Multivariate Time Series Anomaly Detection

UNIVERSIDAD
POLITÉCNICA
DE MADRID

POLITÉCNICA

Anomaly Detection on a Server Machine Dataset

Authors
Emanuele Emilio Alberti
Leandro Duarte
Ottavia Biagi

UPM — Data Mining & Time Series

Project Roadmap

Pipeline

Core Evaluation

Experiements results

Diagnostics

Studying our results

Methodology & Modelling

Designing the Mean Baseline, LSTM, and Transformer Autoencoders.

Context & Data

Why we chose machine-1-1 and how we prepared the data .

Context & Data

Problems and Dataset

- **Goal**: **Detect anomalies** in the **Server Machine Dataset (SMD)** (**28 servers**, **38 KPIs each**) without human supervision

- **Challenge**: The data is **non-stationary** (changing mean/variance) and **unlabeled** in the **training set**

- **Strategy**: Use **"Reconstruction-Based Detection",** train a model to learn "normal" behavior and flag what it cannot reproduce

Context & Data

Problem & Dataset

Dataset Information

Dataset name	Number of entities	Number of dimensions	Training set size	Testing set size
SMAP	55	25	135183	427617
MSL	27	55	58317	73729
SMD	28	38	708405	708420

OmniAnomaly / ServerMachineDataset / **train** /

smallcowbaby init

Name	Last comm
..	
machine-1-1.txt	init
machine-1-2.txt	init
machine-1-3.txt	init
machine-1-4.txt	init
machine-1-5.txt	init
machine-1-6.txt	init
machine-1-7.txt	init

- **28 Server Machine** Dataset (**SMD**)
- **Train-set**: only normal behaviour
- **Test-set**: normal + anomalies
- Labels at timestamp level only for test
- **38 KPIs** per **machine**
- Sampling every “few” seconds
- **Highly multivariate** & **non-stationary**

Context & Data

Machine Selection

We analyzed all **28 machines** to find the **best candidate**

Criterio 1: Stability
Bursty and irregular system behaviour

Criterion 2: Correlation
Correlation comparison across 28 machines and their 38 KPIs/Features

	machine	mean_corr
0	machine-1-1	0.457749
1	machine-1-2	0.113721
2	machine-1-3	0.243654
3	machine-1-4	0.266576
4	machine-1-5	0.304304
5	machine-1-6	0.277832
6	machine-1-7	0.250052
7	machine-1-8	0.346967
8	machine-2-1	0.212146
9	machine-2-2	0.252188

