PatchTrAD: A Patch-Based Transformer focusing on Patch-Wise Reconstruction Error for Time Series Anomaly Detection

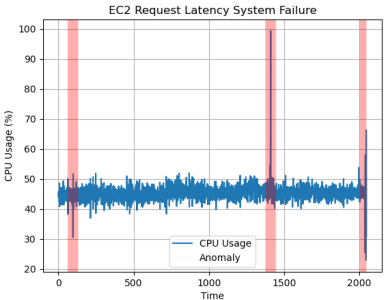
Vilhes Samy-Melwan
Joint work with Gasso Gilles and Mokhtar Z. Alaya

1. Time Series Anomaly Detection (TSAD)

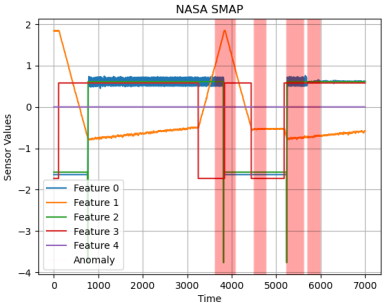
Introduction

What is TSAD?

TSAD refers to the task of identifying whether new observations from a data stream significantly differ from normal behaviour.



Univariate signal

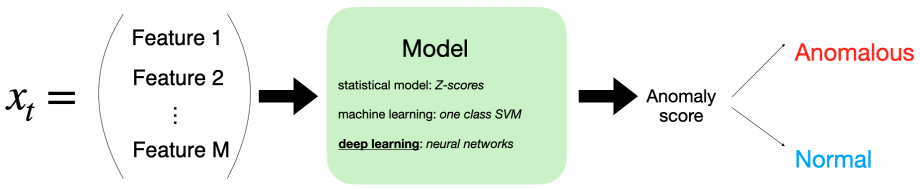


Multivariate signal

1. Time Series Anomaly Detection (TSAD)

Unsupervised Learning for TSAD

- Anomalies are **rare events**. We train the model exclusively on normal (clean) samples.
- At inference time, when anomalies may occur, the model outputs an **anomaly score**.
- Higher scores indicate higher likelihood of abnormal behavior.



Where M denotes the number of modalities and $x_t \in \mathbb{R}^M$ the observation at time t.

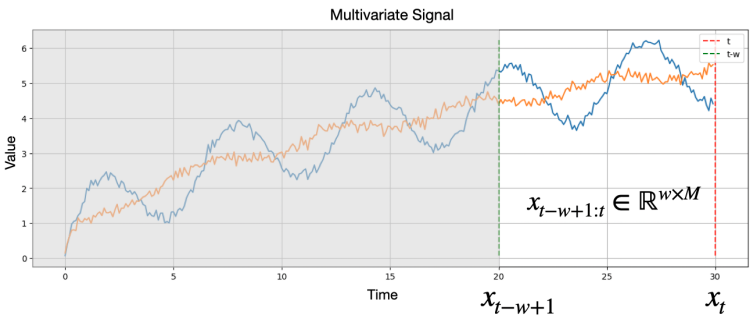
In the following, our attention is focused on deep learning–based models.

2. Deep Learning for TSAD: related works

Sliding window

Sliding window

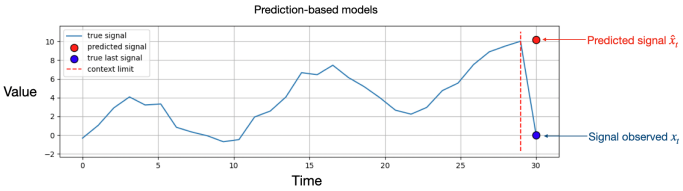
When dealing with $x_t \in \mathbb{R}^M$, we do not consider the whole past but w previous observations.



- w: sliding window length,
- $x_{t-w+1:t} \in \mathbb{R}^{w \times M}$: signals from time $t - w + 1$ to t.

