EBES: Easy Benchmarking for Event Sequences

Dmitry Osin*, Igor Udovichenko*,
Viktor Moskvoretskii, Egor Shvetsov, Evgeny Burnaev

Contacts: dima.tina2013@gmail.com
Telegram: @xaosina

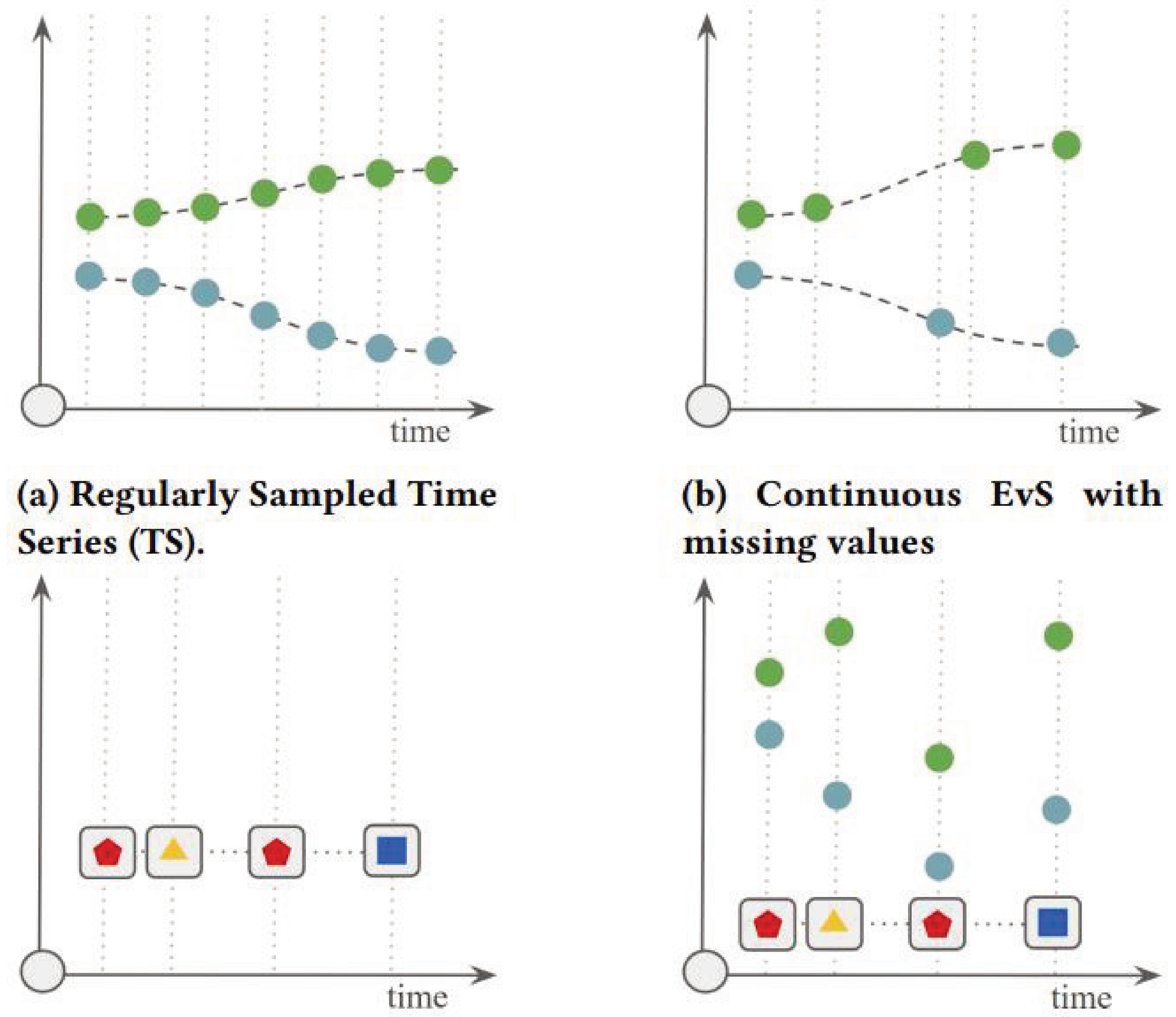
Skoltech

What is Event Sequences?