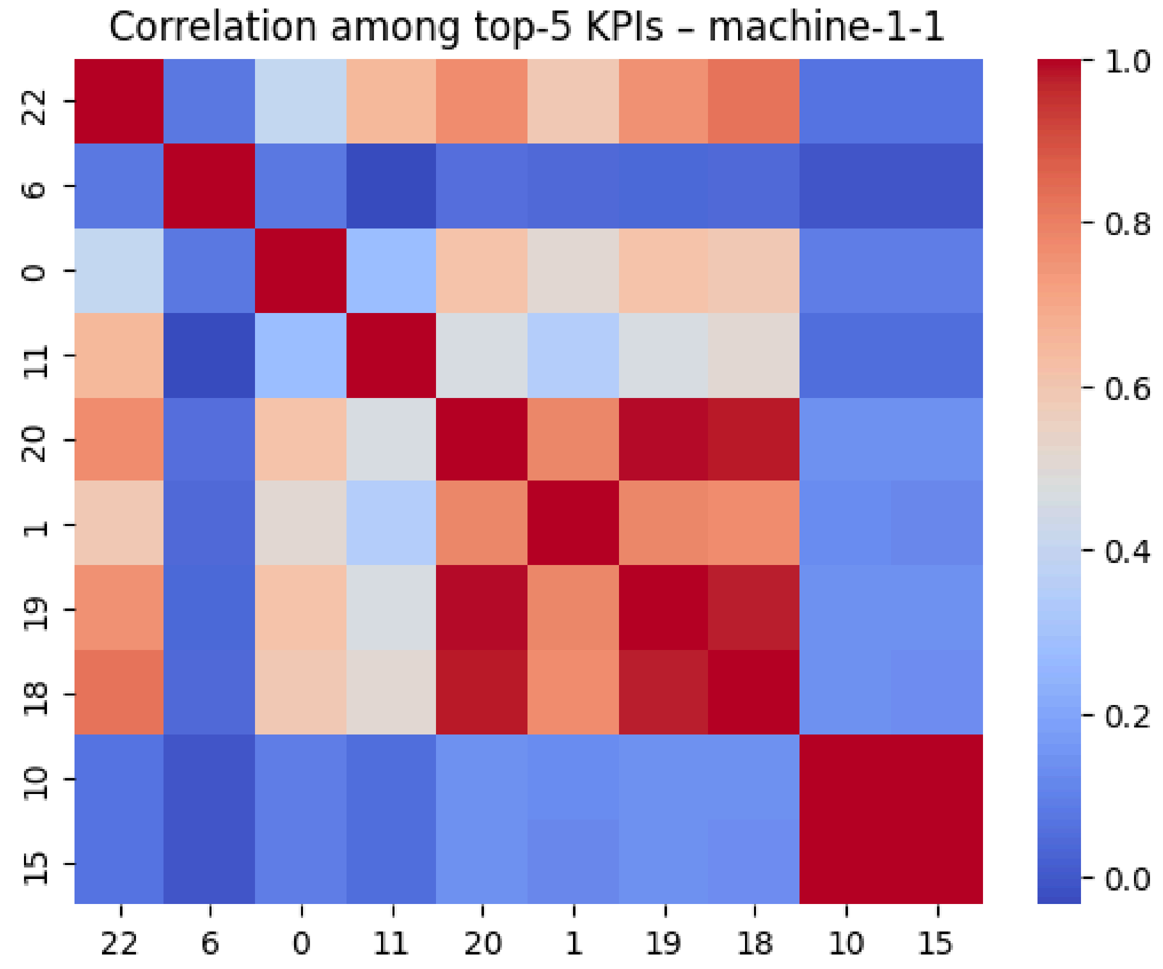
Context & Data
Variables Selection

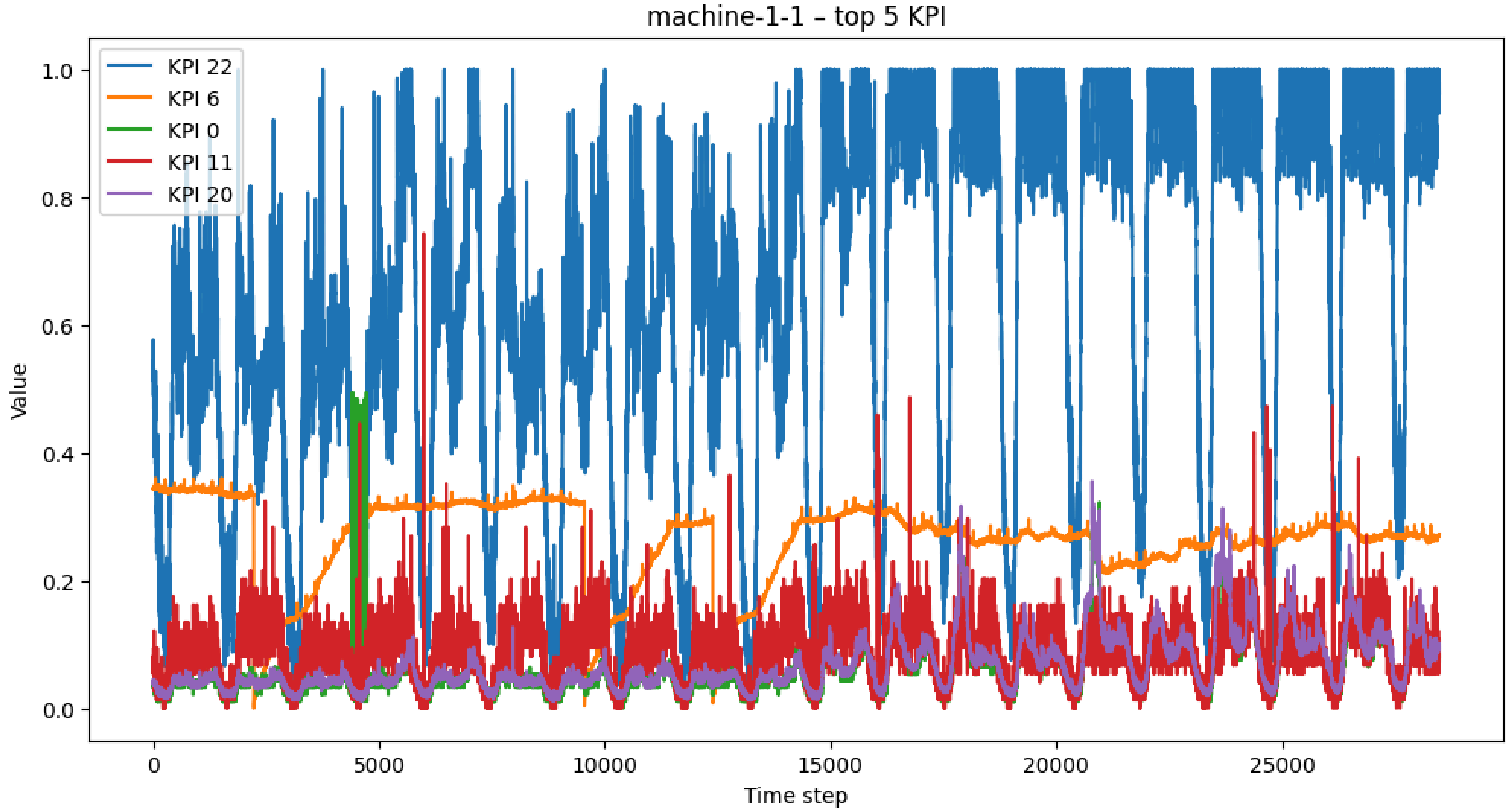
Machine 1-1 KPIs

- High but not maximal KPI variance
- Non-trivial correlation structure
- Suitable trade-off between complexity and interpretability

