Testing EvoForest Beyond Structural Breaks: An Independent Evaluation on Atrial-Fibrillation Detection

A sealed transfer study on PhysioNet/CinC 2017

Gabriel Kahen
Independent researcher · 16 July 2026

Abstract. EvoForest frames learning as large-language-model (LLM) guided search over a multi-alternative computational graph, evaluated by a cross-validated Ridge readout. Its authors reported 0.9413 mean cross-validated ROC–AUC after 600 steps on a structural-break task. We independently reimplemented the published architecture and tested whether its representation search transferred to atrial-fibrillation (AF) discrimination in short single-lead ECGs. After excluding 279 noisy-label records, exact-waveform groups were divided into 6,599 development and 1,650 sealed-test recordings. Four CPU islands generated 600 proposals against fixed three-fold development ROC–AUC; seven predictors were frozen before one authenticated test-label opening. EvoForest improved its 43-feature seed from 0.7131 to 0.7698 test ROC–AUC, a paired gain of 0.0567 (95% CI, 0.0268–0.0878), and improved average precision by 0.0720 (0.0374–0.1112). A fixed 2,048-feature random-convolution representation ranked higher at 0.8784 ROC–AUC and 0.4966 average precision. The evolved 152-feature matrix had effective rank 27.48; 95.9% of its development gain appeared by proposal 53, and 56.3% of proposals failed. The architecture discovered transferable signal. In this run, the strong fixed representation performed better, and open-ended search imposed substantial operational burden. Clinical utility and any LLM-specific advantage remain untested.

1 EvoForest and its original test

1.1 Architecture

Most supervised pipelines optimize parameters within a fixed set of computations. EvoForest makes the computations themselves the main search object [1]. A shared directed acyclic graph (DAG) stores an internal population: intermediate nodes hold alternative tensor computations; callable nodes hold reusable transformations such as projections, gates, and activations; output alternatives produce candidate feature columns; fitting nodes may alter sample or residual weighting; and a persistent global store holds low-dimensional trainable parameters.

A graph configuration selects one alternative at each reachable intermediate, callable, and fitting node; all compatible output alternatives form the feature matrix. Continuous globals can first be refined on a fixed path with L-BFGS. After refinement, they are frozen and configurations are scored by cross-validated Ridge regression. The linear evaluator serves as judge and sensor: feature importance, redundancy, residual association, and fold stability are compressed into diagnostics. An LLM scientist proposes hypotheses from those diagnostics; an LLM engineer converts selected hypotheses into executable graph edits; and a memorandum agent preserves search history. Multiple asynchronous islands explore at different temperatures, exchange strong graphs, and alternate expansion with pruning and repair.

This architecture combines program search, feature discovery, limited gradient refinement, and a simple predictive readout. Its distinctive claim centers on persistence and reuse: LLM-generated computations remain available across proposals, and competing alternatives coexist inside one graph.

1.2 Reported structural-break result

The original paper evaluated EvoForest on the 2025 ADIA Lab Structural Break Challenge: binary detection of a change at a specified boundary in univariate time series. It reports 10,001 labeled series, four asynchronous GPU islands, at most 64 configurations per evaluation, and 600 evolution steps scored by mean three-fold ROC–AUC. The final development score was 0.9413. The paper compared that value with the public winner's reported leaderboard score of 0.9014 [1].

The two scores arise from different evaluations. EvoForest's value came from development cross-validation reused throughout adaptive search. The public winner's value came from a leaderboard. We therefore treat 0.9413 as an adaptive development estimate. Numerical reproduction of the private run falls outside this study. Our question was whether the public architectural pattern could add held-out signal on a substantially different time-series problem.

2 Independent reimplementation

We independently implemented EvoForest as faithfully as possible from the architecture documented in the paper. The source is public at https://github.com/Gabriel-Kahen/evoforest-reimplementation [2]. It follows the published multi-alternative graph model, configuration search, callable and fitting nodes, persistent global parameters, two-phase refinement, Ridge evaluation, diagnostic feedback, structured scientist–engineer mutations, memoranda, graph maintenance, persistent asynchronous islands, and migration. Given the information public in the paper, we consider this the best possible architecture-level reimplementation.

Several components remain private: the authors' prompts, LLM stack, scheduler, evolved graph, competition pipeline, and source code. Our AF evaluation therefore concerns the reconstructed architecture. Exact software behavior and numerical reproduction of the structural-break result fall outside the study's scope. The run used four persistent CPU actors and Gemini 3.1 Flash-Lite with independently written prompts based on the public paper.