EXAMPLES:

1. Customer transactions
2. Medical measurements
3. Credit payments
4. E-commerce

The input consists of an **irregularly sampled** sequence of events containing both **categorical and numerical** features, with a **single label** assigned to the entire sequence.



(a) Regularly Sampled Time Series (TS).

(b) Continuous EvS with missing values

(c) A stream of discrete events, usually, modeled by Temporal Point Process (TPP).

(d) Discrete EvS with 2 numerical and 1 categorical features.

Why need benchmark?

Current issues:

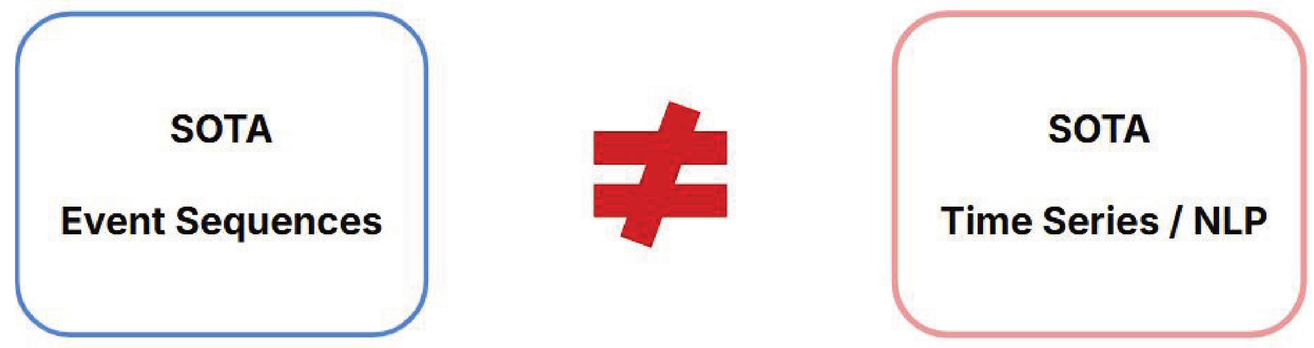
1. **Different data preprocessing** per paper.
2. **No proper HPO**. Architectures overfit to the test set.
3. **Noisy datasets** hinder method comparison.

The value of a **high-quality benchmark**:

1. **Simplifies** model selection for practical
2. **Reveals** the strengths and weaknesses of methods.
3. **Enables assessment** of the contribution of individual method components.

GRU are better than Transformer

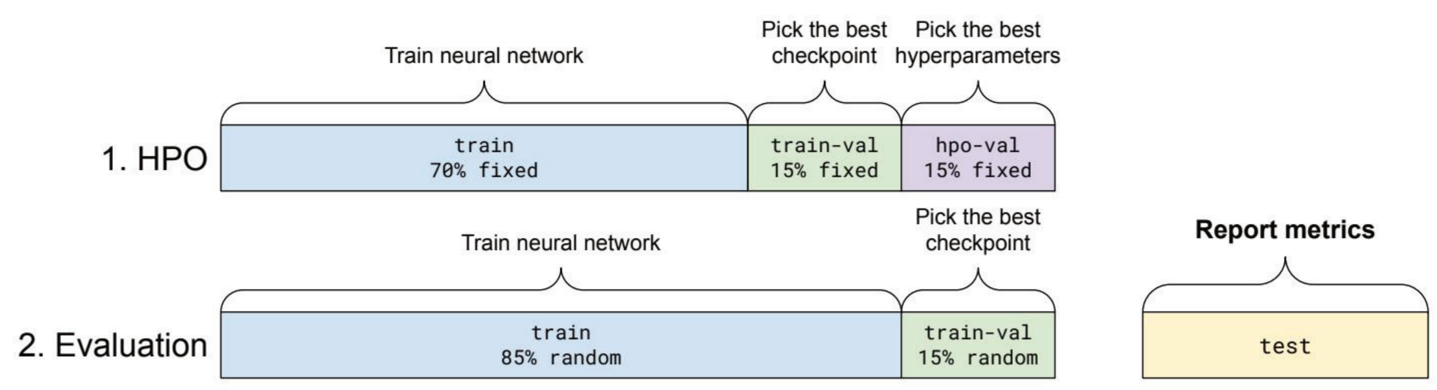
(for event sequence classification)



1. GRU-based models are **top-performers.**
2. Transformer go next in ranking.
3. Time Series methods **perform worse** on EvS.
4. MLP is not that bad
5. Physionet2012 **is bad** for model evaluation

Table 2: Model performance obtained using EBES. Results are averaged over 20 runs, with the best hyperparameters determined through HPO. Statistically indistinguishable ($p > 0.01$) results share the same superscripts, indicating the method's rank for each dataset. The best-performing methods for each dataset are highlighted. Methods are sorted according to their average rank across all datasets. Note: 4/20 runs of mTAND on the Pendulum dataset were excluded due to non-convergence (<20% accuracy). Number of learnable parameters presented in Table 6.