Machine 1-1's KPIs observations

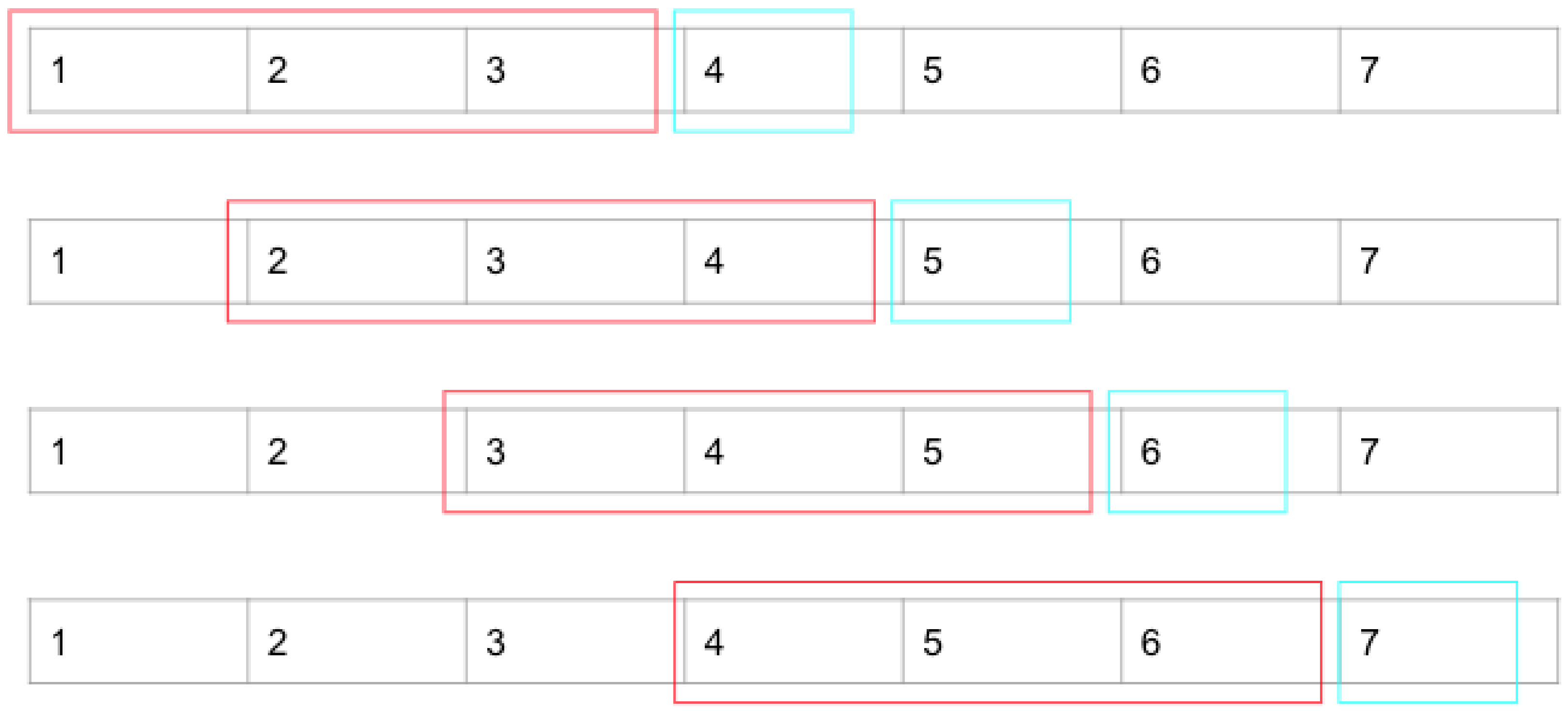
- Top-5 KPIs selected by variance → capture **global instability** and **bursty behaviour**
- Correlation heatmap reveals 2–3 clusters among the top-5 KPIs
- One low-correlated KPI (KPI 30)→ representative of an **independent subsystem**





Context & Data

Preprocessing Strategy



Normalization:

Min–max normalization applied → prevents large-scale KPIs from dominating MSE loss & preserves relative temporal patterns

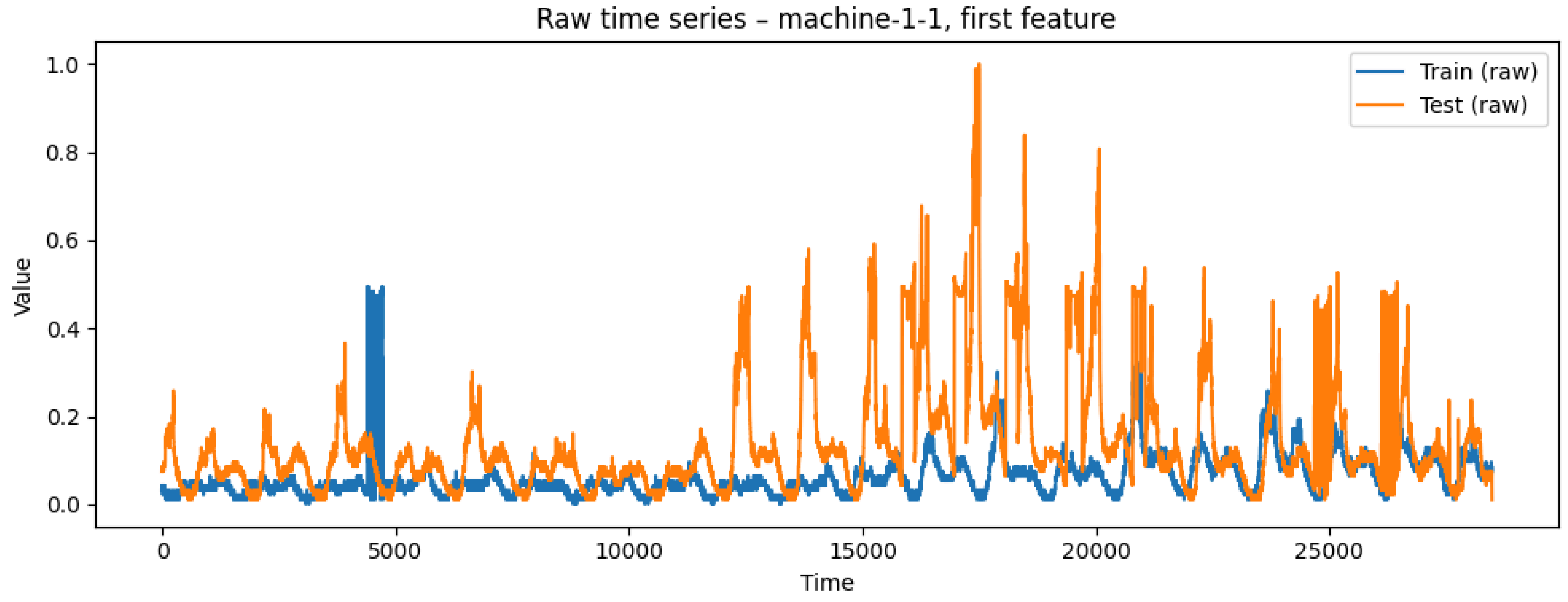
Sliding Windows:

Window sizes: K = 10, 30, 50

- Short windows → local bursts
- Long windows → periodicity and long-range dependencies

Key observations:

- System metrics exhibit rare but **high-magnitude events**
- Strongly non-Gaussian distributions
- Heavy tails and sudden spikes
- Long periods of stability interrupted by bursts



Methodology & Modeling

The models

Baseline (The "Sanity Check"): Predicts the global training mean for every timestamp. Simple, fast, non-temporal.

LSTM Autoencoder: Classical recurrent network. Captures sequences but often struggles with false positives.

Transformer Autoencoder: Modern attention-based architecture. Learns cross-feature dependencies.

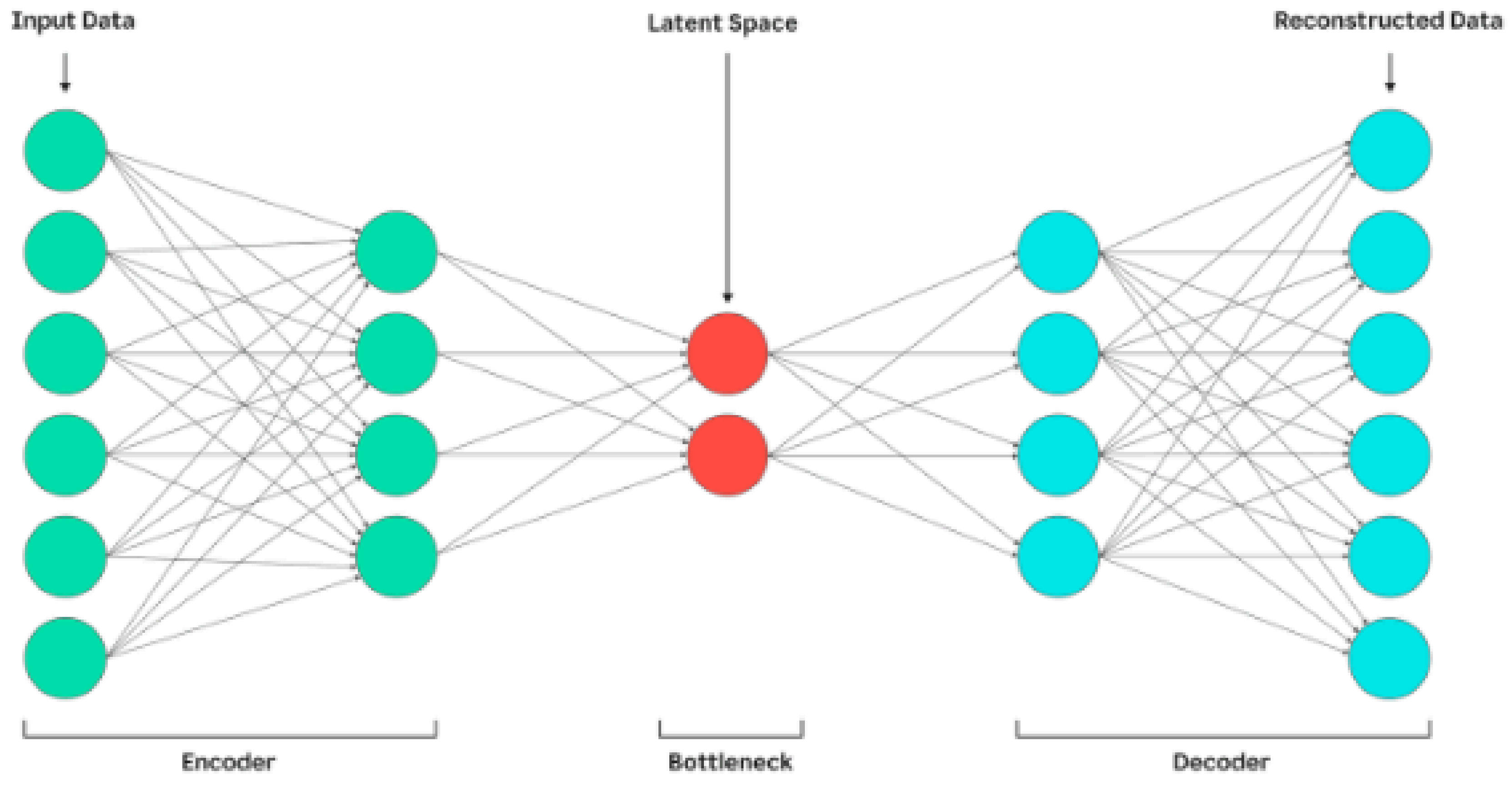


Figure 7: Generic autoencoder architecture (illustrative).

