MoSiLa: Moment–Signature Landmarks for Multivariate Time Series Classification

Anonymous authors
Paper under double-blind review

Abstract

Regional compression reduces the cost of time series attention, but the resulting tokens determine which amplitude and order information remains available. We introduce MoSiLa (Moment–Signature Landmarks), a classifier that represents a learned feature field through multiscale means, temporal moments, and low-rank path coordinates. Two relation-aware attention blocks operate on at most 85 landmarks. An independent random-kernel expert provides an alternative representation, with training out-of-fold predictions selecting a single route or a mixture. Across all 30 UEA datasets, the complete MoSiLa system achieves 78.739% macro-average accuracy, the highest among 18 measured baselines under heterogeneous, TEST-selected training recipes. Every listed fixed-recipe ablation has a lower macro-average, with numerical reductions of 0.330–2.248 percentage points. A separate TRAIN-loss-selected evaluation yields 76.570% $\pm$ 0.439% across three seeds, exceeding the three-seed averages of MultiRocket, HYDRA, and their combination under the recorded preprocessing recipes. External results show numerical gains over the additional measured comparators but remain below the strongest included literature reports. These findings distinguish the performance of a complete classifier from the empirical utility of its representation choices.

1 Introduction

Multivariate time series classification must recognize evidence that varies in duration, location, and temporal organization. A recurring waveform can identify an activity, while the progression of similar waveforms can distinguish another. Convolutional classifiers, random-kernel transforms, attention models, and state-space models offer different ways to represent this evidence (Fawaz et al., 2020; Dempster et al., 2023; Foumani et al., 2024; Jung & Kim, 2026). Their coexistence motivates a question beyond backbone selection: which aspects of a sequence should a classifier preserve when it aggregates observations?

For attention-based classifiers, aggregation determines both computational cost and accessible information. Timestamp-level self-attention requires a quadratic number of interactions. Patching reduces that cost, but subsequent attention operates on the patch representation rather than the observations that it replaces (Nie et al., 2023; Moreno-Pino et al., 2024). Averaging a fixed feature field is invariant to permutations of its feature vectors inside a region; attaching a position to that average cannot recover the discarded within-region order. Conversely, path increments encode changes rather than an absolute feature level. A convolutional encoder can already encode local order before aggregation, so explicit order statistics must be evaluated as an additional representation choice, not assumed to be necessary.

A second issue concerns whether one learned representation is appropriate across heterogeneous classification datasets. Fixed random-kernel features with a fitted linear readout remain competitive alternatives to end-to-end feature learning (Tan et al., 2022; Dempster et al., 2023). Requiring every prediction to pass through a learned token representation excludes that alternative; always averaging the two predictors assumes that their combination helps. We therefore study representation construction together with prediction-route selection: first expose regional amplitude and order to attention, then use training predictions to decide whether this route, a pattern route, or a mixture is appropriate for the dataset.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our classifier, MoSiLa (Moment–Signature Landmarks), implements these two decisions separately. A masked convolutional encoder produces a feature field at the original valid timestamps. Four temporal partitions generate at most 85 landmarks, each containing a regional mean, a signed first temporal moment, and a residual map of low-rank path displacement and signed area. Relation-aware attention connects landmarks using their content, location, and scale. The coarsest landmark serves as a global readout. The independent pattern route instead represents the occurrence and transitions of random-filter responses and fits a ridge classifier. Five-fold out-of-fold predictions support a fixed choice among two single experts and three mixture predictors before full-data refitting.

This separation leads to two empirical questions: how the complete procedure compares with other implemented classifiers, and which representation and readout choices contribute within it. Some datasets select only the pattern expert, so system accuracy alone cannot establish landmark utility. We evaluate UEA30 classification, structural and depth ablations, TRAIN-loss checkpoint selection, and two external datasets. The primary comparison uses TEST-selected checkpoints and heterogeneous training recipes; it measures those implemented systems rather than an equal-budget, untouched-test comparison.

The contributions are:

- A fixed-cardinality landmark interface that supplies regional feature level, temporal moment, and low-rank path order to cross-scale attention. Its bounded token count separates attention resolution from sequence length.
- A complete two-route fitting procedure that selects among neural, pattern, and mixture predictors using paired training out-of-fold records. The route decision is explicit and can be examined alongside the final accuracy.
- A full UEA30 evaluation separating system performance from component utility. MoSiLa reaches 78.739% macro-average against 18 measured baselines; fixed-recipe and retuned ablations examine component utility, while repeated TRAIN-based-selection runs and external comparisons quantify sensitivity to evaluation conditions.

Section 3 defines the representation and selection procedure. Sections 4 and 5 describe the evaluation conditions and examine the resulting evidence.

2 Related work

Classification backbones. InceptionTime develops multiscale convolutions for classification (Fawaz et al., 2020); ConvTran studies convolution–attention models and positional encoding (Foumani et al., 2024); TSLANet combines spectral and convolutional processing (Eldele et al., 2024). MambaSL examines a shallow selective state-space backbone, whereas TSCMamba uses multiple feature views (Jung & Kim, 2026; Ahamed & Cheng, 2025). General-purpose time series architectures extend the comparison beyond classification-specific designs (Wu et al., 2023; Luo & Wang, 2024; Wang et al., 2025). Our focus is the interface between local features and regional aggregation, rather than replacing every feature extractor with a single preferred backbone.

Patches, signatures, and attention. Patching and multigranularity representations limit the objects processed by attention (Vaswani et al., 2017; Nie et al., 2023; Wang et al., 2024b). Rough Transformers are a close predecessor: they already connect local and global signature views with attention (Moreno-Pino et al., 2024). MoSiLa instead forms moments of a learned convolutional field, adds 36 residual coordinates from an eight-dimensional learned path, and relates a fixed four-resolution partition through relative geometry. Its contribution lies in this particular representation interface and its evaluation, not in inventing signatures or their use with Transformers. A direct empirical comparison with Rough Transformers is absent from the current benchmark.

Pattern transforms and prediction routes. MultiRocket aggregates random convolutional responses through several pooling operators (Tan et al., 2022). HYDRA counts competitions among random kernels and complements Rocket-family transforms (Dempster et al., 2023). MoSiLa uses a HYDRA-inspired expert with centered winner-transition features and a ridge readout, alongside the learned landmark route. Out-of-fold selection determines whether either route should be used alone

or combined. This distinguishes the proposed system from both a neural-only architecture and an unconditional ensemble; the constituent random-kernel and expert-selection ideas remain established tools.

3 METHOD

The complete MoSiLa classifier separates a learned landmark representation from an independently fitted pattern representation. Four design questions organize the construction: what to preserve within each region, how to relate regions, when to retain a fixed-feature alternative, and how to select the prediction route. They motivate the choices below; Section 5 examines their empirical utility.

3.1 TASK AND LOCAL FEATURE FIELD

An example is $X = (x_0, \ldots, x_{L-1}) \in \mathbb{R}^{L \times C}$ with label $y \in \{1, \ldots, K\}$. Batches contain padded sequences and explicit valid-prefix masks. Channel normalization statistics are fitted on the current training subset. Three Inception-style units, each using kernel lengths 40, 20, and 10 plus a max-pooling branch, produce $H = (h_t)_{t=0}^{L-1} \in \mathbb{R}^{L \times d}$ with $d = 128$. These units supply local waveform context before regional compression. A projected input shortcut is added after the three units. Padding is masked between operations and excluded from batch-normalization statistics. Figure 1 shows the two independent prediction routes.

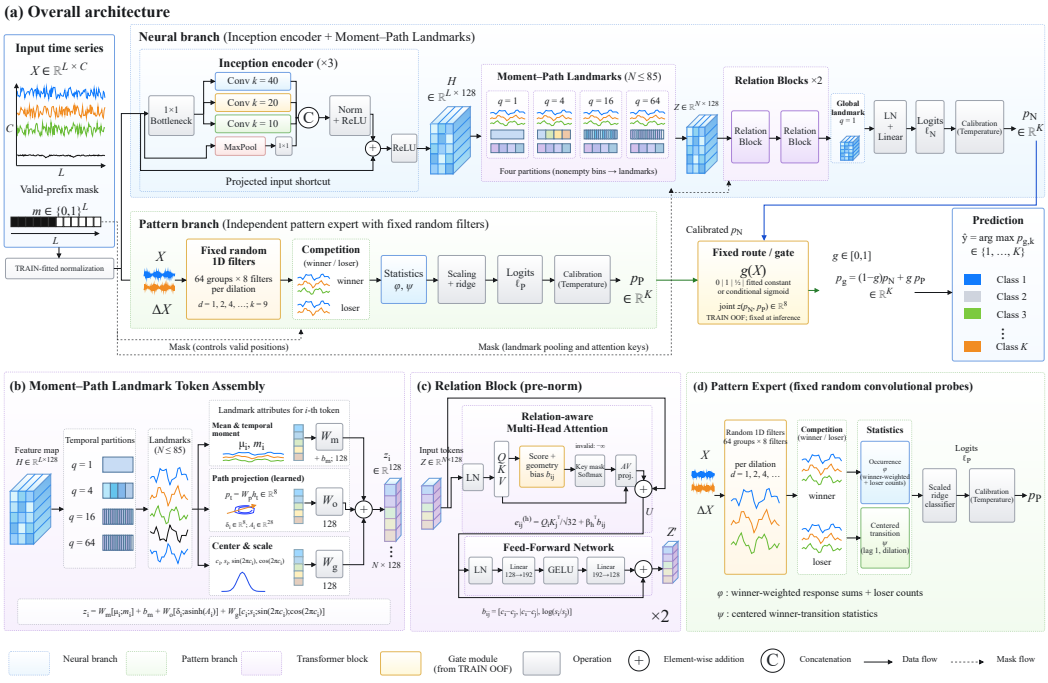


Figure 1: MoSiLa architecture. (a) Independent neural and pattern experts meet at a prediction rule selected from TRAIN out-of-fold predictions and fixed at inference. Solid arrows show data flow; dashed arrows show mask control. (b) Regional mean, temporal moment, path, and geometry coordinates form at most 85 landmarks. (c) Two relation blocks combine content attention with relative geometry; the updated $q = 1$ landmark supplies the neural readout. (d) Fixed random filters produce occurrence and centered-transition features for a scaled ridge classifier. Waveforms and tensor grids are schematic.