Category	Discrete EvS					Continuous EvS			Time Series	
Dataset	**MBD**	**Retail**	**Age**	**Taobao**	**BPI17**	**PhysioNet2012**	**MIMIC-III**	**Pendulum**	**ArabicDigits**	**ElectricDevices**
Metric	Mean ROC AUC	Accuracy	Accuracy	ROC AUC	ROC AUC	ROC AUC	ROC AUC	Accuracy	Accuracy	Accuracy
CoLES	0.826 ± 0.001^{2}	$\mathbf{0.553 \pm 0.002^{1}}$	$\mathbf{0.634 \pm 0.005^{1}}$	$\mathbf{0.713 \pm 0.002^{1}}$	$0.742 \pm 0.010^{3,4}$	$0.840 \pm 0.004^{2,3}$	$\mathbf{0.902 \pm 0.001^{1}}$	0.740 ± 0.013^{2}	$\mathbf{0.983 \pm 0.004^{1,2}}$	$\mathbf{0.729 \pm 0.019^{1,2}}$
GRU	$\mathbf{0.827 \pm 0.001^{1}}$	0.543 ± 0.002^{2}	0.626 ± 0.004^{2}	$\mathbf{0.713 \pm 0.004^{1}}$	$\mathbf{0.754 \pm 0.004^{1}}$	$\mathbf{0.846 \pm 0.004^{1}}$	$\mathbf{0.901 \pm 0.002^{1}}$	0.683 ± 0.031^{3}	0.975 ± 0.003^{4}	$\mathbf{0.741 \pm 0.013^{1}}$
MLEM	0.824 ± 0.001^{3}	0.544 ± 0.002^{2}	$\mathbf{0.634 \pm 0.003^{1}}$	$\mathbf{0.713 \pm 0.004^{1}}$	$\mathbf{0.753 \pm 0.005^{1,2}}$	$\mathbf{0.846 \pm 0.007^{1}}$	0.899 ± 0.002^{2}	0.676 ± 0.017^{3}	0.978 ± 0.002^{3}	$\mathbf{0.736 \pm 0.014^{1}}$
Transformer	0.821 ± 0.002^{4}	$0.536 \pm 0.006^{3,4}$	0.621 ± 0.006^{2}	$0.692 \pm 0.013^{3,4}$	$\mathbf{0.749 \pm 0.006^{1,2,3}}$	$0.838 \pm 0.008^{2,3,4}$	0.894 ± 0.002^{3}	0.658 ± 0.019^{4}	$\mathbf{0.986 \pm 0.004^{1,2}}$	0.710 ± 0.024^{2}
Mamba	0.820 ± 0.003^{4}	0.538 ± 0.003^{3}	0.609 ± 0.006^{3}	$0.693 \pm 0.023^{2,3}$	$0.737 \pm 0.012^{4,5}$	$0.835 \pm 0.006^{3,4}$	0.895 ± 0.002^{3}	0.687 ± 0.017^{3}	0.983 ± 0.005^{2}	0.716 ± 0.022^{2}
ConvTran	0.816 ± 0.002^{5}	0.534 ± 0.005^{4}	0.603 ± 0.006^{4}	0.703 ± 0.009^{2}	$0.748 \pm 0.006^{2,3}$	$0.837 \pm 0.006^{2,3,4}$	$0.892 \pm 0.005^{3,4}$	$0.674 \pm 0.028^{3,4}$	$\mathbf{0.986 \pm 0.003^{1}}$	0.711 ± 0.019^{2}
mTAND	0.798 ± 0.002^{7}	0.519 ± 0.003^{6}	0.582 ± 0.009^{5}	0.672 ± 0.010^{5}	0.738 ± 0.005^{4}	0.841 ± 0.005^{2}	$0.888 \pm 0.003^{4,5}$	$\mathbf{0.777 \pm 0.031^{1}}*$	0.951 ± 0.010^{5}	0.631 ± 0.019^{3}
PrimeNet	0.780 ± 0.006^{8}	0.521 ± 0.003^{6}	0.583 ± 0.011^{5}	0.681 ± 0.010^{4}	0.730 ± 0.006^{5}	$0.839 \pm 0.004^{2,3,4}$	0.887 ± 0.004^{5}	0.600 ± 0.026^{5}	0.958 ± 0.009^{5}	0.636 ± 0.016^{3}
MLP	0.809 ± 0.001^{6}	0.526 ± 0.002^{5}	0.581 ± 0.007^{5}	0.659 ± 0.035^{5}	0.737 ± 0.004^{4}	0.835 ± 0.004^{4}	0.881 ± 0.001^{6}	0.186 ± 0.006^{6}	0.760 ± 0.011^{6}	0.437 ± 0.019^{4}