Design Choice:
We used **Huber Loss** for the Transformer to ignore outliers and stabilize training

$$Huber = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(y_i - \hat{y}_i)^2 \qquad |y_i - \hat{y}_i| \le \delta$$

$$Huber = \frac{1}{n}\sum_{i=1}^{n}\delta\left(|y_i - \hat{y}_i| - \frac{1}{2}\delta\right) \qquad |y_i - \hat{y}_i| > \delta$$

Methodology & Modeling

Thresholding

Robust thresholding is required to handle noisy and bursty reconstruction errors.

Scoring: Anomaly Score = Reconstruction Error (MSE).

$$s_t = \mathrm{MSE}(X_t, \hat{X}_t)$$

Threshold: A timestamp is an anomaly if its error is in the top 0.5% of training errors (99.5th percentile).

$$\hat{y}_t = \mathbb{I}[s_t > \tau]$$

Refinement: We applied Score Smoothing (window of 5) to stop the model from panicking over single noisy spikes .

$$\tilde{s}_t = \frac{1}{5} \sum_{i=t-2}^{t+2} s_i$$

Core Evaluation

Main Model Results - machine-1-1 - smoothed threshold

Baseline and Neural Models – Quantitative Comparison

- Mean baseline: F1 = 0.52 (surprisingly strong)
- LSTM AE: high recall, low precision
- Transformer (LN): balanced

Model	Precision	Recall	F1	ROC-AUC	PR-AUC
Mean reconstruction	**0.568**	0.472	**0.516**	**0.911**	**0.577**
LSTM AE (MSE)	0.187	**0.99**	0.314	0.878	0.516
Transformer AE (LN, MSE, 20 ep.)	0.339	0.449	0.386	0.866	0.432
Transformer AE (LN, MSE, 40 ep.)	**0.356**	0.541	**0.43**	0.878	0.463
Transformer AE (LN, Huber, 20 ep.)	0.349	0.486	0.406	0.867	0.44

Mean baseline sets a very strong reference: simple global mean → F1 0.52

Core Evaluation

Effect of the window size - Machine 1-1

- Transformer improves significantly from K=10 → K=30
- K=50 degrades (noise + overfitting)
- Transformers require sufficient context but degrade without meta-learning

K	Model	Precision	Recall	F1	ROC-AUC
10	LSTM AE	0.185	0.984	0.312	0.872
10	Transformer (LN)	0.352	0.448	0.394	0.844
30	LSTM AE	0.184	1	0.31	0.893
30	Transformer (LN)	0.374	0.536	**0.44**	0.896
50	LSTM AE	0.212	0.957	**0.347**	0.896
50	Transformer (LN)	0.347	0.501	0.41	0.881

Increasing window size improves attention models up to K = 30.
Larger windows degrade performance without meta-learning, consistent with findings in TranAD

Core Evaluation

Main Model Results - machine-1-1 - K=10 - smoothed threshold

- **Transformer (LN):** balanced
- **Transformer LN + Huber + Dropout, 40 epoch, K=10:** best learned model → **F1 = 0.53**, **ROC-AUC = 0.915**

Model	Precision	Recall	F1	ROC-AUC	PR-AUC
Transformer (LN + Dropout + Huber)	0.392	0.824	**0.531**	**0.915**	0.6
Transformer (High Dropout)	0.36	0.475	0.41	0.882	0.457
Transformer (Mixed PosEnc)	0.354	0.474	0.406	0.872	0.448
Transformer (NormOut)	0.322	0.437	0.371	0.754	0.375

Mean baseline sets a very strong reference: simple global mean → F1 0.52

Core Evaluation

Main Model Results - machine-1-1

Best Transformer and Stabilisation Strategies

Model Variant	Precision	Recall	F1	ROC-AUC	PR-AUC
Transformer AE (LN + Huber + Dropout, K10, 40 ep.)	**0.392**	0.824	**0.531**	0.915	0.46
Stabilised Transformer (Huber + Dropout + Sigmoid + clipping)	**0.416**	0.329	0.368	0.672	0.367
+ Segment cleaning	0.415	0.326	0.365	0.672	0.367
+ Self-conditioned scoring	0.415	0.339	**0.373**	0.676	0.371
Semi-adversarial training	0.359	0.284	0.317	0.729	0.31
Semi-adv. + self-cond. + cleaning	0.36	0.288	0.32	0.732	0.313

Stabilisation techniques completely remove numerical explosions, but do not automatically improve detection metrics

Core Evaluation

Multi-Machine Generalisation

Performance when training jointly on 4 machines:

- machine-1-1: F1 = 0.405
- machine-1-2: F1 = 0.361
- machine-1-3: F1 = 0.062
- machine-2-1: F1 = 0.083