2. Deep Learning for TSAD: related works

Prediction-based models



- f_θ: predictive model
- $\hat{x}_t = f_\theta(x_{t-w+1:t-1}) \in \mathbb{R}^M$: model's prediction of the true signal x_t given the context.

Anomaly score

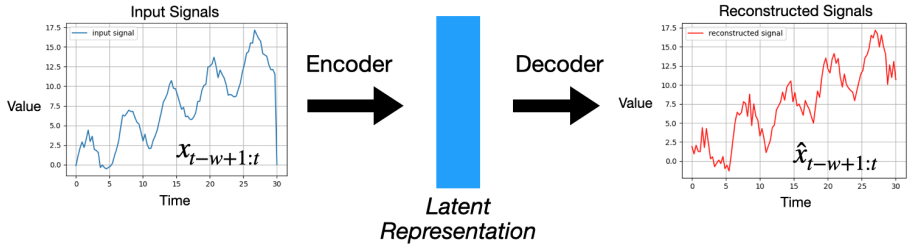
The anomaly score of prediction-based models corresponds to the prediction error given by

$$\text{anomaly score} = ||\hat{x}_t - x_t||^2.$$

According to the model, the higher the anomaly score, the more likely x_t is abnormal.

2. Deep Learning for TSAD: Related works

Reconstruction-based models



$$\hat{x}_{t-w+1:t} = \text{Decoder}(\text{Encoder}(\hat{x}_{t-w+1:t})) \in \mathbb{R}^{w \times M}.$$

Anomaly score

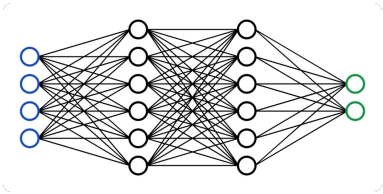
The anomaly score of reconstruction-based models corresponds to the reconstruction error given by

$$\text{anomaly score} = ||\hat{x}_{t-w+1:t} - x_{t-w+1:t}||^2.$$

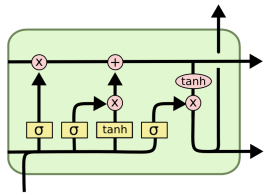
According to the model, the higher the anomaly score, the more likely x_t is abnormal.

2. Deep Learning for TSAD: related works

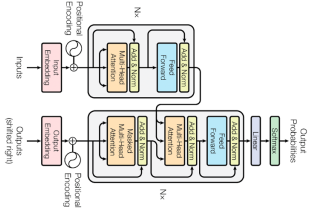
Architectures



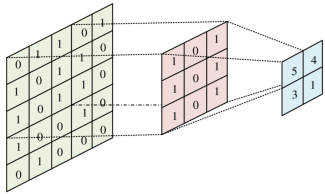
MLP (Zhong et al. 2024)



LSTM (Malhotra et al. 2016)



Transformer (Nie et al. 2023)



CNN (Ismail Fawaz et al. 2020)

2. Deep Learning for TSAD: related works

Benchmarks and experimental protocol

Dataset	# Feats	Train Size	Test Size	% Anomaly Test
NYC Taxi	1	5570	4750	0.11
EC2	1	1984	2049	0.15
SWAT	51	495000	449919	12.13
SMD	38	708405	708420	4.16
MSL	55	58317	73729	10.48
SMAP	25	140825	444035	12.85

Datasets Statistics:

1. Train a model on a training set containing only normal observations.
2. Evaluate its performance on a test set, which includes both normal and anomalous samples, using the ROC-AUC metric.

Why the ROC-AUC?

The ROC-AUC provides a robust estimation and ranking of classifier performance across different class imbalances (Richardson et al. 2024).