The Role of Time and Order

Table 3: Testing on Permuted Sequences. Models were trained on non-permuted data; only the test set was permuted. We report performance difference relative to metrics obtained on not permuted sequences. Only values with statistically significant difference ($p < 0.01$) in performance are highlighted.

Category	Discrete EvS					Continuous EvS			Time Series	
Dataset	**MBD**	**Retail**	**Age**	**Taobao**	**BPI17**	**PhysioNet2012**	**MIMIC-III**	**Pendulum**	**ArabicDigits**	**ElectricDevices**
Metric	Mean ROC AUC	Accuracy	Accuracy	ROC AUC	ROC AUC	ROC AUC	ROC AUC	Accuracy	Accuracy	Accuracy
CoLES	−0.09%	−1.57%	−1.63%	−0.49%	−4.66%	−2.36%	−1.86%	−84.49%	−33.86%	−68.79%
GRU	−0.10%	−2.25%	−1.15%	−0.67%	−4.46%	−1.49%	−4.24%	−76.09%	−46.88%	−69.46%
MLEM	−0.30%	−2.57%	−1.52%	−0.89%	−3.80%	−1.71%	−1.43%	−81.84%	−37.81%	−65.17%
Transformer	−0.00%	−0.09%	−0.00%	−0.05%	−0.00%	0.03%	−0.00%	−0.00%	−15.12%	−25.26%
Mamba	−0.06%	−2.44%	−1.20%	−0.00%	−9.56%	−0.65%	−3.04%	−82.14%	−53.37%	−54.18%
ConvTran	−7.28%	−29.02%	−9.55%	−4.51%	−17.04%	−0.47%	−8.21%	−77.61%	−60.45%	−68.66%
mTAND	−5.05%	−28.09%	−8.95%	−4.13%	−9.07%	−4.13%	−5.05%	−82.57%	−59.12%	−56.04%
PrimeNet	−4.08%	−26.41%	−7.82%	−2.12%	−4.73%	−3.95%	−3.72%	−75.88%	−53.38%	−54.38%
MLP	−0.00%	−0.00%	−0.00%	−0.00%	−0.00%	−0.00%	−0.00%	−0.00%	−0.00%	−0.00%

Table 5: Training on Permuted Sequences without Timestamps. The GRU model with the best hyperparameters had the time feature removed and was then trained from scratch in two settings: *with* and *without* permuting both the training and test sequences. We report performance difference relative to metrics obtained on original sequences. Only values with statistically significant difference ($p < 0.01$) in performance are highlighted.

Category	Discrete EvS					Continuous EvS			Time Series	
Dataset	**MBD**	**Retail**	**Age**	**Taobao**	**BPI17**	**PhysioNet2012**	**MIMIC-III**	**Pendulum**	**ArabicDigits**	**ElectricDevices**
Metric	Mean ROC AUC	Accuracy	Accuracy	ROC AUC	ROC AUC	ROC AUC	ROC AUC	Accuracy	Accuracy	Accuracy
GRU w/o time	−0.89%	−0.00%	−0.44%	−3.85%	−0.00%	−0.00%	−0.27%	−59.43%	0.04%	−0.00%
GRU w/o time w/ perm.	−0.96%	0.50%	0.62%	−1.54%	−0.45%	−0.22%	−1.25%	−63.87%	−1.28%	−16.00%

On-Point-RND
(our team)

code

datasets