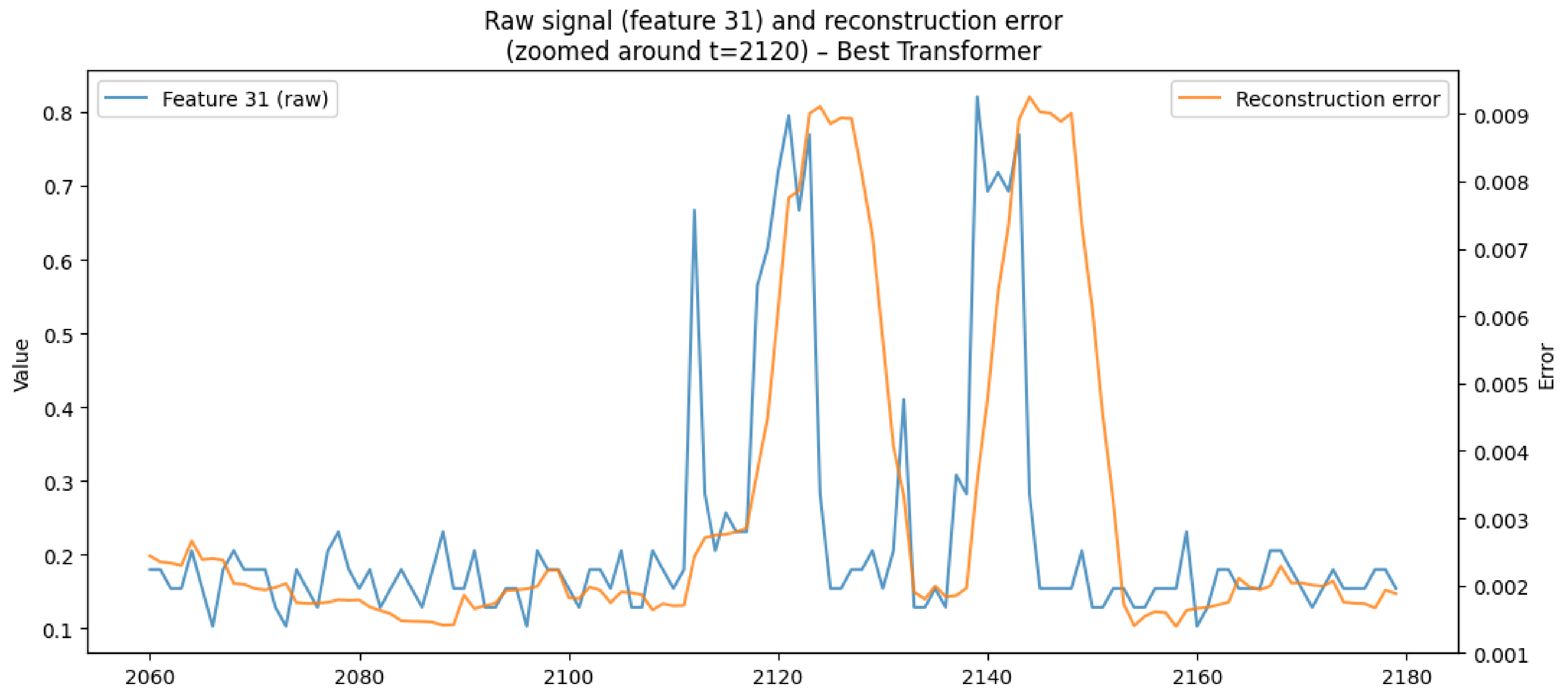
Metric	Average across
Precision	**0.497**
Recall	0.306
F1	**0.309**
ROC-AUC	0.741

This confirms that the latent space is machine-specific; shared training collapses

Family	Best Model	F1 Score	Notes
Baseline	Mean Reconstruction	**0.52**	Surprisingly strong
LSTM	LSTM AE (K=10)	0.31	Very high recall, low prec.
Transformer	Transformer (LN + Dropout + Huber, 40 epoc, K=10)	**0.53**	**Best learned model**
Transformer MSE	Transformer LN (40 epoc, K=30)	0.44	Best MSE-based transformer