3 Why atrial-fibrillation detection?

PhysioNet/CinC 2017 offers a useful cross-domain stress test: public, variable-length, single-lead ECG recordings collected outside the structural-break setting [3, 4, 5]. AF discrimination plausibly rewards exactly the kind of compositional search EvoForest is designed to perform. Useful representations may combine level distributions, beat-scale irregularity, local change, autocorrelation, entropy, and multi-scale summaries; ROC–AUC also supplies a direct non-differentiable search target.

The dataset presents a demanding setting. AF is uncommon, recordings vary in length and acquisition quality, and the negative "other rhythm" class can contain irregular rhythms that resemble AF [4]. This makes generic variability an incomplete shortcut. It also creates a meaningful contest between adaptive computation discovery and a broad fixed basis: if dense local shape coverage is already sufficient, random convolution should be hard to beat [6].

We chose the task as a transfer study. The official Challenge evaluated four classes on a private hidden test; our study used binary AF versus normal/other rhythm on an internal sealed partition of the public cohort. Our evidence therefore applies to this architecture and internal split. Official Challenge standing and clinical validation require separate studies.

4 Experimental run

4.1 Cohort, preprocessing, and seal

All 8,528 release-1.0.0 recordings were reconciled to final V3 labels: 758 AF, 5,076 normal, 2,415 other rhythm, and 279 too noisy to classify. We excluded the noisy class before splitting, leaving 8,249 records. Exact-waveform groups were kept together in a seeded 80/20 split: development contained 6,599 records (602 AF) and the sealed test 1,650 (156 AF). Grouping prevented exact-waveform leakage. Patient disjointness remains uncertain because identifiers were unavailable.

WFDB physical calibration was retained. Signals were resampled from 300 to 100 Hz, zero-right-padded, and paired with valid-length masks. Preprocessing stopped there; filtering, detrending, amplitude normalization, cropping, and augmentation were omitted. Test labels were unavailable to search and fitting. Frozen model hashes and predictions were authenticated before labels were read; the sealed evaluator then aligned record identifiers, scored once, and blocked reuse.

4.2 Search and comparators

The 43-feature seed combined robust distribution, first-difference, fixed-lag autocorrelation, gated, and projected summaries. Four asynchronous CPU islands generated 600 LLM-guided proposals, migrating every ten proposals. Candidate fitness was mean ROC–AUC on three fixed development folds. Each evaluation considered at most 64 exact configurations; L-BFGS refinement preceded standardized Ridge scoring. Across 425 completed train evaluations, the run scored 19,004 exact configurations. It accepted 19 proposals, rejected 243, and failed 338. Search took 29 h 54 min and 2,943 LLM calls (US$12.03 under the recorded pricing assumptions).

Before opening the test set, we froze seven development-fitted predictors: a constant; two duration features; Catch22's 22 fixed summaries [7]; 42 generic waveform summaries; the unevolved seed; full EvoForest; and a ROCKET-like transform with 1,024 fixed random kernels, each summarized by maximum response and proportion positive, for 2,048 features [6]. The random-convolution baseline was intentionally strong and used 13.5 times as many features as full EvoForest.

4.3 Outcomes and uncertainty

The prespecified primary contrast was full EvoForest minus its unevolved seed in sealed ROC–AUC; average precision (AP) was secondary. We used 10,000 class-stratified record bootstraps, with paired resamples for that contrast. Intervals condition on one record split and one stochastic search. All other between-model, threshold, and subgroup comparisons are descriptive.

5 Results

5.1 Evolution added held-out signal

Full EvoForest achieved sealed ROC–AUC 0.7698 versus 0.7131 for the seed, a paired gain of 0.0567 (95% CI, 0.0268–0.0878). AP rose from 0.1867 to 0.2588, a gain of 0.0720 (0.0374–0.1112). The seed nearly matched the generic 42-feature summary model and provided a credible starting point for the observed gain.

5.2 The strongest fixed basis ranked higher

Random convolution exceeded EvoForest's point estimates by 0.1086 ROC–AUC and 0.2378 AP (Table 1). The comparison was descriptive, and random convolution used 13.5 times as many features. The evidence here is the observed ranking: among the tested strategies, adaptive graph evolution was intermediate between compact fixed summaries and a high-dimensional fixed transform.

The ranking mattered at a concrete operating point. At thresholds chosen on development to target 95% specificity, the seed, full EvoForest, and random convolution detected 24, 41, and 88 of 156 AF cases, with 65, 73, and 75 false positives. These fixed-threshold descriptive values require independent clinical validation. Random convolution also had higher point estimates in AF-versus-normal, AF-versus-other, and every defined duration and signal-quality subgroup; the subgroup analysis reports point estimates only.