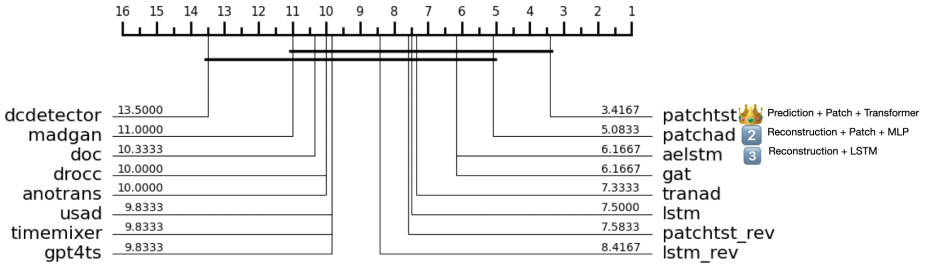
2. Deep Learning for TSAD: related works

Experiments: ROC-AUC scores (red: first, blue: second, green: third)

	Dataset	NYC-Taxi	EC2	MSL	SWaT	SMAP	SMD
Category	**Model**				**ROC-AUC**		
Others	DC-Detector	0.498	0.827	0.537	0.435	0.566	0.530
	AnomalyTransformer	0.491	0.994	0.553	0.819	0.621	0.678
	DOC	0.704	0.804	0.538	0.404	0.634	0.766
	DROCC	0.529	0.886	0.593	0.751	0.705	0.638
Prediction-based	PatchTST-revin	0.552	0.999	0.626	0.233	0.537	0.873
	LSTM-revin	0.646	0.998	0.627	0.238	0.586	0.858
	LSTM	0.511	0.999	0.595	0.842	0.604	0.833
	GAT	0.689	0.999	0.617	0.816	0.646	0.820
	PatchTST	0.696	0.999	0.626	0.843	0.622	0.882
Reconstruction-based	MADGAN	0.782	0.011	0.460	0.791	0.568	0.708
	USAD	0.669	0.977	0.684	0.255	0.547	0.605
	TimeMixer	0.523	0.942	0.681	0.235	0.536	0.900
	GPT4TS	0.272	0.954	0.699	0.235	0.544	0.890
	TranAD	0.551	0.967	0.644	0.815	0.581	0.884
	AE-LSTM	0.716	0.998	0.612	0.840	0.618	0.828
	PatchAD	0.972	0.998	0.622	0.822	0.671	0.818

2. Deep Learning for TSAD: related works

Experiments: critical difference diagram for ROC-AUC scores using the post-hoc Nemenyi test with $\alpha = 5\%$, where better-ranked methods appear on the upper right.

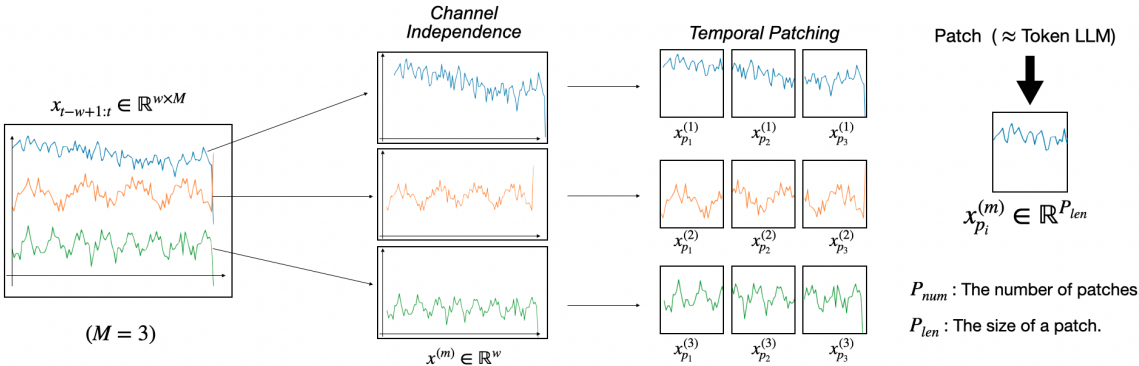


Conclusion: PatchTST (Nie et al. 2023), a prediction-based model leveraging transformers (Vaswani et al. 2017), patching, and channel independence, achieves the best performance. The second (Zhong et al. 2024) and third (Malhotra et al. 2016) models are reconstruction-based.