3.2 MOMENT–PATH LANDMARKS

D1: Preserve feature level and coarse order. A regional mean retains feature level but removes the order of an already constructed feature field. A temporal moment adds signed association with

time; path displacement and area add low-dimensional change and interaction coordinates. Their joint use exposes these distinctions without retaining every timestamp. Four nested resolutions provide both whole-sequence context and finer regions under a fixed token budget. Let $u_t = (t + \frac{1}{2})/L$. For each partition size $q \in \{1, 4, 16, 64\}$ and bin $j \in \{0, \dots, q-1\}$, define $I_{qj} = \{t : \lfloor qu_t \rfloor = j\}$, center $c_{qj} = (j + \frac{1}{2})/q$, and width $s_{qj} = 1/q$. An empty bin is masked, not interpreted as zero-valued evidence. For a nonempty bin with n_{qj} elements, the amplitude and temporal moment are

$$\mu_{qj} = \frac{1}{n_{qj}} \sum_{t \in I_{qj}} h_t, \qquad m_{qj} = \frac{\sqrt{12}}{n_{qj}} \sum_{t \in I_{qj}} q(u_t - c_{qj}) h_t. \tag{1}$$

The moment measures a signed within-bin temporal trend rather than a least-squares slope for every finite sampling pattern. For two different feature values, swapping their order preserves the mean but changes the moment. This statement concerns H: permuting raw observations can also change the preceding convolutional responses.

A learned linear map forms $p_t = W_p h_t \in \mathbb{R}^r$, with $r = 8$. For the contiguous bin endpoints a, b, set $\delta_{qj} = p_b - p_a$ and

$$A_{qj} = \frac{1}{2} \sum_{t=a}^{b-1} (p_t - p_a) \wedge (p_{t+1} - p_t), \quad (v \wedge w)_{ij} = v_i w_j - v_j w_i \quad (i < j). \tag{2}$$

These are the displacement and independent antisymmetric level-two coordinates of a piecewise-linear path. Signed area distinguishes some paths sharing the same endpoints, while the mean retains feature-level information. They have $8 + \binom{8}{2} = 36$ dimensions. Singleton bins have zero displacement and area. The implementation uses the standard prefix composition identity, given in Appendix B.4, and applies asinh coordinatewise to the area. Truncation retains selected order information, not an exact reconstruction of the path.

The token for bin $i = (q, j)$ is

$$z_i = W_m[\mu_i; m_i] + b_m + W_o[\delta_i; \operatorname{asinh}(A_i)] + W_g[c_i; s_i; \sin(2\pi c_i); \cos(2\pi c_i)]. \tag{3}$$

We initialize $W_o = 0$ so that optimization starts from the mean–moment representation and learns the path residual. Its output weights can learn from the first update. Gradients into W_p through this route become available as W_o moves away from zero. This starting point does not constrain subsequent performance relative to the base representation. The construction produces at most $N = 1 + 4 + 16 + 64 = 85$ tokens, regardless of sequence length.

Figure 5 in Appendix B.3 illustrates equal-mean fields with opposite moments and equal-displacement paths with opposite areas.

3.3 Relations and readout

D2: Relate content at different temporal scales. Similar content can occur at different locations or cover different durations. Relative geometry supplies this distinction directly to attention. The $q = 1$ landmark already summarizes the whole feature field and becomes a global readout after attending to finer regions; it is data-derived, not a free classification token. Mean, maximum, and last-token readouts test this choice in Section 5.2. Each of two pre-normalized attention blocks has four heads of width 32 and a feed-forward hidden width of 192. For head h, its score is

$$e_{ij}^{(h)} = \frac{Q_i^{(h)} K_j^{(h)\top}}{\sqrt{32}} + \beta_h^\top \left[c_i - c_j,\ |c_i - c_j|,\ \log(s_i/s_j) \right]. \tag{4}$$

Invalid keys receive $-\infty$ before softmax, and invalid token outputs are zeroed after the block. Learned relation coefficients $\beta_h \in \mathbb{R}^3$ start at zero. Residual attention and GELU feed-forward updates preserve the token width. MoSiLa reads the updated $q = 1$ token, applies layer normalization and a linear classifier, and produces logits $\ell_N(X)$. The neural parameters minimize ordinary cross-entropy. The two experts share neither gradients nor an auxiliary loss.

3.4 Independent pattern expert

D3: Retain an alternative to learned regional features. Random-filter statistics describe local pattern occurrence without requiring the landmark encoder to discover each filter from labels. Keeping

this route independent allows it to remain available when learning the regional representation is unhelpful. The expert is HYDRA-inspired (Dempster et al., 2023), but includes centered transition features in addition to occurrence statistics. The pattern route applies groups of fixed random length-nine filters at dyadic dilations to raw and first-difference views. Within each group, eight filters compete at each valid timestamp. Winner-weighted response sums and loser counts form occurrence features $\phi(X)$. Directed winner-transition counts at lags 1 and the dilation, centered by their lag-specific marginal products, form order features $\psi(X)$. Random channel subsets allow the same construction to handle different channel counts. Appendix B.5 specifies the transformation and fitting grids.

After training-fitted feature scaling, a ridge classifier combines the two feature blocks. In the exact backend, its kernel is

$$G_\omega = \Phi\Phi^\top + \omega\Psi\Psi^\top, \qquad B = (G_\omega + \lambda I)^{-1}(Y - \mathbf{1}\bar{y}^\top). \tag{5}$$

The fitted prediction includes the target mean $\bar{y}$. Fixed-design leave-one-out predictions select ω, λ, and a temperature. This calculation holds the fitted feature design fixed; it is not a leave-one-out refit of the entire preprocessing pipeline. When the total TRAIN cohort exceeds 2,000 examples, each feature block is projected to 512 Gaussian features and fitted in the primal. The backend is fixed from the dataset's TRAIN size before constructing folds; no examples are discarded. The resulting expert produces $\ell_P(X)$ independently of the neural optimizer.

3.5 Training-only selection of the prediction route

D4: Select the route from held-out training predictions. Different representations need not help every dataset, and an unconditional average can dilute the stronger predictor. The candidate set therefore includes both single experts as well as mixtures. Paired out-of-fold errors compare their predictions on the same examples; a fold-variation penalty favors an improvement that is not concentrated in one fold. Five class-preserving folds produce paired out-of-fold logits $(\ell_{N,i}, \ell_{P,i})$. Each neural fold fit uses 100 epochs. Temperature-scaled probabilities p_N, p_P enter

$$p_g(y \mid X) = (1 - g(X))p_N(y \mid X) + g(X)p_P(y \mid X). \tag{6}$$

The candidates are neural-only, pattern-only, an equal mixture, a fitted constant mixture, and a conditional sigmoid gate with eight probability-summary inputs. A three-fold meta-level split fits temperatures and mixture parameters without using a row's own label to fit that row's meta-predictor. For a candidate and the current reference, let d_i be their paired difference in zero–one error. A candidate replaces the reference only when

$$\bar{d} + \frac{\mathrm{SD}(\bar{d}_1, \ldots, \bar{d}_5)}{\sqrt{5}} < 0, \tag{7}$$

where the five means group observations by their original base-model fold. Candidates are examined in the fixed order above; accepted choices are refitted to all out-of-fold records. The threshold is an empirical stability heuristic, not a confidence bound: fold training sets overlap, and the meta-level records share fitted base models. Appendix B.6 gives the exact features and regularization.

The gate is then fixed, and both experts are refitted on all TRAIN examples. At inference, their calibrated predictions are combined by Equation 6. The implemented full-system evaluation computes both routes even when one receives zero weight. Checkpoint selection during the final neural fit follows the evaluation condition in Section 4. Algorithm 1 gives the complete sequence.

3.6 Representation and computational scope

The construction separates representational capacity from empirical utility. A feature-field mean, temporal moment, and signed area encode different properties, but the encoder may already make some order information available through its local responses. D1–D4 are design rationales whose utility is examined by the ablations, not assumptions that every component must improve accuracy.

The attention interactions cost $O(N^2 d)$ per block, with bounded N, instead of $O(L^2 d)$ timestamp attention. The convolutional field still depends on L. The implemented masked reductions cost $O(NLd)$, and its deterministic prefix scan costs $O(Lr^2 \log L)$; therefore, we do not describe the

complete implementation as strictly linear-time. The exact pattern backend requires a dense training Gram matrix and a cubic eigendecomposition in the number of fitting examples; the large-cohort backend replaces this with a fixed-dimensional primal solve. Five additional neural fits and the pattern computations are material training costs. No matched wall-clock speedup is inferred from the token bound.

4 Experimental setup

Datasets and metrics. We evaluate all 30 UEA multivariate classification datasets with their official TRAIN/TEST partitions and all 45,947 TEST examples (Bagnall et al., 2018). Accuracy is the fraction of correct TEST labels. We report its unweighted dataset mean, average within-dataset ranks with average ranks for ties, and exact-count wins/ties/losses. Top-N counts include a method when fewer than N methods have strictly higher accuracy, including ties at the boundary. The common ten-dataset subset is a secondary summary. Appendix A provides dataset metadata, complete scores, and paired statistics.

Baseline selection. The 18 measured methods cover classification-oriented convolution and attention (ConvTran, TSLANet), general-purpose processing (TimesNet, ModernTCN, TimeMixer++), state-space models (MambaSL, TSCMamba), and random-kernel alternatives (MultiRocket, HYDRA, MR+Hydra). Eight forecasting-origin adaptations broaden the comparison: DLinear, PatchTST, Crossformer, FEDformer, iTransformer, Non-stationary Transformers, TimeMixer, and TimeFilter (Zeng et al., 2023; Nie et al., 2023; Zhang & Yan, 2023; Zhou et al., 2022; Liu et al., 2024; 2022; Wang et al., 2024a; Hu et al., 2025). This set tests learned and fixed-feature prediction routes; it is not an exhaustive ranking of classification methods. All 540 baseline cells are executed runs. Literature values remain in separate tables.