5.3 Search growth outpaced information

The global-best development score rose from 0.7270 to 0.7724. Proposal 53 had already reached 0.7705, or 95.9% of the total gain; the remaining 547 proposals added 0.0019 (Figure 1). Proposal 53 has no sealed-test score. The trajectory documents a development plateau, and the test performance of an early-stopped search remains unknown.

The final graph produced 152 features with effective rank 27.48 and mean maximum absolute interfeature correlation 0.9862. Its largest development contributions came from gated distribution, first-difference, interaction, and autocorrelation summaries, many in near-duplicate families. The search discovered useful structure. Graph size substantially overstated independent information. Operationally, 56.3% of proposals failed and only 3.2% were accepted, making proposal validity

Table 1: Sealed-test discrimination. Model intervals are 95% stratified record-bootstrap intervals. Only the paired full-minus-seed contrast was prespecified for inferential comparison. AP chance level equals the test AF prevalence, 0.0945.

Representation / predictor	Features	ROC–AUC (95% CI)	AP (95% CI)
Constant prevalence	–	0.5000 (0.5000–0.5000)	0.0945 (0.0945–0.0945)
Duration only	2	0.5511 (0.5120–0.5892)	0.1128 (0.1003–0.1377)
Catch22	22	0.6481 (0.6028–0.6911)	0.1449 (0.1258–0.1758)
Generic waveform summaries	42	0.7082 (0.6674–0.7477)	0.1853 (0.1586–0.2290)
Unevolved EvoForest seed	43	0.7131 (0.6721–0.7516)	0.1867 (0.1600–0.2303)
Full EvoForest	**152**	**0.7698 (0.7295–0.8079)**	**0.2588 (0.2187–0.3154)**
Fixed random convolution	2,048	0.8784 (0.8478–0.9064)	0.4966 (0.4271–0.5841)
Full minus seed (paired)	–	**+0.0567 (0.0268–0.0878)**	**+0.0720 (0.0374–0.1112)**

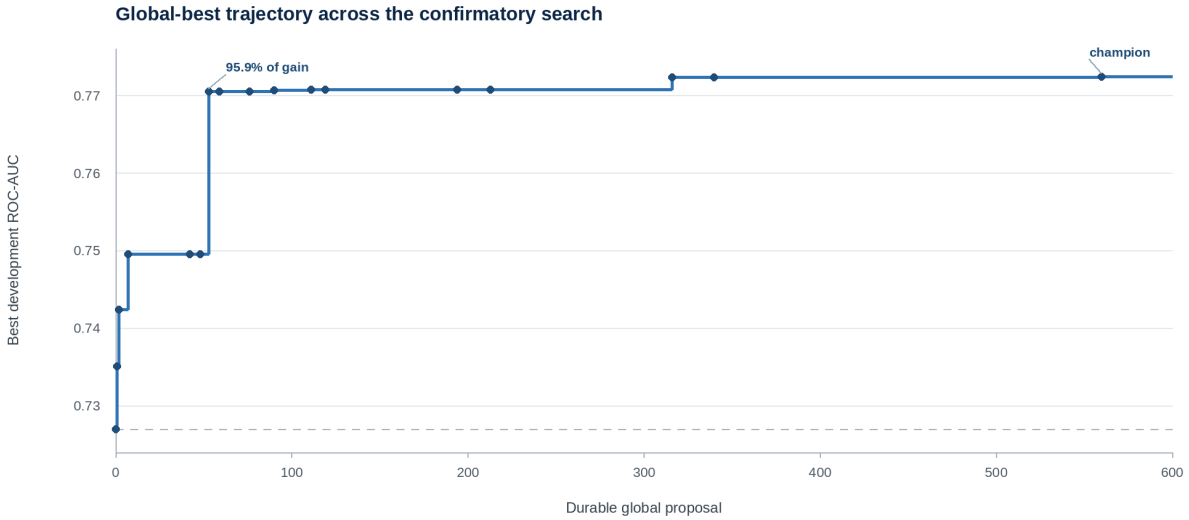

Figure 1: Development trajectory of durable global-best graphs. The dashed line marks the seed. Search repeatedly reused the same folds, so this curve records optimization on those folds. Generalization was measured separately on the sealed test. The frozen champion first appeared at proposal 560.

and budget allocation central weaknesses of this run.

6 What the results imply

6.1 What transferred

The positive paired intervals support a narrow conclusion: this implementation of EvoForest added held-out AF signal beyond its own credible seed. That is meaningful evidence that the architecture can transfer beyond structural-break detection. The result also illustrates a genuine strength of the design: it can optimize a task metric and assemble task-specific computations into a compact linear-readout basis.