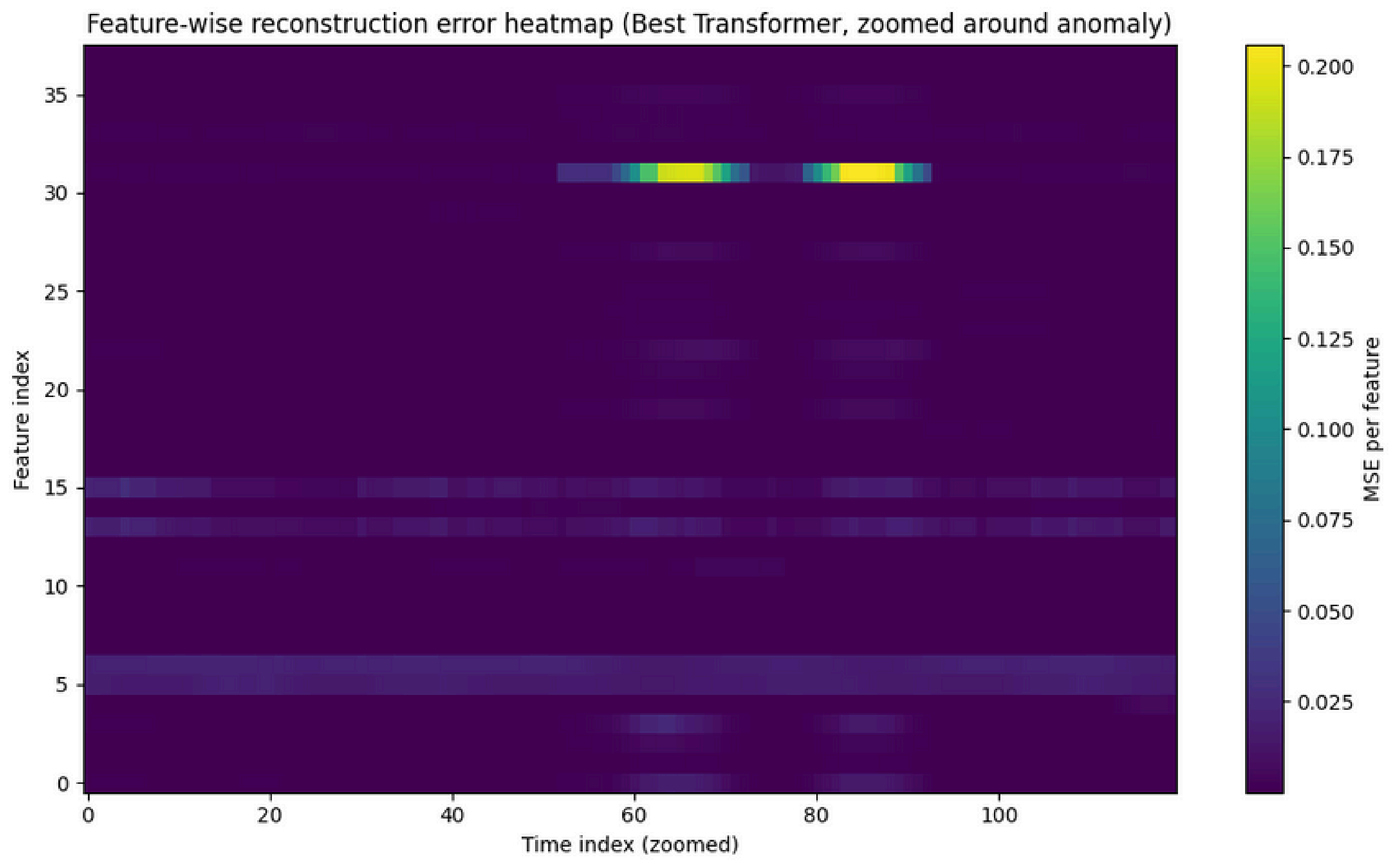
Diagnostics

Qualitative Diagnostics - Transformer AE (LN + Dropout + Huber)



- Error spikes align with ground-truth anomalies
- Anomalies affect a limited subset of KPIs, not the full system

- Feature-wise heatmaps identify which KPIs drive the anomaly
- Dominant features correspond to disk / network throughput metrics