Primary training condition. MoSiLa uses seed 42, batch 16, Adam at 10^{-3}, and cosine decay to 5×10^{-5}. Each of five neural folds runs for 100 epochs. Full refitting permits 500 epochs and selects the checkpoint by the combined predictor's TEST correct count; ties update the checkpoint and reset a patience of 50 strictly worse epochs. Thus, gradients use TRAIN, but primary scores are TEST-selection-biased. Deep baselines use the same seed and FP32 tensors with TF32-enabled kernels, without AMP, but retain their recorded optimizer, stopping, and preprocessing recipes. Some baseline loaders normalize TRAIN and TEST separately, unlike MoSiLa's training-fitted scaling. The larger MoSiLa fitting procedure and these differences preclude an equal-budget architectural attribution. Appendix B specifies the deviations, including the unresolved TSCMamba/InsectWingbeat evaluation-loss anomaly.

Supplementary conditions. The TRAIN-loss supplement selects the lowest full-TRAIN neural cross-entropy in evaluation mode, introduces no validation split, and reports seeds 42–44 separately and as mean $\pm$ sample SD. MultiRocket, HYDRA, and MR+Hydra are also summarized over seeds 42–44 using their TRAIN-based ridge selection. Retuned structural variants are reported separately with their actual search counts. External MoSiLa runs use the retained Medformer-derived splits and validation macro-F1/SWA selection, with five model seeds 42–46. Additional measured external comparators and literature reports are tabulated separately. These conditions are analyzed separately from the primary TEST-selected results.

5 Results and discussion

5.1 Classification results on UEA30

MoSiLa reaches 78.739% UEA30 accuracy, 78.661% on the common ten datasets, and mean rank 3.73 (Figure 2). Its UEA30 mean exceeds measured TSLANet and ConvTran by 3.790 and 4.532 percentage points. It attains or ties the best score on 11 datasets and enters the top three on 19. Against measured MambaSL, its wins/ties/losses are 26/1/3. Sixteen of 18 paired comparisons remain below 0.05 after Holm adjustment; TSLANet and ConvTran both have adjusted $p = 0.074$ (Table 5). These results describe the recorded systems under the primary TEST-selected condition.

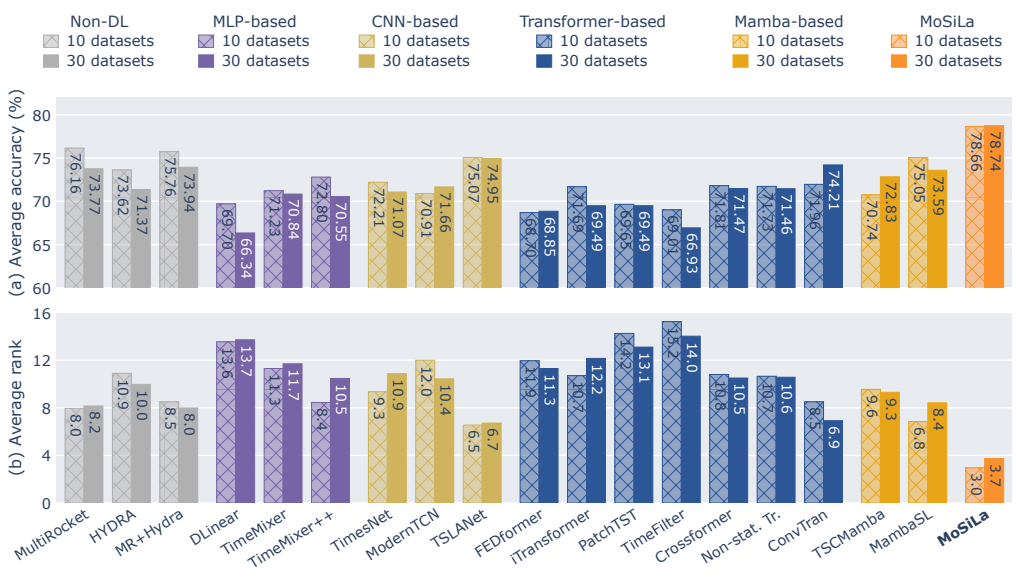


Figure 2: Primary seed-42 comparison, grouped by model family. (a) Macro-average accuracy and (b) mean rank on the common ten datasets (hatched) and UEA30 (solid). Both rankings compare all 19 measured systems, with midranks for ties; lower is better. Training recipes, including TEST-selected neural checkpoints, are specified in Section 4. Exact values and model abbreviations appear in Appendix A.

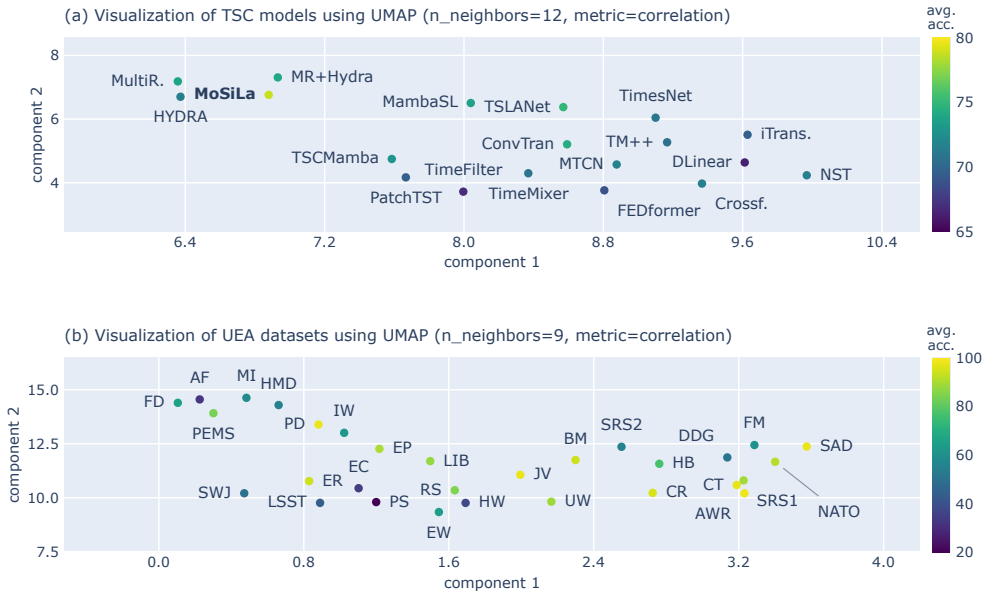


Figure 3: Two views of the same 19×30 primary accuracy matrix using UMAP. (a) Models represented by their 30 dataset accuracies; MoSiLa is labeled in bold. (b) Datasets represented by their 19 model accuracies. Color gives mean accuracy (%) along the respective axis. Both use correlation distance and seed 42; full settings appear in Appendix A.1.

Figure 3 complements the ranking with performance profiles. MoSiLa lies near the random-kernel systems in the model view, while the dataset view separates several difficult datasets from the high-

accuracy group. These are descriptive neighborhoods of measured accuracies, not evidence of shared internal representations. The selected full model uses the neural expert on 15 datasets, the pattern expert on seven, and mixtures on eight; paired ablations below assess component effects.

Measured versus reported performance. Measured MambaSL reaches 73.587%, compared with its published 79.819%. MoSiLa exceeds this remeasurement but remains below that published score. All 13 full-UEA30 reference comparisons from MambaSL Table 5 are lower in our baseline reruns. Appendix E retains the complete comparison; recovered settings, different seeds, arithmetic, and search conditions do not isolate a single cause of these gaps.

5.2 Structural ablations

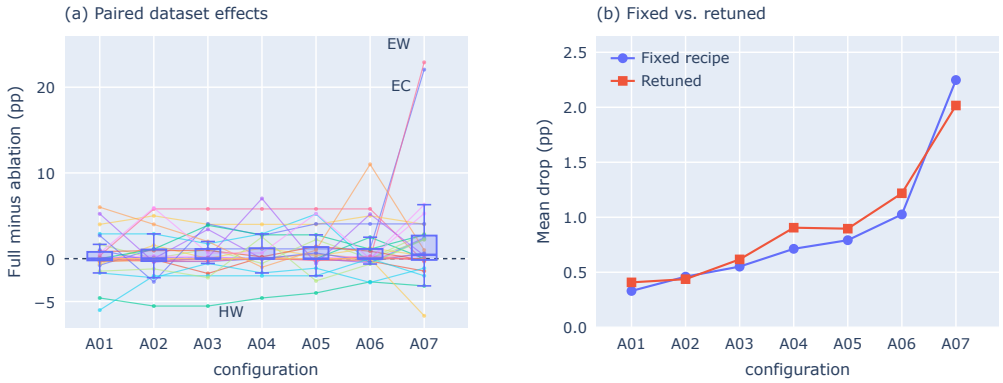


Figure 4: Component effects. (a) Full-model accuracy minus each ablation on 30 datasets. Lines follow individual datasets across variants; boxes span the interquartile range, center lines mark medians, and whiskers extend to observations within 1.5 times the interquartile range. All observations, including zeros and negative effects, are retained. (b) Mean drops under the fixed recipe and separate retuning, with 18–35 searched configurations per variant. Variants and full numerical results appear in Appendix C.

Every fixed-recipe ablation has lower mean accuracy than A00, but the dataset-level effects vary (Figure 4a). Removing path, moment, or relation bias lowers the mean by 0.330, 0.460, and 0.551 points. Path removal yields 13/9/8 wins/ties/losses for A00; its benefit on LSST and MI coexists with losses on HW. Mean, max, and last-token readouts lower the mean by 0.713, 0.791, and 1.026 points. Table 8 retains accuracies and paired tests, and Figure 6 gives all dataset-level differences.

