Position: Quo Vadis, Unsupervised Time Series Anomaly Detection?

M. Saquib Sarfraz [1,2], Mei-Yen Chen [2], Lukas Layer [2], Kunyu Peng [1], Marios Koulakis [1]

[1] Karlsruhe Institute of Technology [2] Mercedes-Benz Tech Innovation

Position

The current state of machine learning scholarship in Timeseries Anomaly Detection (TAD) is plagued by the persistent use of:

- Flawed evaluation metrics
- Inconsistent benchmarking practices
- Overemphasis on complex models without real improvement

We advocate for a shift in focus from solely pursuing novel model designs to improving benchmarking practices, creating non-trivial datasets, and critically evaluating the utility of complex methods against simpler baselines.

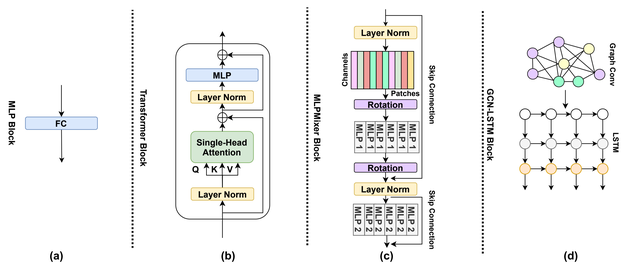
Simple Baselines

We've set simple baselines that challenge the necessity of complexity in state-of-the-art models.

- Sensor range deviation $f(\hat{\mathbf{x}}_t) = 0 \; if \; \hat{\mathbf{x}}_t \in [\min(\mathbf{X}), \max(\mathbf{X})] \; else \; 1$
- L2-norm $f(\hat{\mathbf{x}}_t) = \|\hat{\mathbf{x}}_t\|_2$
- Nearest Neighbor distance $f(\hat{\mathbf{x}}_t) = \min_{\mathbf{x} \in \mathbf{X}}(\|\hat{\mathbf{x}}_t - \mathbf{x}\|_2)$
- PCA reconstruction error $f(\hat{\mathbf{x}}_t) = |\hat{\mathbf{x}}_t - \mathbf{U}^T\mathbf{U}\hat{\mathbf{x}}_t|$

We discover that stripping the large neural network methods to a single layer and single building block suffices to achieve comparable performance.

- 1-Layer MLP
- Single block MLP Mixer
- Single Transformer block
- 1-Layer GCN-LSTM block



Proposed simple neural-network baselines

Key Insights

- Simpler methods not only rival but often surpass the performance of complex, state-of-the-art methods.
- Our findings suggest that many advanced models are performing tasks no more complex than linear mappings.
- The over engineered novel design choices in current methods are without much utility and rational.

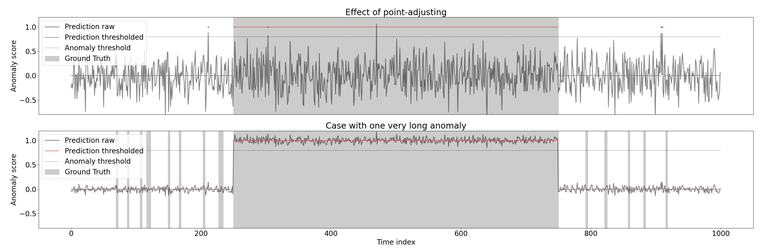
Benchmarking & Datasets Issues

- **Datasets:** Previous studies have identified many of the commonly used datasets for time series anomaly detection as unreliable.
- **Benchmarking:** There are glaring issues in benchmarking when reporting comparisons; use of different subsets of sensors of the dataset, use of flawed evaluation protocol, different pre/post processing.

Evaluation Metrics & Protocols Issues

- **$\mathbf{F1}$ score point-wise:** Standard point-wise metrics, captures the quality of detection of individual time stamps anomalies.
- **$\mathbf{F1}$ score with point-adjustment ($\mathbf{F1_{PA}}$):** Commonly used & flawed protocol. Using ground truth corrects the prediction for a whole anomaly interval based on a single hit. ***Random prediction can achieve a high score under this evaluation.***
- **$\mathbf{F1}$ score range-wise:** Time series range-wise metrics, capture the quality of anomaly interval coverage.

