ADIA LAB STRUCTURAL BREAK CHALLENGE · REAL-TIME EDITION · 2026

Know what normal looks like

Real-time structural-break detection by testing each online stream against a model fitted to its own history

Harshveer harshveer321.code@gmail.com · first competition, joined 25 August (week 16 of 21)

64.66%

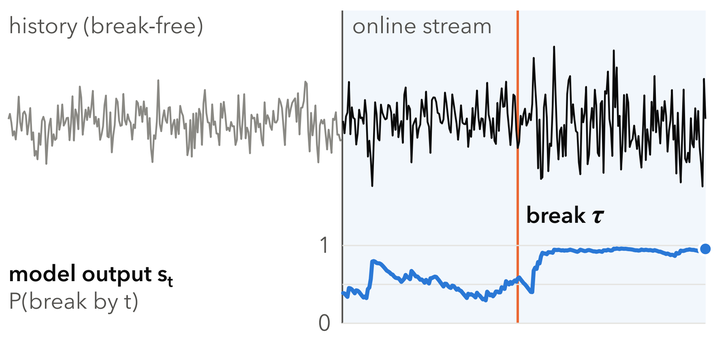
public TS-AUC · **rank 55 of 810 live models**

public board: #1 71.68 · #10 67.43 · #50 64.83

private TS-AUC 65.31% · **rank 36 of ~1,690 teams · top 2%**

1 The task

Each series is a break-free **history** (1,000–5,000 points), then an **online** stream (10–1,000 points) revealed one point at a time. After each point the model outputs s_t = P(a break has already happened). Half the series break, at a uniformly random time.

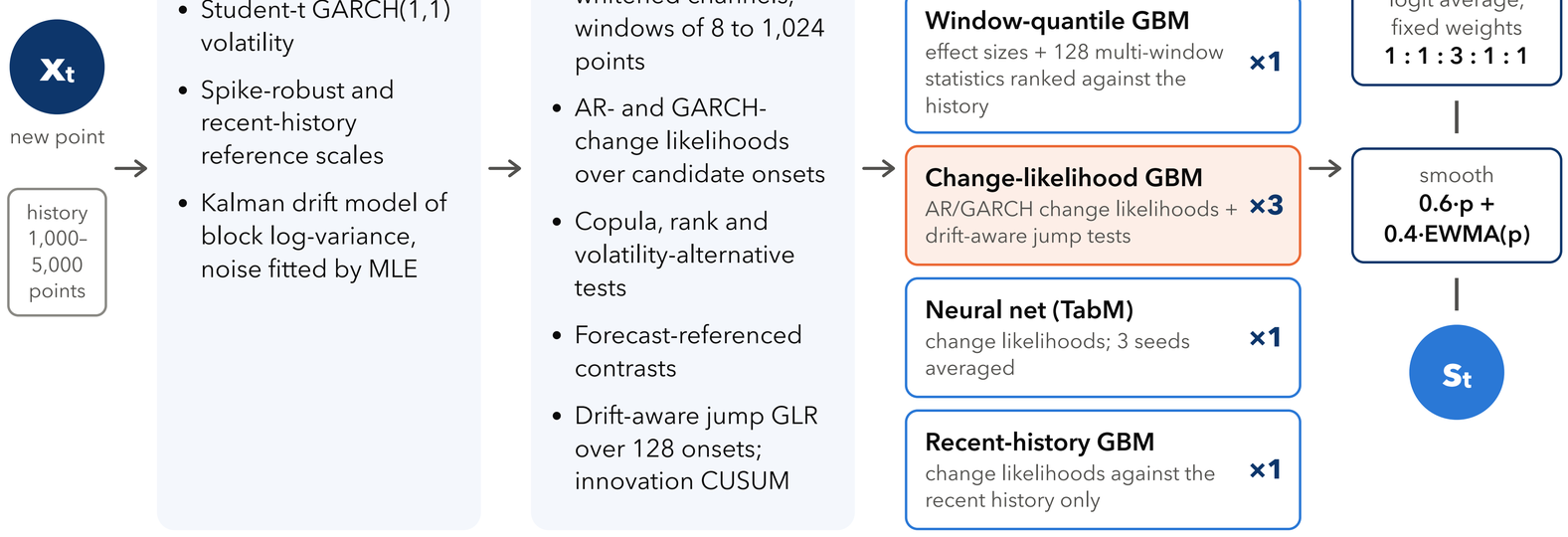


each step t is scored against all other series at the same t

TS-AUC compares broken with unbroken series *at the same step t*, pooled over all steps, so detection speed counts. Inference must be causal and deterministic.

Breaks change variance, dependence and tail shape, not the mean: a mean-shift detector scores 50.85, where chance is 50.

2 Method

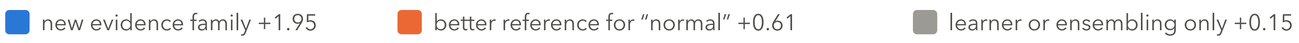


Training. 10,000 labelled series, 48 random steps per series. Each row is weighted by the number of opposite-class series alive at the same t, an exact pointwise surrogate of TS-AUC. Training features are verified identical to the streaming ones.

Weights. Fixed shares, not fitted: each new model took a pre-declared half of an existing slot. The ×3 was picked from {2, 3, 4, 6} on local cross-validation (+0.04, all 5 folds), never on the public board.