Removing the pattern expert produces the largest mean drop, 2.248 points. The large effects on EW (22.901 points) and EC (22.053) coexist with an improvement on SRS1 after removal (91.809 versus 91.468%). Its exploratory, unadjusted paired $p = 0.020$ is reported separately from the Holm-corrected primary comparisons.

After separate retuning, A00 reaches 79.152% and every ablation remains lower by 0.408–2.016 points (Figure 4b). This preserves the direction of the mean effects, without establishing equal-budget optimized ceilings. Two relation blocks also have the highest tested mean (78.739%, versus 77.850% and 77.801% for one and three); neither paired depth comparison reaches $p < 0.05$ (Appendix D).

5.3 Additional evaluation conditions

TRAIN-loss checkpoint selection. MoSiLa obtains 76.570% $\pm$ 0.439% over seeds 42–44, 2.169 points below its primary score, while exceeding the three classical baselines in mean accuracy (Table 1). Its seed-42 result is 76.469%. This quantifies the changed checkpoint condition, not removal of earlier TEST-label influence on development. Appendix F retains all dataset-level repeats and seed-42 paired comparisons.

Table 1: TRAIN-based selection, seeds 42–44. Accuracy: UEA30 macro-average, mean ± sample SD across seeds; advantage: difference of seed means in percentage points. MoSiLa uses TRAIN-loss checkpoints and classical methods TRAIN-based ridge selection; preprocessing differences remain (Appendix B).

Model	Accuracy: mean ± SD	MoSiLa advantage
MoSiLa	$\mathbf{76.570 \pm 0.439}$	–
MultiRocket	73.557 ± 0.445	3.013
Hydra	71.436 ± 0.119	5.134
MR+Hydra	73.755 ± 0.287	2.816

Table 2: Measured external results (percent, mean ± SD). MoSiLa uses five seeds and validation macro-F1 selection. Comparators retain recorded statistics; NR: unavailable, †: external-only comparator. Bold: highest measured mean. Literature reports: Table 17.

	ADFTD		FLAAP	
Measured system	Accuracy	Macro-F1	Accuracy	Macro-F1
MoSiLa	$\mathbf{51.473 \pm 4.590}$	$\mathbf{48.647 \pm 5.401}$	$\mathbf{69.250 \pm 0.601}$	$\mathbf{69.061 \pm 0.590}$
Crossformer	49.718 ± 1.267	46.917 ± 1.405	67.763 ± 0.747	66.849 ± 1.240
Reformer†	47.009 ± 1.462	45.246 ± 1.505	64.308 ± 0.705	63.526 ± 0.635
Transformer†	48.411 ± 0.745	45.777 ± 1.230	65.512 ± 0.711	65.397 ± 0.681
TCN†	NR	NR	64.812 ± 0.707	65.005 ± 0.906
ModernTCN	NR	NR	67.816 ± 0.760	67.770 ± 0.727
Mamba	NR	NR	68.297 ± 0.905	68.539 ± 0.640

External datasets. On ADFTD, MoSiLa obtains 51.473% accuracy and 48.647% macro-F1; on FLAAP it obtains 69.250% and 69.061% (Table 2). Accuracy exceeds the strongest additional measured comparator by 1.755 points on ADFTD (Crossformer) and 0.953 on FLAAP (Mamba). However, FLAAP remains below the reported MambaSL and Medformer accuracies, 77.47% and 76.44%. ADFTD is numerically close to MambaSL's reported 51.68% accuracy and 48.73% macro-F1. These unpaired literature comparisons use different searches and seeds, so external gains over measured comparators do not imply a uniform advantage over published results.

6 Conclusion

We studied regional representation and prediction-route selection as separate decisions in multivariate time series classification. MoSiLa exposes feature level, temporal moment, and low-rank path coordinates through a bounded multiscale token set, while an independent pattern expert supplies an alternative prediction route. The complete A00 system uses the global landmark readout and reaches 78.739% macro-average accuracy, the highest among the 18 baselines measured under the recorded TEST-selected recipes. Every listed fixed-recipe structural ablation has a lower macro-average, with the largest reduction after removing the pattern expert. Separately retuned variants preserve this full-versus-ablated ordering, and three-seed TRAIN-based-selection comparisons favor MoSiLa over the three classical baselines in mean accuracy.

Three limits define this conclusion. TEST reuse, heterogeneous budgets, and baseline reproduction gaps restrict the primary comparison to the recorded systems; the differences from TSLANet and ConvTran do not survive Holm adjustment. Fixed normalized-time regions and low-order statistics can discard distinctions, and their utility depends on information already learned by the encoder. Finally, the empirical gate is not a risk guarantee, and external performance is mixed relative to literature reports. The TRAIN-loss-selected A00 repeats (76.570% ± 0.439%) also remain below the TEST-selected primary score. These results support the evaluated system under its declared conditions without establishing an equal-budget or cross-protocol advantage.

AI USE STATEMENT

The authors acknowledge the use of ChatGPT-5 and Google Translate to improve the clarity and concision of academic English. OpenAI Codex also assisted with programming and debugging, literature retrieval, results analysis, manuscript drafting and revision, and figure and table preparation. The authors are responsible for the final manuscript and research findings.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices B.4–B.6 define the model and fitting procedure. Section 4 and Appendix B distinguish all checkpoint-selection conditions and implementation recipes. The appendices retain per-dataset measurements, paired statistics, repeated runs, and separate measured and reported comparisons. The prepared anonymous MoSiLa code package contains the proposed-method implementation, experiment entry points, dependencies, and third-party notices. It excludes datasets, pretrained weights, and the 18 third-party baseline implementations; baseline configurations are supplied separately from source-code redistribution. Main training checkpoints were not retained, so reproducing predictions requires retraining. Package smoke checks verify execution paths, not reproduction of the full benchmark.

REFERENCES

Md Atik Ahamed and Qiang Cheng. TSCMamba: Mamba meets multi-view learning for time series classification. *Information Fusion*, 120:103079, 2025. doi: 10.1016/j.inffus.2025.103079. URL https://doi.org/10.1016/j.inffus.2025.103079.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. arXiv:1811.00075, 2018. URL https://arxiv.org/abs/1811.00075.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. Hydra: Competing convolutional kernels for fast and accurate time series classification. *Data Mining and Knowledge Discovery*, 37:1779–1805, 2023. doi: 10.1007/s10618-023-00939-3. URL https://doi.org/10.1007/s10618-023-00939-3.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, and Xiaoli Li. TSLANet: Rethinking transformers for time series representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12409–12428, 2024. URL https://proceedings.mlr.press/v235/eldele24a.html.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34:1936–1962, 2020. doi: 10.1007/s10618-020-00710-y. URL https://doi.org/10.1007/s10618-020-00710-y.

Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I. Webb, and Mahsa Salehi. Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery*, 38(1):22–48, 2024. doi: 10.1007/s10618-023-00948-2. URL https://doi.org/10.1007/s10618-023-00948-2.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24893–24911, 2025. URL https://proceedings.mlr.press/v267/hu25ac.html.

Yoo-Min Jung and Leekyung Kim. MambaSL: Exploring single-layer mamba for time series classification. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YDl4vqQqGP.

Prabhat Kumar and S. Suresh. FLAAP: An open human activity recognition (HAR) dataset for learning and finding the associated activity patterns. *Procedia Computer Science*, 212:64–73, 2022. doi: 10.1016/j.procs.2022.10.208. URL https://doi.org/10.1016/j.procs.2022.10.208.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0718. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6):95, 2023. doi: 10.3390/data8060095. URL https://doi.org/10.3390/data8060095.

Fernando Moreno-Pino, Álvaro Arroyo, Harrison Waldon, Xiaowen Dong, and Álvaro Cartea. Rough transformers: Lightweight and continuous time series modelling through signature patching. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3371. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bfe167fee3be2d862a56af82dee77720-Abstract-Conference.html.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://github.com/yuqinie98/PatchTST.

Chang Wei Tan, Angus Dempster, Christoph Bergmeir, and Geoffrey I. Webb. MultiRocket: Multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, 36:1623–1646, 2022. doi: 10.1007/s10618-022-00844-1. URL https://doi.org/10.1007/s10618-022-00844-1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=7oLshfEIC2.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2b187165e28fdfdc0ffb34d1bfff2b0c-Abstract-Conference.html.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36314–36341, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023. URL https://github.com/Thinklab-SJTU/Crossformer.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

A DATA AND COMPLETE PRIMARY RESULTS

Reporting conventions. Accuracy and macro-F1 are reported in percent. Measured results, retuned variants, repeated runs, and literature values are tabulated under their respective evaluation conditions. NR denotes a value absent from the corresponding result record or reference; n/a denotes a self-comparison. Neither denotes zero performance. Table 3 gives the UEA30 metadata and abbreviations used throughout the supplementary tables. The first ten rows define the common ten-dataset subset. Dataset names, sizes, and sequence ranges follow the archive and the reference benchmark description (Bagnall et al., 2018; Jung & Kim, 2026); TEST counts are also checked against our completed result records. Variable-length examples retain their actual lengths in MoSiLa. Missing entries within a channel are interpolated within that sequence; a wholly missing sample channel uses the current fitting subset's channel mean. Non-missing channel means and sample standard deviations are fitted only on the fitting subset. Channels with near-zero standard deviation use scale one. Padding is not included in these statistics.

Table 3: Summary of the 30 UEA datasets. Length gives the observed sequence-length range. The first ten datasets define the common subset.