Our Idea (PatchTrAD): combine the strengths of patch-based transformers considering channel independence and reconstruction-based approaches for TSAD.

3. PatchTrAD

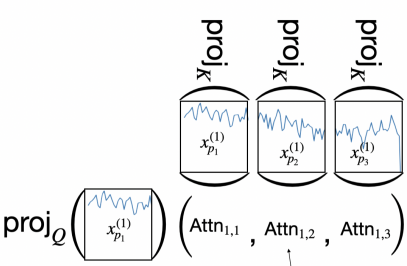
Channel Independence and Patching



Channel independence means that each patch contains information from a single modality without cross-modality sharing. Empirical studies show that this design improves model robustness while maintaining performance (Han, Ye, and Zhan 2023).

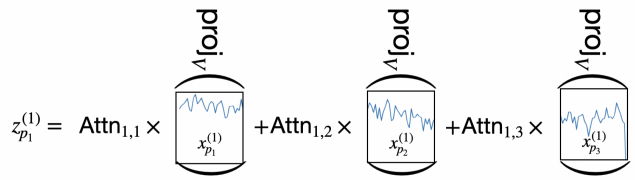
3. PatchTrAD

Transformer Attention Mechanism



$$Attn_{i,j} = Softmax\left(\frac{proj_Q(x_{p_i}^{(1)})^T \cdot proj_K(x_{p_j}^{(1)})}{\sqrt{dim}}\right)$$

$$z_i^{(1)} = \sum_{j=1}^{P_{num}} Attn_{i,j} \times proj_V(x_{p_j}^{(1)})$$



dim: the dimension of the model

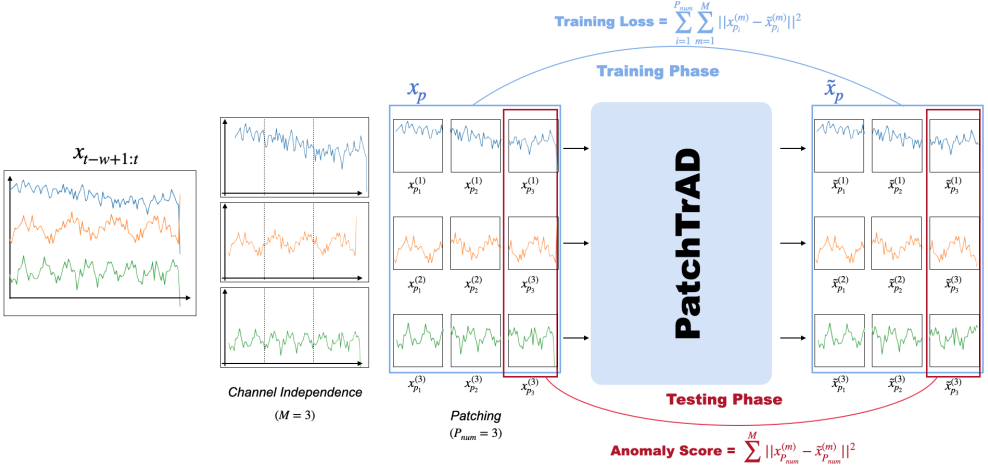
$proj_Q : \mathbb{R}^{P_{len}} \mapsto \mathbb{R}^{dim}$

$proj_K : \mathbb{R}^{P_{len}} \mapsto \mathbb{R}^{dim}$

$proj_V : \mathbb{R}^{P_{len}} \mapsto \mathbb{R}^{dim}$

3. PatchTrAD

Training strategy and inference



By construction, the test observation x_t always belongs to the last patch of each modality. Therefore, during inference, we focus on the error of this final patch.