3 What moved the score in the final week

Public-leaderboard gain of each change, 24 Sep → 1 Oct



		PUBLIC GAIN	LOCAL OOF
25 Sep	Effect sizes of each channel vs its history	+0.65	+1.08
25 Sep	AR- and GARCH-change likelihoods	+0.75	+0.74
25–27 Sep	Copula, rank and volatility-alternative tests	+0.55	+0.97
27 Sep	Recent-history reference (+ linear leaves)	+0.17	+0.22
28 Sep	Forecast-referenced contrasts	+0.07	+0.15
28 Sep–1 Oct	Neural head; seed, logit and weight averaging	+0.15	+0.14
30 Sep–1 Oct	Spike-robust reference scale	+0.15	+0.17
1 Oct	Drift-aware jump detector	+0.22	+0.18
7 days	**Total: 61.95 → 64.66 public**	**+2.71**	**+3.65**

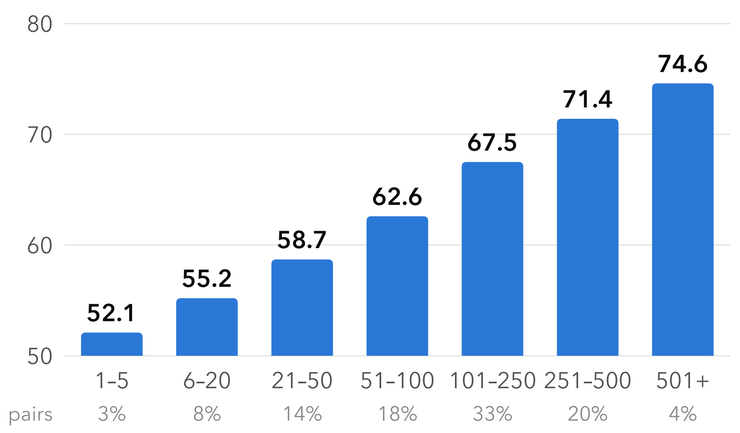
Most of the gain came from **new evidence**, meaning new tests against a model of the series' history. Better **references for "normal"** added +0.61. **Learner and ensembling changes** added +0.15 in total, and none added more than +0.07 alone. Steps under about 0.12 are not resolved one by one (see 6), so they are grouped here. Every row is positive on both the public board and local cross-validation.

4 How it performed

65.1
local 5-fold OOF, all 10,000 training series

0.97×
cloud gain ÷ local OOF gain, 27 Sep–1 Oct (+0.58 vs +0.60)

AUC by steps since the break



Near chance for 5 steps, 74.6 after 500; a third of all pairs sit at 101–250. Penultimate submission (64.43 public), 5-fold OOF.

5 Ideas that mattered most

1. **Optimise the metric exactly.** TS-AUC only compares series at the same t, so each training row is weighted by the count of opposite-class series alive at that t.
2. **Series, not rows, are the sample.** Boosters memorised series from ~500 correlated rows each (98.5 train, 52.3 validation). 48 random steps per series train as well as all 5 million.
3. **Whiten, then test against the history.** An AR(5) filter fitted on the history exposes variance, dependence and tail changes that raw values hide.
4. **Model each series' normal behaviour.** Real break-free series drift and carry rare volatility spikes. A spike-robust reference (+0.15 public) and a drift-aware Kalman jump detector (+0.22) were the largest late gains.
5. **Measure before believing.** One 7,500/2,500 split read 63.39 while 5-fold over all 10,000 series read 61.16. Every later decision used paired 10,000-series OOF.

6 Limits and noise

How far to trust a step

- **Public board:** four seeds of one model scored 61.86–61.95 (SD 0.04), so the difference between two runs carries SD ≈ 0.06.
- **Local:** paired 5-fold OOF over 10,000 series has an SE of 0.05–0.12 per change.
- **Selection:** a +0.20 "winner", the best of ~25 candidates, flipped sign on disjoint data. The final pick was the best of three near-equal models, worth about +0.04 of luck.

What did not help

Sequence models (GRU, TCN, LRU)	**no mixture gain**
Chronos-Bolt forecast features	**±0.04**
Simulated series as training rows	**−0.13**
Stacking; learned combiners	**−1.7; −0.09 to −0.25**
HMM / Bayes filtering of the score	**−1.5**

7 The gap to the leaders

The top ten scored 67.4–71.7. After the close, top teams described three things this pipeline did not do:

Learned null. A neural next-point model trained on all break-free history, with likelihood ratios summed over onsets. **+1.7**

Shift augmentation. Part of each history moved into the online segment, for 3× the data. **+0.8**

Synthetic pre-training. A deep sequence model pre-trained on simulated series, then fine-tuned. **68.1**

Effects as the teams reported them. This entry tested only weaker versions: augmentation on a noisy 2,500-series harness, simulations as extra booster rows, and a hand-built drift null. It does share the leaders' AR pre-whitening, row subsampling and plain averaging.

Takeaway. Every signal here is a test against a model fitted to the series' own history. New tests (+1.95) and better references (+0.61) gave 94% of the final week's gain. Learners and ensembling gave +0.15.

References Willsky & Jones 1976 (GLR for abrupt jumps) · Basseville & Nikiforov 1993 · Romano et al. 2022 (DeCAFS) · Ke et al. 2017 (LightGBM) · Gorishniy et al. 2025 (TabM)

Thanks to ADIA Lab and CrunchDAO