Dataset	Code	TRAIN	TEST	Length	Channels	Classes
EthanolConcentration	EC	261	263	1751	3	4
FaceDetection	FD	5890	3524	62	144	2
Handwriting	HW	150	850	152	3	26
Heartbeat	HB	204	205	405	61	2
JapaneseVowels	JV	270	370	7–29	12	9
PEMS-SF	PEMS	267	173	144	963	7
SelfRegulationSCP1	SRS1	268	293	896	6	2
SelfRegulationSCP2	SRS2	200	180	1152	7	2
SpokenArabicDigits	SAD	6599	2199	4–93	13	10
UWaveGestureLibrary	UW	120	320	315	3	8
ArticularyWordRecognition	AWR	275	300	144	9	25
AtrialFibrillation	AF	15	15	640	2	3
BasicMotions	BM	40	40	100	6	4
CharacterTrajectories	CT	1422	1436	60–182	3	20
Cricket	CR	108	72	1197	6	12
DuckDuckGeese	DDG	50	50	270	1345	5
EigenWorms	EW	128	131	17984	6	5
Epilepsy	EP	137	138	206	3	4
ERing	ER	30	270	65	4	6
FingerMovements	FM	316	100	50	28	2
HandMovementDirection	HMD	160	74	400	10	4
InsectWingbeat	IW	25000	25000	2–22	200	10
Libras	LIB	180	180	45	2	15
LSST	LSST	2459	2466	36	6	14
MotorImagery	MI	278	100	3000	64	2
NATOPS	NATO	180	180	51	24	6
PenDigits	PD	7494	3498	8	2	10
PhonemeSpectra	PS	3315	3353	217	11	39
RacketSports	RS	151	152	30	6	4
StandWalkJump	SWJ	12	15	2500	4	3

Ranks and W/T/L use exact correct counts. Paired tests use dataset accuracy differences, omit zeros, assign midranks to tied absolute differences, and evaluate the two-sided conditional sign distribution exactly. The comparison unit is the dataset. Supplementary ablation, depth, and seed-42 classical-baseline tests use the same procedure and report unadjusted exploratory p values.

Table 4: Primary measured comparison. Accuracies are percentages. W/T/L counts MoSiLa wins, ties, and losses against the listed method, using exact correct counts rather than rounded table entries. Mean ranks are across these 19 systems. Every row uses a completed seed-42 run on each of the 30 datasets, with the recipe differences in Appendix B.

Model	Avg. acc. (30 datasets)	Avg. acc. (10 datasets)	Avg. rank	W/T/L
MoSiLa	**78.739**	**78.661**	**3.73**	–
TSLANet	74.949	75.069	6.72	15/7/8
ConvTran	74.208	71.958	6.92	18/5/7
MR+Hydra	73.938	75.761	7.98	22/3/5
MultiRocket	73.766	76.160	8.15	22/3/5
MambaSL	73.587	75.051	8.42	26/1/3
TSCMamba	72.834	70.737	9.30	23/2/5
ModernTCN	71.663	70.907	10.43	24/1/5
Crossformer	71.475	71.809	10.50	24/2/4
Non-stationary Tr.	71.457	71.727	10.57	27/2/1
Hydra	71.368	73.623	9.97	24/3/3
TimesNet	71.073	72.206	10.87	28/0/2
TimeMixer	70.839	71.228	11.70	27/2/1
TimeMixer++	70.547	72.802	10.47	28/1/1
PatchTST	69.494	69.646	13.10	26/1/3
iTransformer	69.491	71.688	12.15	28/0/2
FEDformer	68.851	68.695	11.30	26/1/3
TimeFilter	66.933	69.010	14.02	27/0/3
DLinear	66.342	69.696	13.72	25/2/3

Table 5: Statistical companion to Table 4. Top-N includes ties at the boundary: fewer than N methods have strictly higher accuracy. Paired two-sided signed-rank tests compare MoSiLa with each baseline across the 30 datasets. Holm adjustment covers these 18 comparisons; n/a is the self-comparison.

Model	Top-1	Top-3	Paired p	Holm p
MoSiLa	**11**	**19**	n/a	n/a
TSLANet	6	13	0.041	0.074
ConvTran	6	12	0.037	0.074
MR+Hydra	5	12	2.10×10^{-4}	0.001
MultiRocket	3	10	4.79×10^{-4}	0.002
MambaSL	0	5	2.76×10^{-5}	3.03×10^{-4}
TSCMamba	2	8	5.83×10^{-4}	0.002
ModernTCN	1	4	2.37×10^{-4}	0.001
Crossformer	1	2	8.34×10^{-5}	6.67×10^{-4}
Non-stationary Tr.	1	1	3.33×10^{-6}	4.66×10^{-5}
Hydra	4	9	2.83×10^{-7}	4.81×10^{-6}
TimesNet	0	1	5.59×10^{-7}	8.94×10^{-6}
TimeMixer	1	4	3.33×10^{-6}	4.66×10^{-5}
TimeMixer++	0	2	1.64×10^{-6}	2.46×10^{-5}
PatchTST	1	2	8.72×10^{-5}	6.67×10^{-4}
iTransformer	0	1	1.64×10^{-7}	2.95×10^{-6}
FEDformer	2	5	3.63×10^{-5}	3.27×10^{-4}
TimeFilter	1	1	1.44×10^{-5}	1.73×10^{-4}
DLinear	1	2	3.18×10^{-5}	3.18×10^{-4}

Table 6: Classification accuracy (%) of all 19 measured systems on UEA30, seed 42. Bold and underline mark the best and second-best distinct scores, including ties. Model families are grouped in the header; MoSiLa appears in both panels. Summary rows use the same exact-count results and paired tests as Tables 4–5.

(a) Non-DL, MLP-based, and CNN-based models.

	Non-DL			MLP-based			CNN-based			Proposed
Dataset	Multi-Rocket	HYDRA	MR+ Hydra	DLinear	TimeMixer	Time-Mixer++	TimesNet	Modern TCN	TSLANet	**MoSiLa (proposed)**
EC	53.232	52.091	52.091	27.757	27.757	31.939	28.897	31.559	28.897	**56.274**
FD	60.726	52.866	59.024	67.963	68.700	67.310	68.530	65.579	62.940	69.098
HW	52.000	47.529	52.118	23.176	30.588	32.588	35.529	26.353	**55.529**	52.588
HB	76.098	78.049	77.561	76.098	76.585	78.537	75.610	75.122	**80.976**	78.537
JV	98.919	98.378	98.919	96.486	93.514	97.568	97.297	97.838	98.919	**99.459**
PEMS	79.769	80.347	80.347	80.347	87.283	86.127	82.081	82.659	82.081	**89.017**
SRS1	91.468	85.324	89.078	**93.515**	89.420	90.102	89.078	92.833	87.713	91.468
SRS2	55.556	52.778	55.000	53.333	52.778	53.333	57.778	51.667	60.556	57.222
SAD	99.454	98.863	99.409	96.407	99.090	98.954	99.136	98.272	**99.955**	99.818
UW	94.375	90.000	94.062	81.875	86.562	91.562	88.125	87.188	93.125	93.125
AWR	**99.333**	**99.333**	**99.333**	97.333	97.000	98.667	98.000	98.333	99.000	99.000
AF	6.667	6.667	6.667	40.000	40.000	26.667	33.333	26.667	46.667	40.000
BM	**100.000**	**100.000**	**100.000**	85.000	**100.000**	95.000	95.000	97.500	**100.000**	**100.000**
CT	99.304	98.816	99.304	97.772	98.468	98.816	97.632	98.538	99.373	**99.652**
CR	97.222	98.611	98.611	91.667	95.833	93.056	91.667	97.222	98.611	98.611
DDG	50.000	56.000	46.000	66.000	50.000	54.000	62.000	50.000	**68.000**	66.000
EW	77.863	64.885	83.969	42.748	50.382	58.779	60.305	56.489	75.573	**97.710**
EP	**100.000**	**100.000**	**100.000**	46.377	93.478	83.333	89.130	95.652	97.826	**100.000**
ER	98.519	98.519	**99.630**	90.000	89.630	93.704	94.074	92.222	86.667	98.148
FM	66.000	56.000	61.000	60.000	62.000	60.000	59.000	66.000	59.000	63.000
HMD	35.135	22.973	36.486	63.514	48.649	66.216	59.459	59.459	62.162	54.054
IW	59.548	59.472	61.928	19.188	65.012	63.864	63.512	**75.200**	55.860	69.552
LIB	92.778	92.778	92.778	73.333	81.111	84.444	85.000	88.333	**96.111**	**96.111**
LSST	46.715	54.866	47.324	33.090	36.577	29.521	40.065	40.998	38.281	**65.450**
MI	51.000	50.000	49.000	59.000	56.000	60.000	59.000	64.000	63.000	61.000
NATO	93.889	86.111	93.333	95.556	95.000	91.667	94.444	94.444	96.111	96.111
PD	96.827	96.398	96.884	91.995	98.027	97.970	98.599	97.885	98.628	**98.999**
PS	32.419	**33.254**	32.806	6.472	14.793	15.151	14.345	13.630	24.068	32.747
RS	88.158	90.132	88.816	80.921	80.921	84.211	82.237	81.579	86.184	**92.763**
SWJ	60.000	40.000	**66.667**	53.333	60.000	33.333	33.333	46.667	46.667	46.667
Avg. acc. (10)	76.160	73.623	75.761	69.696	71.228	72.802	72.206	70.907	75.069	**78.661**
Avg. acc. (30)	73.766	71.368	73.938	66.342	70.839	70.547	71.073	71.663	74.949	**78.739**
Avg. rank (10)	7.95	10.90	8.50	13.55	11.30	8.45	9.35	12.00	6.55	**2.95**
Avg. rank (30)	8.15	9.97	7.98	13.72	11.70	10.47	10.87	10.43	6.72	**3.73**
Top-1	3	4	5	1	1	0	0	1	6	11
Top-3	10	9	12	2	4	2	1	4	13	19
W/T/L	22/3/5	24/3/3	22/3/5	25/2/3	27/2/1	28/1/1	28/0/2	24/1/5	15/7/8	–
Paired p	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.041	–
Holm p	0.002	< 0.001	0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001	0.074	–

MoSiLa makes 33207 correct predictions among 45,947 TEST examples, yielding 72.272% micro-accuracy. Macro and micro averages weight different units: one mistake changes accuracy by 6.667 points on a 15-example dataset, while InsectWingbeat contributes 25,000 examples to the micro-average. Compared with measured MambaSL, MoSiLa is higher on EC (56.274 versus 31.939%), EW (97.710 versus 78.626%), and LSST (65.450 versus 41.281%), but lower on HMD (54.054 versus 67.568%) and SRS1 (91.468 versus 92.150%). These comparisons describe the full predictors and do not isolate a component's contribution.