The source of the gain remains unresolved. A budget-matched random, heuristic, or non-LLM graph search would be needed to isolate it. We therefore attribute the gain to adaptive graph expansion as a whole; the specific role of LLM reasoning remains unresolved. High linear contribution alone offers weak evidence of physiological interpretability: severe collinearity makes component-level stories unstable, and several evolved names overstated what their underlying primitives computed.

6.2 Strengths, weaknesses, and fit

EvoForest appears most promising where the right computations are genuinely uncertain, data are moderate in scale, the evaluation target is awkward to differentiate, and reusable domain structure is worth discovering. Examples may include structured scientific time series, industrial signals, or tabular tasks where a compact auditable basis has value and existing fixed transforms leave room for improvement. Diagnostics, explicit graph structure, and retained competing computations offer useful research affordances alongside a simple final predictor.

This run also exposed substantial costs. Search incurred LLM expense, long wall time, code-generation failures, adaptive reuse of development folds, and a highly redundant champion. On short raw ECG waveforms, a fixed random-convolution basis supplied broader local pattern coverage, generated its feature set directly, and achieved much stronger discrimination. These results favor other first-line methods for raw waveforms and for tight-compute or high-reliability settings. Clinical use requires independent patient-level validation. Practical remedies include redundancy penalties, typed mutation libraries, adaptive stopping, budget-matched controls,

and hybrid seeds that expose strong fixed convolutional features to graph search.

6.3 Limits on generalization

This was one public dataset, one recording-level split, and one stochastic search trajectory. Patient overlap is unknown. The sealed set was drawn internally from public data; the official Challenge test remained separate. The binary target also differs from the original four-class task. The same development folds were optimized over 600 proposals. Bootstrap intervals capture record-resampling uncertainty. Patient, split, search, prompt, and model uncertainty remain unmeasured. The CPU topology and independently developed Gemini agents differed from the private original system. Feature and compute budgets were unmatched for the strongest comparator. These constraints bound both positive and negative claims.

7 Conclusion

An independently implemented EvoForest architecture transferred useful signal to AF detection and passed its prespecified comparison with an unevolved seed. It ranked well below a fixed random-convolution representation. Its feature matrix was redundant, and the failure-prone search had nearly plateaued by proposal 53. These findings position EvoForest as a flexible computation-discovery scaffold. Future work must establish competitive AF performance and isolate the contribution of LLM guidance. Its best next test is a repeated, patient-disjoint, matched-budget comparison against non-LLM search and strong hybrid fixed representations.

Availability and integrity

The public architecture code is available at https://github.com/Gabriel-Kahen/evoforest-reimplementation. The archival study bundle identifies run `run_20260712T142307Z` and source commit `253be7f014a53a978045a1118cf835e2e45fbdaf`; it contains split manifests, frozen models and predictions, metrics, bootstrap outputs, and run logs. A post-search `score-to-basic` analysis-compatibility mapping reproduced all stored development predictions and fold scores. Frozen predictors and sealed outputs remained unchanged. Manuscript drafting and typesetting used automated assistance; all claims and artifacts were checked against the recorded run.

References

[1] K. A. Yuksel and H. Sawaf. EvoForest: A novel machine-learning paradigm via open-ended evolution of computational graphs. *arXiv:2604.19761*, 2026. https://doi.org/10.48550/arXiv.2604.19761.

[2] G. Kahen. EvoForest reimplementation: an independent implementation of the published software architecture. GitHub repository, 2026. https://github.com/Gabriel-Kahen/evoforest-reimplementation.

[3] G. D. Clifford, C. Liu, B. Moody, L.-W. H. Lehman, I. Silva, A. Johnson, and R. Mark. AF classification from a short single lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017, version 1.0.0. *PhysioNet*, 2017. https://doi.org/10.13026/d3hm-sf11.

[4] G. D. Clifford, C. Liu, B. Moody, L.-W. H. Lehman, I. Silva, Q. Li, A. E. W. Johnson, and R. G. Mark. AF classification from a short single-lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017. *Computing in Cardiology*, 44:1–4, 2017. https://doi.org/10.22489/CinC.2017.065-469.

[5] A. L. Goldberger et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. https://doi.org/10.1161/01.CIR.101.23.e215.

[6] A. Dempster, F. Petitjean, and G. I. Webb. ROCKET: Exceptionally fast and accurate time-series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020. https://doi.org/10.1007/s10618-020-00701-z.

[7] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones. catch22: CAnonical Time-series CHaracteristics. *Data Mining and Knowledge Discovery*, 33:1821–1852, 2019. https://doi.org/10.1007/s10618-019-00647-x.