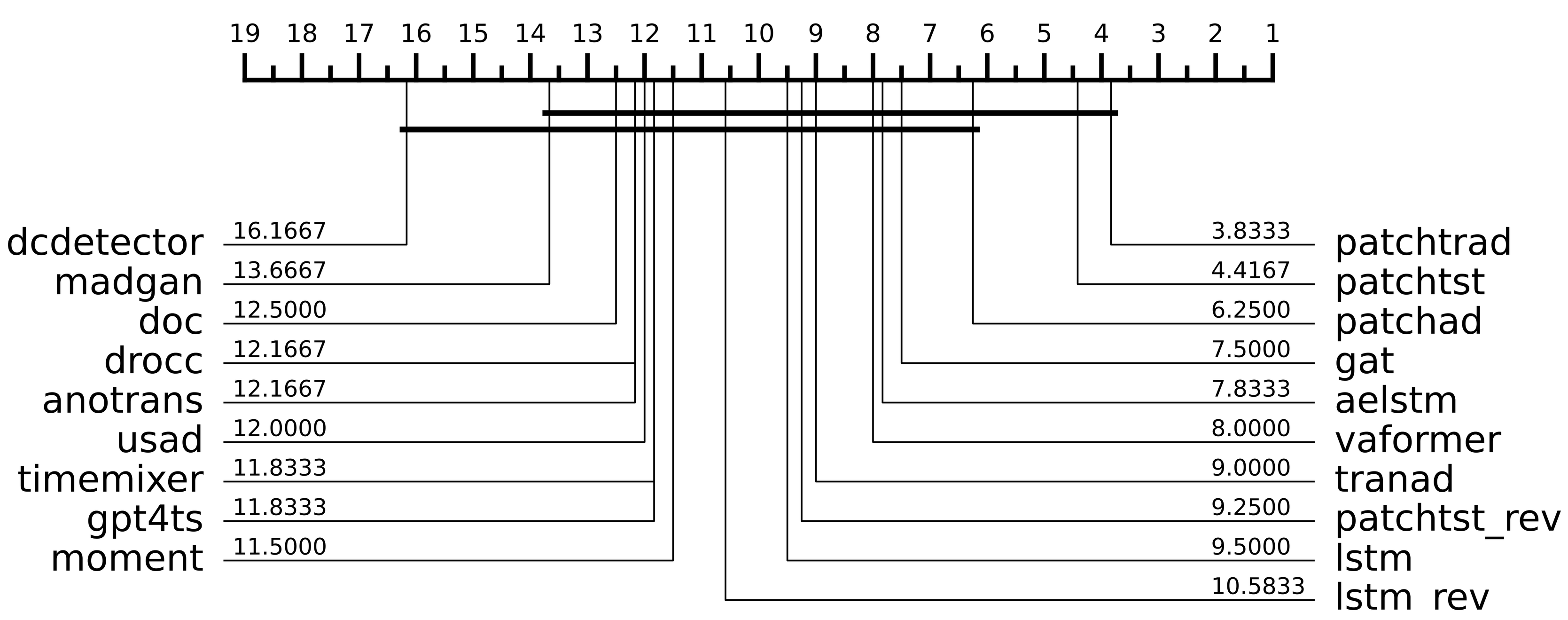


Fig. 2. Critical difference diagram based on the post-hoc Nemenyi test with $\alpha = 5\%$. Bars connect models that are not significantly different. Higher-ranked methods appear toward the upper right.

To ensure a fair and interpretable comparison, we evaluate using the ROC-AUC score, effective for evaluating models across datasets with varying class imbalance. This metric eliminates the need for threshold selection, as it is handled intrinsically. We compare PatchTrAD to 18 state-of-the-art models across 6 univariate and multivariate datasets. For each dataset, training is performed only on normal data, while testing includes both normal and anomalous observations.

6. Conclusion

PatchTrAD is a Transformer-based model leveraging patches for TSAD based on reconstruction error. It competes with state-of-the-art approaches and performs well across diverse datasets, both univariate and multivariate. PatchTrAD remains efficient during inference, making it suitable for a wide range of TSAD problems.

References

[1] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6):100994, 2024.

[3] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006.

[4] Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting, 2023.