Compact model labels in Figures 2–3 are MultiR. (MultiRocket), MTCN (ModernTCN), TM++ (TimeMixer++), iTrans. (iTransformer), Crossf. (Crossformer), and NST (Non-stationary Transformer). All dataset abbreviations are defined in Table 3.

Table 6: (Continued.) Classification accuracy (%) on UEA30. Highlighting and rankings still use all 19 systems, including models in panel (a). W/T/L compares MoSiLa with each method; Holm adjustment covers the 18 primary comparisons.

(b) Transformer-based and Mamba-based models.

	Transformer-based							Mamba-based		Proposed
Dataset	FED-former	iTrans-former	PatchTST	Time-Filter	Cross-former	Non-stat. Trans.	ConvTran	TSC-Mamba	MambaSL	**MoSiLa (proposed)**
EC	27.757	27.757	27.757	28.137	36.502	26.996	28.517	26.236	31.939	**56.274**
FD	**69.268**	66.175	68.473	64.898	68.275	67.423	66.657	67.764	66.941	69.098
HW	30.353	30.000	29.765	24.706	29.059	34.000	26.235	41.176	55.176	52.588
HB	74.146	77.073	69.756	71.220	78.049	77.561	78.537	77.073	78.537	78.537
JV	95.405	97.297	94.865	95.135	97.297	97.838	98.108	97.568	96.486	**99.459**
PEMS	87.861	86.127	83.815	85.549	79.191	84.393	81.503	84.971	83.237	**89.017**
SRS1	69.625	91.809	84.983	85.666	91.468	90.785	91.468	92.150	92.150	91.468
SRS2	55.556	55.000	53.333	52.778	55.556	53.333	**61.111**	54.444	55.556	57.222
SAD	98.545	98.454	98.090	97.635	98.317	99.000	99.636	70.987	99.545	99.818
UW	78.438	87.188	85.625	84.375	84.375	85.938	87.812	**95.000**	90.938	93.125
AWR	90.333	98.333	98.000	98.000	98.667	98.333	98.333	98.000	97.667	99.000
AF	40.000	33.333	**60.000**	26.667	33.333	26.667	46.667	46.667	33.333	40.000
BM	92.500	92.500	77.500	65.000	**100.000**	**100.000**	**100.000**	**100.000**	82.500	**100.000**
CT	68.593	97.841	98.259	97.354	98.398	98.468	99.513	99.164	99.304	**99.652**
CR	83.333	93.056	91.667	93.056	95.833	94.444	**100.000**	98.611	94.444	98.611
DDG	30.000	56.000	26.000	28.000	56.000	62.000	66.000	26.000	60.000	66.000
EW	61.069	54.962	51.908	54.962	54.198	54.198	59.542	80.916	78.626	**97.710**
EP	94.928	78.986	93.478	61.594	90.580	90.580	97.101	97.101	97.101	**100.000**
ER	95.926	93.704	94.074	92.963	94.815	93.333	91.481	92.593	91.852	98.148
FM	61.000	56.000	57.000	59.000	56.000	58.000	**71.000**	57.000	60.000	63.000
HMD	**71.622**	48.649	59.459	62.162	59.459	62.162	50.000	52.703	67.568	54.054
IW	65.100	73.124	59.768	54.456	71.948	65.448	71.440	71.024	65.392	69.552
LIB	90.000	89.444	78.333	82.222	86.667	88.889	91.111	86.667	86.111	**96.111**
LSST	36.861	37.794	55.637	52.920	37.429	36.659	40.511	58.110	41.281	**65.450**
MI	58.000	59.000	61.000	62.000	62.000	60.000	**65.000**	59.000	55.000	61.000
NATO	87.778	86.111	79.444	78.889	88.889	93.889	**98.333**	85.556	95.000	96.111
PD	98.914	98.199	97.570	97.885	98.285	98.313	98.799	96.827	98.628	**98.999**
PS	11.035	10.498	15.001	11.065	5.458	13.510	28.959	22.905	29.824	32.747
RS	81.579	76.974	80.921	73.026	84.868	84.868	86.184	88.816	90.132	**92.763**
SWJ	60.000	33.333	53.333	**66.667**	53.333	46.667	46.667	60.000	33.333	46.667
Avg. acc. (10)	68.695	71.688	69.646	69.010	71.809	71.727	71.958	70.737	75.051	**78.661**
Avg. acc. (30)	68.851	69.491	69.494	66.933	71.475	71.457	74.208	72.834	73.587	**78.739**
Avg. rank (10)	11.95	10.70	14.25	15.25	10.80	10.65	8.50	9.55	6.85	**2.95**
Avg. rank (30)	11.30	12.15	13.10	14.02	10.50	10.57	6.92	9.30	8.42	**3.73**
Top-1	2	0	1	1	1	1	6	2	0	11
Top-3	5	1	2	1	2	1	12	8	5	19
W/T/L	26/1/3	28/0/2	26/1/3	27/0/3	24/2/4	27/2/1	18/5/7	23/2/5	26/1/3	–
Paired p	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.037	< 0.001	< 0.001	–
Holm p	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.074	0.002	< 0.001	–

A.1 Performance-profile visualization

Figure 3 uses the same exact-count accuracy matrix as the primary tables: 19 systems by 30 datasets, with no excluded observations or additional column scaling. UMAP (McInnes et al., 2018), implemented with `umap-learn` 0.5.8, embeds its rows and columns separately. Both fits use correlation distance, two output dimensions, spectral initialization, minimum distance 0.1, 500 optimization epochs, and random seed 42 with one execution thread. The model view uses 12 neighbors and the dataset view uses nine, following the visualization settings of Jung & Kim (2026). The coordinates are fitted once under these settings, without selecting a favorable layout across random seeds. Color shows the model's mean across datasets or the dataset's mean across models; each view has its own color scale. These projections summarize performance profiles, not learned sample representations or evidence of causal complementarity.

B IMPLEMENTATION AND FITTING DETAILS

B.1 MOSILA CONFIGURATION AND REPRODUCIBILITY SCOPE

MoSiLa combines a neural landmark classifier, an independent pattern expert, and a prediction-selection stage. The neural classifier uses a convolutional feature field, explicit landmark statistics, relation blocks, and a global landmark readout, as described in Section 3. The pattern expert and selection stage are included in the complete system reported as A00 in the ablation tables.

The main experiment uses one architecture across datasets: encoder width 128, bottleneck/branch width 32, three local units, partitions $(1, 4, 16, 64)$, path rank eight, two relation blocks, four attention heads, and feed-forward width 192. No attention dropout is present in these relation blocks. TRAIN examples are shuffled each epoch. The final partial batch is retained; on a dataset with fewer than 16 fitting examples, its single batch necessarily contains fewer than 16. Gradients are clipped to Euclidean norm four. Adam uses its implementation defaults apart from learning rate 0.001. Each fold's cosine schedule has a 100-epoch horizon; the primary full fit has a 500-epoch horizon. The full-fit seed is 42 and fold fits use 42 plus the fold index. Table 7 reports actual selected epochs, completed epochs, and trainable neural parameter counts; fixed random filters and ridge state are additional non-neural state.

Runs were scheduled as independent jobs on RTX 3090 servers, not as a single distributed-data-parallel fit. FP32 tensors with TF32-enabled matrix and convolution kernels are distinct from BF16/FP16 automatic mixed precision. Feature scaling, ridge eigensolves, and gate fitting use double precision where implemented; saying "TF32" does not imply that every operation uses Tensor Cores or low-precision arithmetic. CUDA graph replay and chunked evaluation were used on supported paths. This paper makes no equal-hardware timing comparison from logs gathered under different concurrency and offloading conditions. Result records, resolved settings, and saved predictions provide the evidence for the tables; deleted inference checkpoints cannot be reconstructed without retraining.

B.2 BASELINE RECIPES AND DEVIATIONS

The measured set contains ten models evaluated through the MambaSL author adapter: MambaSL, iTransformer, TimesNet, ModernTCN, TimeMixer++, TSCMamba, PatchTST, Crossformer, FEDformer, and DLinear. Their dataset-specific recovered configurations are used rather than rerunning the reference's full hyperparameter searches. The standard neural author loop uses RAdam, learning rate 0.001, at most 100 epochs, and patience ten; its score is TEST accuracy minus 10^{-5} times TEST cross-entropy. It also permits early exit at perfect TEST accuracy or infinite evaluation loss. Changing the seed to 42 and enabling TF32 means these are new runs of the recovered recipe, not bitwise reproductions of the paper's seed-2021 scores.

TimeFilter, TimeMixer, and Non-stationary Transformer use recorded classification adaptations and fixed configurations. Their inclusion broadens the architectural comparison but does not imply that their original forecasting papers specify a final UEA30 classification configuration. The supplementary configuration file lists the settings used for each model–dataset pair; the original experiment records retain the complete resolved settings.

ConvTran uses its executable classification defaults: embedding width 16, eight heads, feed-forward width 256, dropout 0.01, tAPE positional encoding, eRPE relative encoding, and RAdam at 0.001. These differ from the paper's width-64/Adam description and are disclosed rather than described as exact paper reproduction. TSLANet uses embedding width 128, depth two, patch size eight, masking ratio 0.4, dropout 0.15, adaptive spectral and interactive convolution blocks enabled, and AdamW at 0.001 with weight decay 10^{-4}. It runs 50 reconstruction-pretraining epochs and 100 classification epochs with label smoothing. Its pretraining checkpoint uses TEST reconstruction loss, and its classification checkpoint uses TEST accuracy. Both baselines use a nominal batch size of 16, with the recorded small-dataset exception for drop-last pretraining; classification ties retain the earlier best checkpoint. They have not been silently recast as TRAIN-validation-selected results.