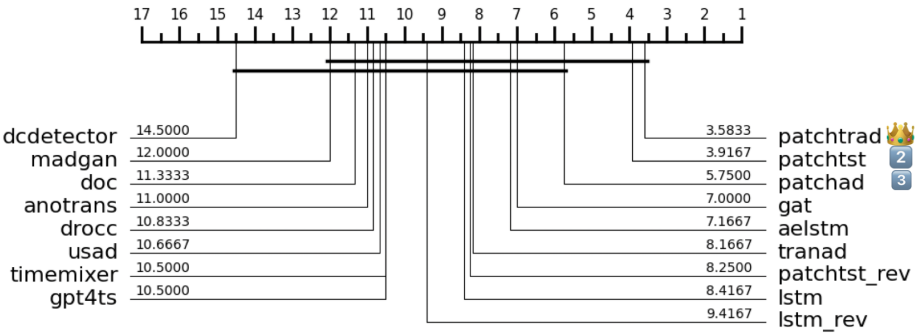
3. PatchTrAD

Final Results: ROC-AUC scores (red: first, blue: second, green: third)

	Dataset	NYC-Taxi	EC2	MSL	SWaT	SMAP	SMD
Category	**Model**	**ROC-AUC**					
Others	DC-Detector	0.498	0.827	0.537	0.435	0.566	0.530
	AnomalyTransformer	0.491	0.994	0.553	0.819	0.621	0.678
	DOC	0.704	0.804	0.538	0.404	0.634	0.766
	DROCC	0.529	0.886	0.593	0.751	0.705	0.638
Predictions-based	PatchTST-revin	0.552	0.999	0.626	0.233	0.537	0.873
	LSTM-revin	0.646	0.998	0.627	0.238	0.586	0.858
	LSTM	0.511	0.999	0.595	0.842	0.604	0.833
	GAT	0.689	0.999	0.617	0.816	0.646	0.820
	PatchTST	0.696	0.999	0.626	0.843	0.622	0.882
Reconstruction-based	MADGAN	0.782	0.011	0.460	0.791	0.568	0.708
	USAD	0.669	0.977	0.684	0.255	0.547	0.605
	TimeMixer	0.523	0.942	0.681	0.235	0.536	0.900
	GPT4TS	0.272	0.954	0.699	0.235	0.544	0.890
	TranAD	0.551	0.967	0.644	0.815	0.581	0.884
	AE-LSTM	0.716	0.998	0.612	0.840	0.618	0.828
	PatchAD	0.972	0.998	0.622	0.822	0.671	0.818
	PatchTrAD (ours)	0.922	0.999	0.661	0.845	0.660	0.869

3. PatchTrAD

Experiments: critical difference diagram for ROC-AUC scores using the post-hoc Nemenyi test with $\alpha = 5\%$, where better-ranked methods appear on the upper right.



4. Conclusion

- We introduced a Patch-Based Transformer leveraging reconstruction error,
- Effective for both univariate and multivariate signal monitoring,
- Efficient and lightweight at inference,
- Achieves competitive performance compared to SOTA methods,
- PatchTrAD shows strong potential for addressing future industrial TSAD challenges.

Future Work: Currently developing foundation models for zero-shot time series anomaly detection.

Thank You!

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

litis

ALLIANCE SORBONNE UNIVERSITÉ

utc Recherche LMAC

anr

References I

- Julien Audibert et al. (2020). *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '20. Virtual Event, CA, USA: Association for Computing Machinery, pp. 3395–3404. ISBN: 9781450379984. DOI: 10.1145/3394486.3403392. URL: https://doi.org/10.1145/3394486.3403392.
- Sachin Goyal et al. (2020). *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.
- Lu Han, Han-Jia Ye, and De-Chuan Zhan (2023). *The Capacity and Robustness Trade-off: Revisiting the Channel Independent Strategy for Multivariate Time Series Forecasting*. arXiv: 2304.05206 [cs.LG]. URL: https://arxiv.org/abs/2304.05206.
- Hassan Ismail Fawaz et al. (Sept. 2020). *Data Mining and Knowledge Discovery* 34.6, pp. 1936–1962. ISSN: 1573-756X. DOI: 10.1007/s10618-020-00710-y. URL: http://dx.doi.org/10.1007/s10618-020-00710-y.
- Taesung Kim et al. (2022). *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=cGDAkQo1C0p.