Standard point-wise metrics also have their own issues and should be complemented with range-wise metrics.



Top: A random prediction with an almost perfect $F1_{PA}$ score **Bottom:** A prediction which captures only the long anomaly with a very high point-wise F1 score

Are Simple Baselines Better than SOTA?

Simple baselines perform on par or better compared to state of the art methods.

		SWaT			WADI_127			WADI_112			SMD		
		$F1_{PA}$	$F1$	$F1_T$	$F1_{PA}$	$F1$	$F1_T$	$F1_{PA}$	$F1$	$F1_T$	$F1_{PA}$	$F1$	$F1_T$
SOTA methods	MERLIN	0.934	0.217	0.286	0.560	0.335	0.354	0.699	0.473	0.503	0.886	0.384	0.473
	DAGMM	0.830	0.770	0.402	0.363	0.279	0.406	0.829	0.520	0.609	0.840	0.435	0.379
	OmniAnomaly	0.831	0.773	0.367	0.387	0.281	0.410	0.742	0.441	0.496	0.804	0.415	0.353
	USAD	0.827	0.772	0.413	0.375	0.279	0.406	0.778	0.535	0.573	0.841	0.426	0.364
	GDN	0.866	0.810	0.385	0.767	0.347	0.434	0.833	0.571	0.588	**0.929**	0.526	0.570
	TranAD	0.865	0.799	0.425	0.671	0.340	0.353	0.680	0.511	0.589	0.827	0.457	0.390
	AnomalyTransformer	0.941	0.765	0.331	0.560	0.209	0.219	0.817	0.503	0.555	0.923	0.426	0.351
Simple baselines	Random	**0.963**	0.218	0.217	**0.783**	0.101	0.106	**0.907**	0.101	0.106	0.894	0.080	0.080
	Sensor range deviation	0.234	0.231	0.230	0.129	0.101	0.098	0.632	0.465	0.526	0.297	0.132	0.116
	L2-norm	0.847	0.782	0.366	0.353	0.281	0.410	0.749	0.513	0.607	0.799	0.404	0.338
	1-NN distance	0.847	0.782	0.372	0.372	0.281	0.410	0.751	0.568	0.618	0.833	0.463	0.384
	PCA Error	0.895	**0.833**	**0.574**	0.621	**0.501**	**0.557**	0.783	**0.655**	**0.699**	0.921	**0.572**	**0.580**
NN baselines	1-Layer MLP	0.856	0.771	0.519	0.295	0.267	0.384	0.601	0.502	0.558	0.829	0.514	0.487
	Single block MLPMixer	0.865	0.780	0.549	0.335	0.275	0.396	0.597	0.497	0.552	0.819	0.512	0.472
	Single Transformer block	0.854	0.787	0.526	0.471	0.289	0.416	0.646	0.534	0.575	0.781	0.489	0.420
	1-Layer GCN-LSTM	0.905	0.829	0.532	0.593	0.439	0.540	0.748	0.596	0.645	0.847	0.550	0.535

$\mathbf{F1_{PA}}$: F1 score with point-adjust $\mathbf{F1}$: the standard point-wise F1 score $\mathbf{F1_T}$: time-series range-wise F1 score

State-of-the-art Models' Learned Function

By distilling the state-of-the-art TAD models to a linear layer, we show evidence that complex models' learned function can be approximated with a linear function.

Methods	SWaT		WADI_112	
	Orig	Line	Orig	Line
Single block MLPMixer	0.780	0.770	0.497	0.500
Single Transformer block	0.787	0.772	0.534	0.521
1-Layer GCN-LSTM	0.829	0.794	0.596	0.587
TranAD	0.799	0.800	0.511	0.572
GDN	0.810	0.808	0.571	0.543

Orig: original model **Line**: linear approximated mode

Quo Vadis?

The perceived progress in Timeseries Anomaly Detection (TAD) is misleading, stemming from the use of inadequate metrics and the lack or low quality of benchmarking with simpler methods. To help improve the research efforts in TAD, we suggest:

- **Develop comprehensive datasets**: Create datasets with a range of anomaly difficulties.
- **Standardize evaluation protocols**: We suggest using both point-wise and range-wise metrics.
- **Run proper benchmarking**: This should be the starting point before building complicated solutions.