MultiRocket, HYDRA, and their combination use the corresponding `aeon` classifiers, random state 42, four CPU jobs, and internal TRAIN-based ridge selection. Inputs are zero-padded to a common train/test maximum length with the recorded minimum length of nine, or ten for the affected short-

series repair. They do not have neural checkpoints or a TF32 execution mode. These distinctions are why the primary comparison is described as a system comparison under documented recipes rather than a perfectly unified training protocol.

The upstream UEA loaders used by the baselines standardize each split separately. This transductive use of TEST input statistics differs from MoSiLa's training-fitted normalization even though it does not directly use TEST labels for scaling. Memory-related implementation repairs included chunked attention/evaluation, CPU-side statistics, and selective offloading; they are not evidence that all methods have identical resource requirements. The completed TSCMamba InsectWingbeat record is 17,756/25,000 correct, but the retained run notes report anomalously large TEST cross-entropy (for example, 32,962.117 at epoch 24). This unresolved anomaly does not invalidate the saved correct count, but the cell must not be treated as evidence of a healthy optimization trace or the architecture's attainable accuracy.

Table 7: MoSiLa's actual primary prediction-route choices and neural fit records. A pattern-only choice still incurred the implemented neural training cost.

Dataset	Prediction route	Epoch selected / run	Neural parameters
EC	pattern	500 / 500	504924
FD	half	3 / 53	531738
HW	half	29 / 79	507762
HB	pattern	500 / 500	515802
JV	neural	10 / 60	507297
PEMS	conditional	25 / 75	689631
SRS1	pattern	500 / 500	505242
SRS2	neural	9 / 59	505434
SAD	half	139 / 189	507618
UW	neural	4 / 54	505440
AWR	neural	496 / 500	508785
AF	neural	3 / 53	504603
BM	neural	500 / 500	505500
CT	neural	88 / 138	506988
CR	neural	500 / 500	506532
DDG	neural	32 / 82	762717
EW	pattern	500 / 500	505629
EP	pattern	500 / 500	504924
ER	pattern	500 / 500	505374
FM	neural	19 / 69	509466
HMD	neural	48 / 98	506268
IW	half	154 / 204	543522
LIB	neural	61 / 111	506151
LSST	half	155 / 205	506790
MI	neural	11 / 61	516378
NATO	neural	81 / 131	509214
PD	conditional	58 / 108	505506
PS	constant	71 / 121	510975
RS	neural	25 / 75	505500
SWJ	pattern	500 / 500	504987

B.3 Illustrative landmark statistics

Figure 5 separates the distinctions encoded by the landmark coordinates. Its examples concern the constructed feature field and projected path, not arbitrary permutations of the raw input: changing the raw order can also change the convolutional responses. They illustrate representational capacity; the ablations evaluate empirical utility.

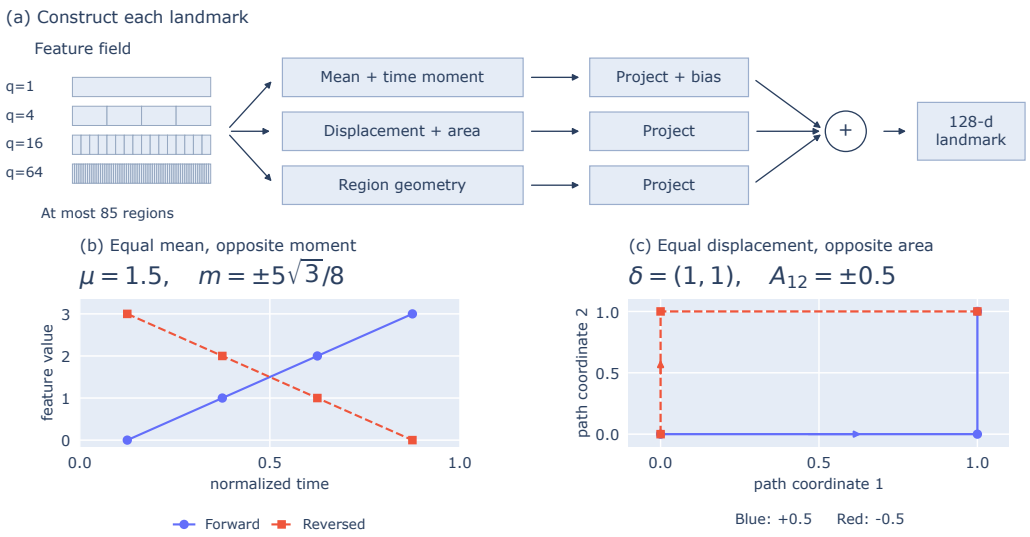


Figure 5: Landmark construction and the distinctions encoded by its statistics. (a) Each nonempty region produces one token by Equation 3. (b) Reversing a four-point feature field preserves its mean and reverses its temporal moment. (c) Two paths with the same endpoints have equal displacement and opposite signed area. Panels (b,c) are analytical examples illustrating Equations 1–2, not measured activations or performance results.

B.4 Path-prefix construction

Let A_{0t} be the signed area of the path from p_0 to p_t and let $\Delta_{ab} = p_b - p_a$. The piecewise-linear concatenation identity gives

$$A_{ab} = A_{0b} - A_{0a} - \tfrac{1}{2}(p_a - p_0) \wedge \Delta_{ab}. \tag{8}$$

To verify the identity, split the prefix sum at a. In the remaining terms write $p_t - p_0 = (p_t - p_a) + (p_a - p_0)$. The first contribution is A_{ab} and the second telescopes to $\frac{1}{2}(p_a - p_0) \wedge (p_b - p_a)$. This is the standard algebra behind the implemented operator, not a new theoretical result. Empty bins are masked and singleton bins give zero. The area transformation is asinh, an invertible coordinate transformation before the learned linear map, not clipping of nonfinite values.

The code implements prefix summation through a deterministic doubling scan. Endpoint selection uses elementwise multiplication and reduction rather than a TF32 matrix multiplication, to avoid rounding two large prefix coordinates before subtracting them. These numerical choices preserve the intended discrete formula within floating-point arithmetic; they do not imply exact invariance to sequence discretization. Means and moments still use masked reductions over valid observations.

B.5 Pattern features and ridge fitting

There are 64 groups in total, split equally between raw and first-difference views; each has eight kernels of length nine. Kernels are sampled from a normal distribution, centered across their nine coefficients, and divided by their absolute coefficient sum. Each group sums $\min(\max(\lfloor C/2 \rfloor, 2), 8)$ sampled channels, with replacement. Dilations are powers of two up to $2^{\max(0, \lfloor \log_2(\max(1, (L_{\max}-1)/8)) \rfloor)}$, determined from the full TRAIN maximum length before folds.

For a group and lag τ, let $J_{ab}^{(\tau)}$ be the empirical frequency of winner a followed by winner b over valid lagged pairs. Its order feature is $J_{ab}^{(\tau)} - J_{a+}^{(\tau)} J_{+b}^{(\tau)}$. Lags are the unique elements of $\{1, \text{dilation}\}$. Occurrence features use the nonnegative square root followed by sparse-count scaling; signed order features use $\operatorname{sign}(x)\sqrt{|x|}$ followed by standardization. Each block is training-centered and normalized by its root-mean-square row norm. Scaling is learned separately within every fit, including the base-model folds.

The weight grid is $\omega \in \{0, 0.1, 1\}$; the ridge grid is $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$. Fixed-design leave-one-out accuracy selects a pair, with mean squared error breaking ties. Temperature is chosen by cross-entropy from $\{0.03125, 0.0625, 0.125, 0.25, 0.5, 1, 2\}$. For an exact fit, if $\hat{Y}$ is the fitted target matrix and h_i includes the unpenalized-intercept leverage, the leave-one-out row is $Y_i - (Y_i - \hat{Y}_i)/(1 - h_i)$. The Gaussian-projection backend uses 512 coordinates per feature block, projection seed 42, and 512-example fitting chunks. Its calculations use all training examples. Factorizing the fitted readout is an inference implementation choice, not a different selection objective.

B.6 Gate details and full fitting algorithm

Let $a = p_N(X)$ and $b = p_P(X)$. The eight gate descriptors are differences in normalized entropy, maximum probability, top-two margin, and squared-probability sum; total variation $\|a - b\|_1/2$; argmax agreement; and total variation multiplied by each of the entropy and maximum-probability differences. Descriptors are standardized using the meta-fitting subset, with population standard deviations floored at 0.1. The conditional gate is $g = \sigma(\gamma_0 + \gamma^\top z)$ and minimizes mean negative log-likelihood plus $0.001\gamma_0^2 + 0.05\|\gamma\|_2^2$. The constant gate has only the intercept and its penalty. L-BFGS uses at most 60 iterations with a strong-Wolfe line search. Expert calibration temperatures are selected from $\{0.25, 0.5, 1, 2, 4, 8\}$ on the same meta-fitting subset.

Base-model folds use seed 2718, distributing each class across distinct holdouts while retaining all classes in every fitting subset. The three meta-folds use shuffled K-fold with seed 1729. Every example receives one base out-of-fold prediction per expert. Meta-fitting uses those records rather than retraining all base models within each meta-fold. Accordingly, base fits contributing to different records can overlap; the procedure is an empirical model-selection device, not fully nested cross-validation. Equation 7 uses a numerical strictness threshold of -10^{-12} in the implementation.

Algorithm 1 MoSiLa fitting and evaluation under an explicitly declared checkpoint rule

Require: TRAIN sequences and labels $\mathcal{D}$; seed s; checkpoint rule r
1: Set pattern backend and dilation metadata from TRAIN size and length.
2: Construct five class-preserving TRAIN folds with seed 2718.
3: **for** fold $f = 0, \ldots, 4$ **do**
4: Fit normalization on $\mathcal{D} \setminus \mathcal{D}_f$.
5: Fit neural MoSiLa for 100 epochs and pattern ridge on that subset, seed $s + f$.
6: Record both experts' logits for all examples in $\mathcal{D}_f$.
7: **end for**
8: Form three meta-folds of the paired records with seed 1729.
9: Evaluate five candidate gates with meta-held-out predictions.
10: Apply the fixed sequential paired-error rule; refit the selected gate on all records.
11: Freeze the gate; fit normalization and the pattern expert on all TRAIN examples.
12: Train the full neural model using the declared rule r:
primary: TEST correct count; supplement: evaluation-mode TRAIN CE;
external: validation macro-F1 of the SWA-based predictor.
13: Restore the selected neural state and compute p_g for every TEST example.
14: Save predictions, correct counts, configuration, and selection metadata.