References II

- Alexander Lavin and Subutai Ahmad (Dec. 2015). *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE. DOI: 10.1109/icmla.2015.141. URL: http://dx.doi.org/10.1109/ICMLA.2015.141.
- Dan Li et al. (2019). *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*. Ed. by Igor V. Tetko et al. Cham: Springer International Publishing, pp. 703–716. ISBN: 978-3-030-30490-4.
- Pankaj Malhotra et al. (2016). *LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection*. arXiv: 1607.00148 [cs.AI]. URL: https://arxiv.org/abs/1607.00148.
- Yuqi Nie et al. (2023). *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Jbdc0vTOcol.
- Eve Richardson et al. (2024). *Patterns* 5.6, p. 100994. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2024.100994. URL: https://www.sciencedirect.com/science/article/pii/S2666389924001090.

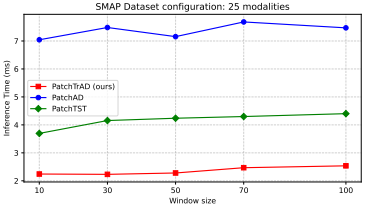
References III

- Lukas Ruff et al. (Oct. 2018). *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4393–4402. URL: https://proceedings.mlr.press/v80/ruff18a.html.
- Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings (Feb. 2022). *Proc. VLDB Endow.* 15.6, pp. 1201–1214. ISSN: 2150-8097. DOI: 10.14778/3514061.3514067. URL: https://doi.org/10.14778/3514061.3514067.
- Ashish Vaswani et al. (2017). *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 6000–6010. ISBN: 9781510860964.
- Shiyu Wang et al. (2024). *TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting*. arXiv: 2405.14616 [cs.LG]. URL: https://arxiv.org/abs/2405.14616.
- Jiehui Xu et al. (2022). *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=LzQQ89U1qm_.

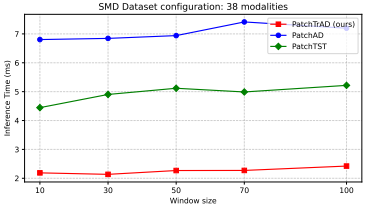
References IV

- Yiyuan Yang et al. (2023). *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '23. Long Beach, CA, USA: Association for Computing Machinery, pp. 3033–3045. ISBN: 9798400701030. DOI: 10.1145/3580305.3599295. URL: https://doi.org/10.1145/3580305.3599295.
- Hang Zhao et al. (2020). *CoRR* abs/2009.02040. arXiv: 2009.02040. URL: https://arxiv.org/abs/2009.02040.
- Zhijie Zhong et al. (2024). *PatchAD: A Lightweight Patch-based MLP-Mixer for Time Series Anomaly Detection*. arXiv: 2401.09793 [cs.LG]. URL: https://arxiv.org/abs/2401.09793.
- Tian Zhou et al. (2023). *One Fits All:Power General Time Series Analysis by Pretrained LM*. arXiv: 2302.11939 [cs.LG]. URL: https://arxiv.org/abs/2302.11939.

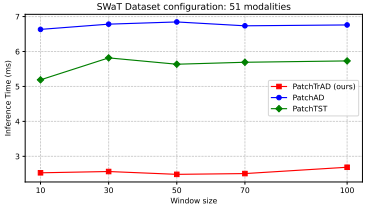
Inference speed comparison I



Soil Moisture Active Passive (SMAP) dataset 25 features.



Server Machine Dataset (SMD) 38 features.



Secure Water Treatment (SWaT) dataset 51 features.

Figure: Inference speed comparison of the 3 best models across 3 datasets.