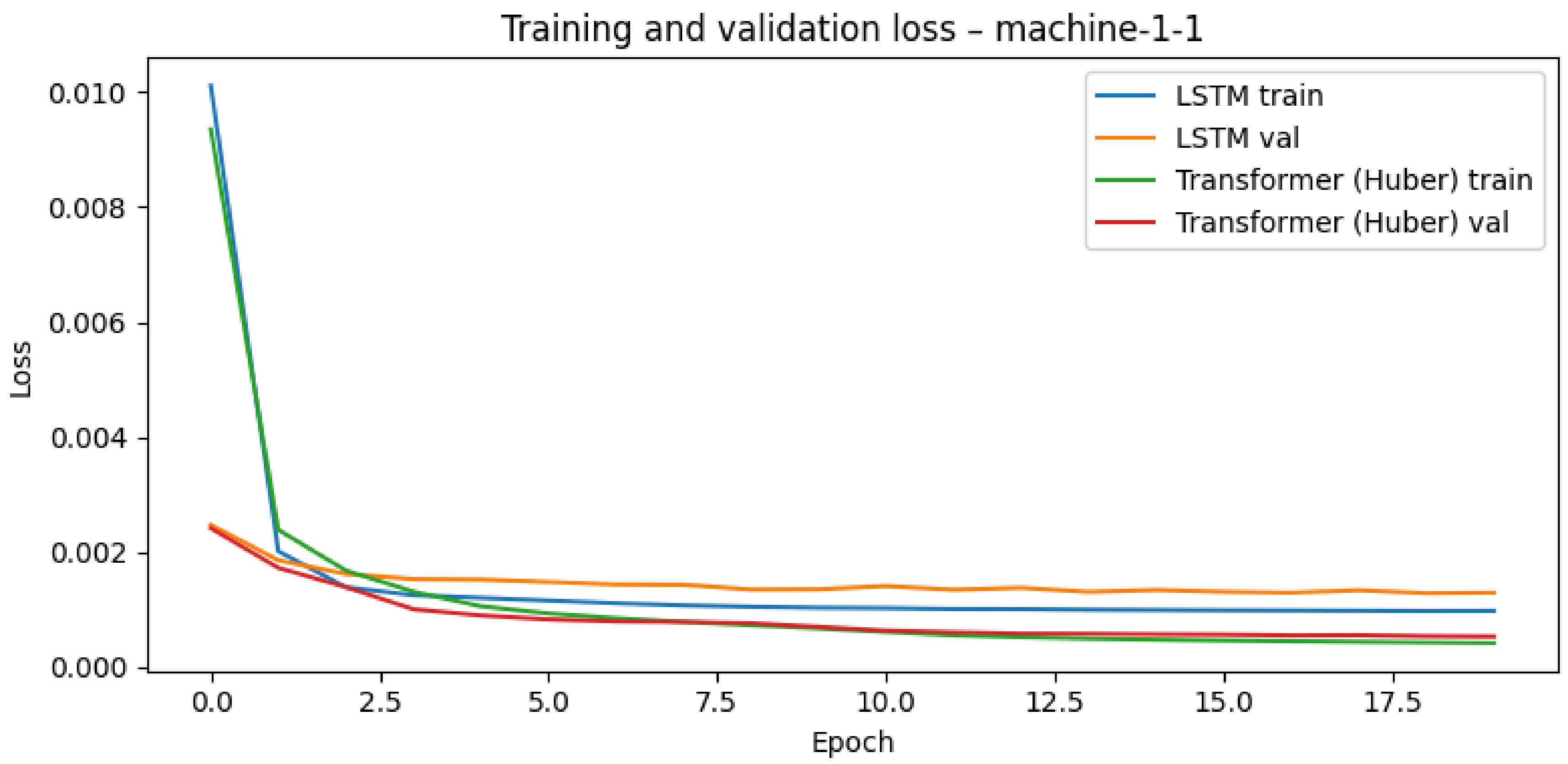


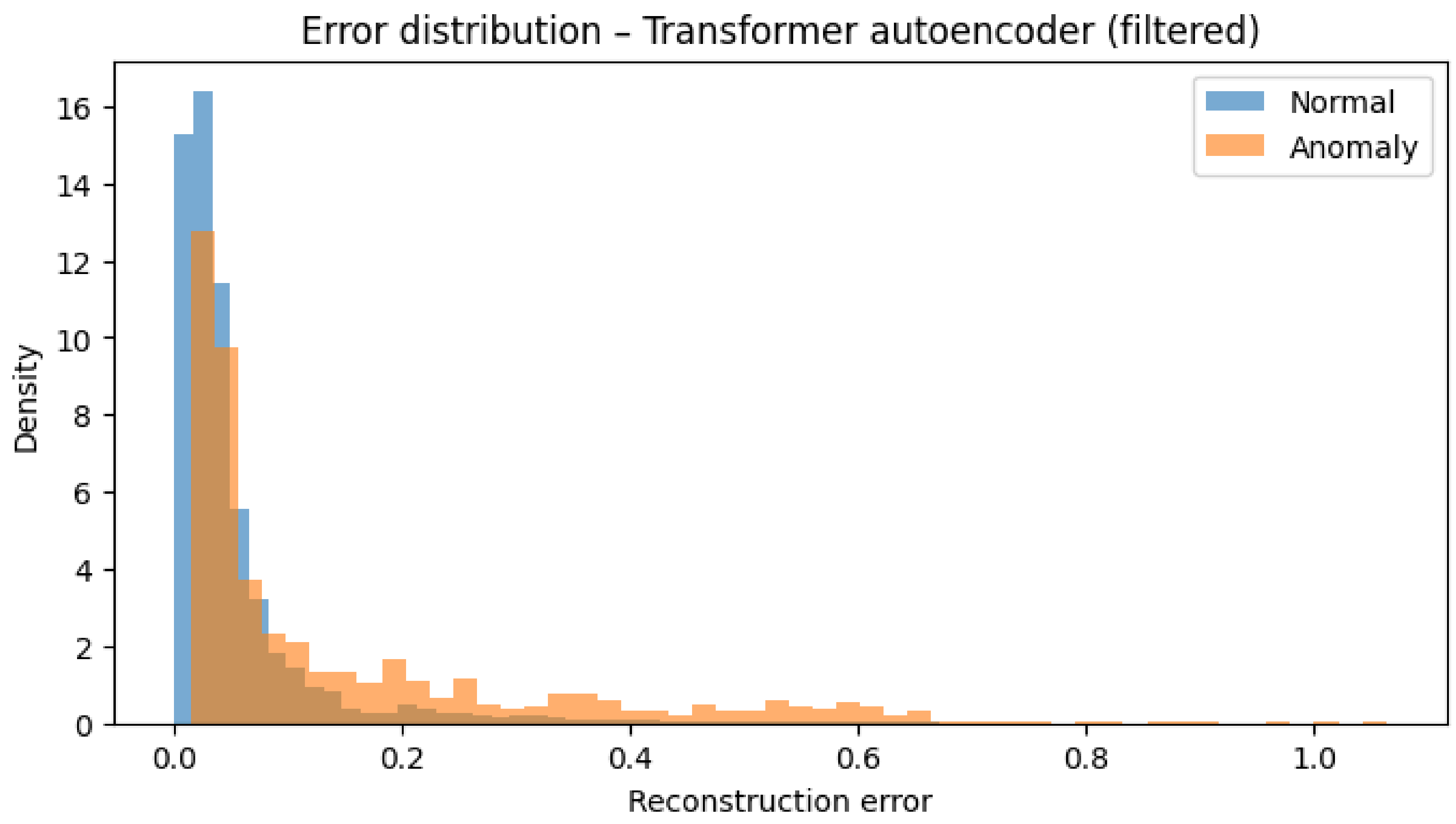
All models operate on 38 KPIs.
Interpretation focuses on a sparse subset of features.

Diagnostics

Training Behaviour & Error Distribution - Transformer (LN + Dropout + Huber)

- Transformer trains slower but reaches lower validation loss
- Overlap between normal and abnormal scores → **limited max F1 achievable**
- Outliers indicate numerical instabilities → motivate Huber loss





Diagnostics

Comparison with State-of-the-Art & Final Takeaways

- Huber transformer with dropout and layernorm is best robust model
- Mean baseline sets high reference
- Feature-level diagnosis is a strength

Model	Window Size	Meta-Learning	Adversarial	Self-Conditioning	Training Scope	F1 Score
OmniAnomaly	100	No	No	No	28 machines	0.94
TranAD	100–500	Yes	Yes	Yes	28 machines	**0.96**
Transformer AE (LN + Drpout + Huber)	10	No	No	No	1 machine	**0.53**

Our model is lightweight (1 machine, K=10, no meta-learning), suitable for real-time costrains. While SOTA methods rely on long context windows and architectural stabilisation.

Diagnostics

Conclusion

Overall Result

A lightweight but interpretable anomaly detection pipeline that:

- Matches strong statistical baselines
- Provides diagnostic insight
- Lays the foundation for deeper architectures like TranAD

Best Performing Model

Transformer AE LN + Huber loss + Dropout + K=10

Best learned model: F1 ≈ 0.53

Current Limitations

Temporal context limited (K=10–50 vs 500 in SOTA)

Architectural complexity limits the perferomaces and stabilization

Cannot generalize well across machines

Sensitive to numerical instabilities without robust losses

Conclusion

UNIVERSIDAD
POLITÉCNICA
DE MADRID

POLITÉCNICA

Thank You