C Complete structural ablations

A01 removes the path projection and residual path-token map. A02 zeroes the first-moment half of the base token while retaining the other coordinates. A03 fixes the relation coefficients at zero. A04, A05, and A06 replace the global-token readout with, respectively, mean, elementwise maximum, and last-valid-token readout over the finest 64-bin partition. They do not pool over all 85 tokens. A07 retains the complete neural architecture but removes the pattern expert and mixture selection. All are seed-42, full-UEA30 runs under the recorded TEST-selected condition; A00 is the full model.

Table 8: Structural ablations on UEA30, seed 42, TEST-selected. The header records retained (✓) and removed (×) components and the readout. Bold/underline mark the best/second-best distinct accuracies (%), including ties. Footer: Drop is A00 minus the variant (pp); W/T/L counts A00 wins/ties/losses; exploratory two-sided signed-rank p values are unadjusted, with zeros omitted.

	Representation			Readout			Expert	Proposed
Dataset	A01	A02	A03	A04	A05	A06	A07	**A00**
Path	×	✓	✓	✓	✓	✓	✓	✓
Moment	✓	×	✓	✓	✓	✓	✓	✓
Relation	✓	✓	×	✓	✓	✓	✓	✓
Pattern	✓	✓	✓	✓	✓	✓	×	✓
Readout	Global	Global	Global	Mean	Max	Last	Global	Global
EC	**57.034**	55.133	55.133	55.133	55.133	55.133	34.221	56.274
FD	69.438	69.183	**70.829**	68.927	69.240	69.381	70.545	69.098
HW	57.176	**58.118**	**58.118**	57.176	56.588	55.294	55.765	52.588
HB	78.537	78.537	75.122	78.537	78.537	**79.024**	76.098	78.537
JV	**99.730**	**99.730**	99.459	99.459	98.919	99.459	98.919	99.459
PEMS	86.127	86.127	87.283	86.127	83.815	**89.595**	86.705	89.017
SRS1	91.126	85.666	85.666	85.666	85.666	85.666	**91.809**	91.468
SRS2	55.556	57.222	56.111	**57.778**	55.000	57.222	55.000	57.222
SAD	99.682	99.500	99.591	**99.909**	99.773	99.500	99.545	99.818
UW	**93.750**	91.562	93.125	91.875	92.812	91.875	92.812	93.125
AWR	98.000	**99.333**	**99.333**	99.000	98.333	**99.333**	98.333	99.000
AF	**40.000**	**40.000**	**40.000**	**40.000**	**40.000**	**40.000**	**40.000**	**40.000**
BM	**100.000**	**100.000**	**100.000**	**100.000**	**100.000**	97.500	**100.000**	**100.000**
CT	99.582	99.513	99.582	99.582	**99.721**	99.582	**99.721**	99.652
CR	**98.611**	**98.611**	**98.611**	**98.611**	**98.611**	**98.611**	**98.611**	**98.611**
DDG	**72.000**	68.000	68.000	68.000	68.000	66.000	68.000	66.000
EW	**97.710**	**97.710**	**97.710**	**97.710**	**97.710**	**97.710**	74.809	**97.710**
EP	**100.000**	**100.000**	**100.000**	**100.000**	**100.000**	**100.000**	97.101	**100.000**
ER	**98.148**	**98.148**	**98.148**	**98.148**	**98.148**	**98.148**	91.852	**98.148**
FM	59.000	58.000	59.000	59.000	59.000	58.000	59.000	**63.000**
HMD	51.351	**56.757**	50.000	51.351	50.000	50.000	50.000	54.054
IW	68.720	68.616	68.592	69.332	68.124	68.684	**69.752**	69.552
LIB	**96.111**	95.000	92.222	93.333	93.333	95.000	93.333	**96.111**
LSST	60.219	65.856	65.328	58.435	**66.018**	60.260	65.085	65.450
MI	55.000	57.000	59.000	**62.000**	60.000	50.000	60.000	61.000
NATO	97.778	98.333	96.667	97.778	97.222	**98.889**	97.222	96.111
PD	98.914	**99.285**	**99.285**	98.942	99.228	98.742	99.171	98.999
PS	34.208	33.940	34.954	30.212	**35.341**	33.373	30.510	32.747
RS	92.105	86.842	92.105	92.105	87.500	**92.763**	87.500	**92.763**
SWJ	46.667	46.667	46.667	46.667	46.667	46.667	**53.333**	46.667
Avg. acc. (30)	78.409	78.280	78.188	78.026	77.948	77.714	76.492	**78.739**
Drop (pp)	0.330	0.460	0.551	0.713	0.791	1.026	2.248	0.000
W/T/L	13/9/8	11/9/10	14/9/7	14/10/6	14/8/8	13/10/7	18/3/9	–
Paired p	0.369	0.452	0.160	0.083	0.075	0.083	0.020	–

Figure 4a shows all 210 dataset–variant differences. Each line follows one dataset across A01–A07; cycling colors separate lines, not dataset groups. Boxes show quartiles and medians; whiskers reach observations within 1.5 interquartile ranges. All points, including zeros, negative effects, and observations beyond the whiskers, are retained; coincident points can overlap. Figure 4b compares fixed-recipe mean effects with retuned aggregates (18–35 configurations per variant; Table 9), not seed replicates or uncertainty estimates.

C.1 DATASET-LEVEL COMPONENT EFFECTS

Figure 6 presents complete accuracy profiles for all 30 datasets. Each panel contains the full model and seven structural variants under the recorded seed-42 recipe. The full-model reference line makes both gains and losses visible. Panel order follows Table 3; route annotations describe the full-model fit and do not isolate a causal contribution of the gate.

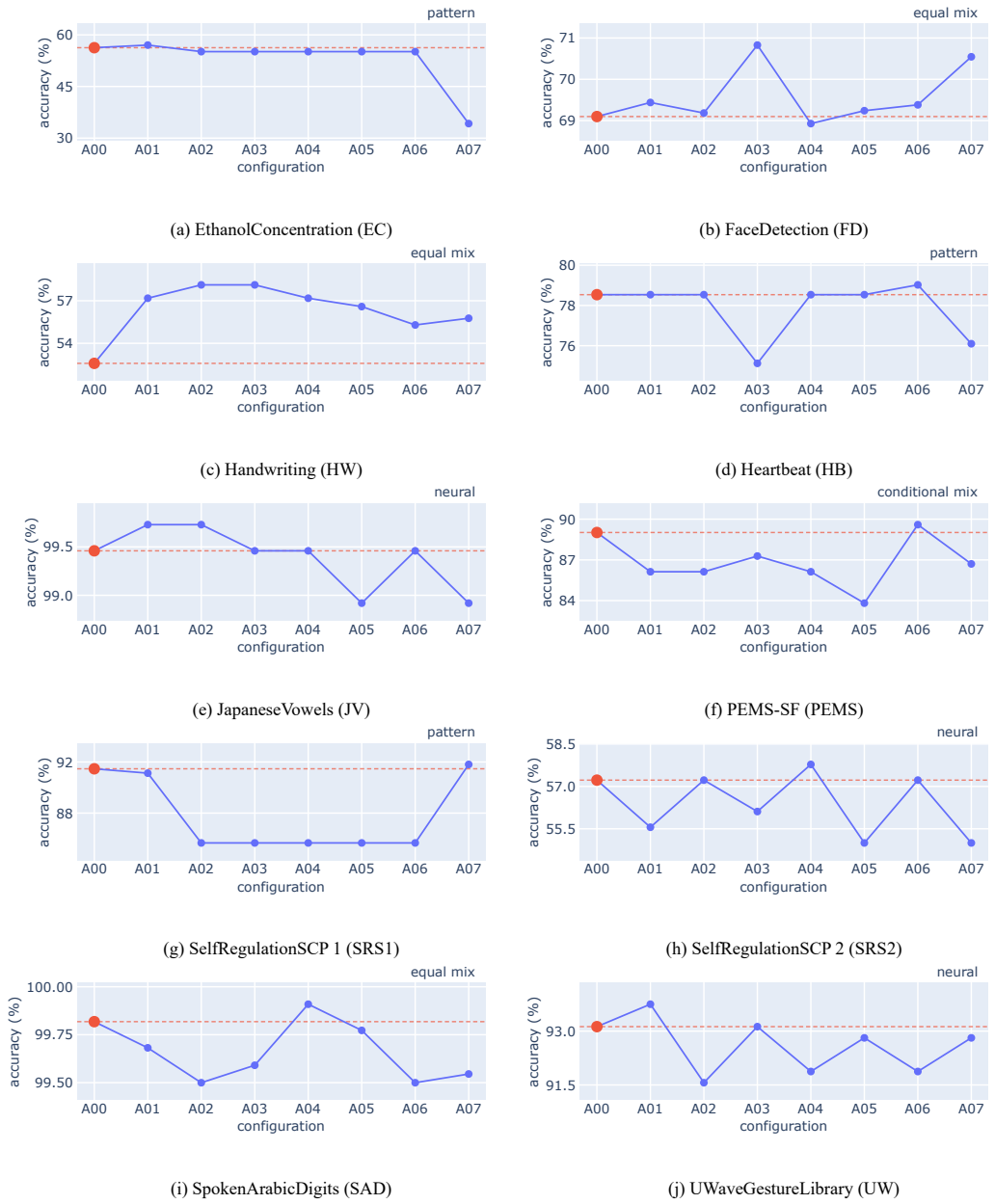
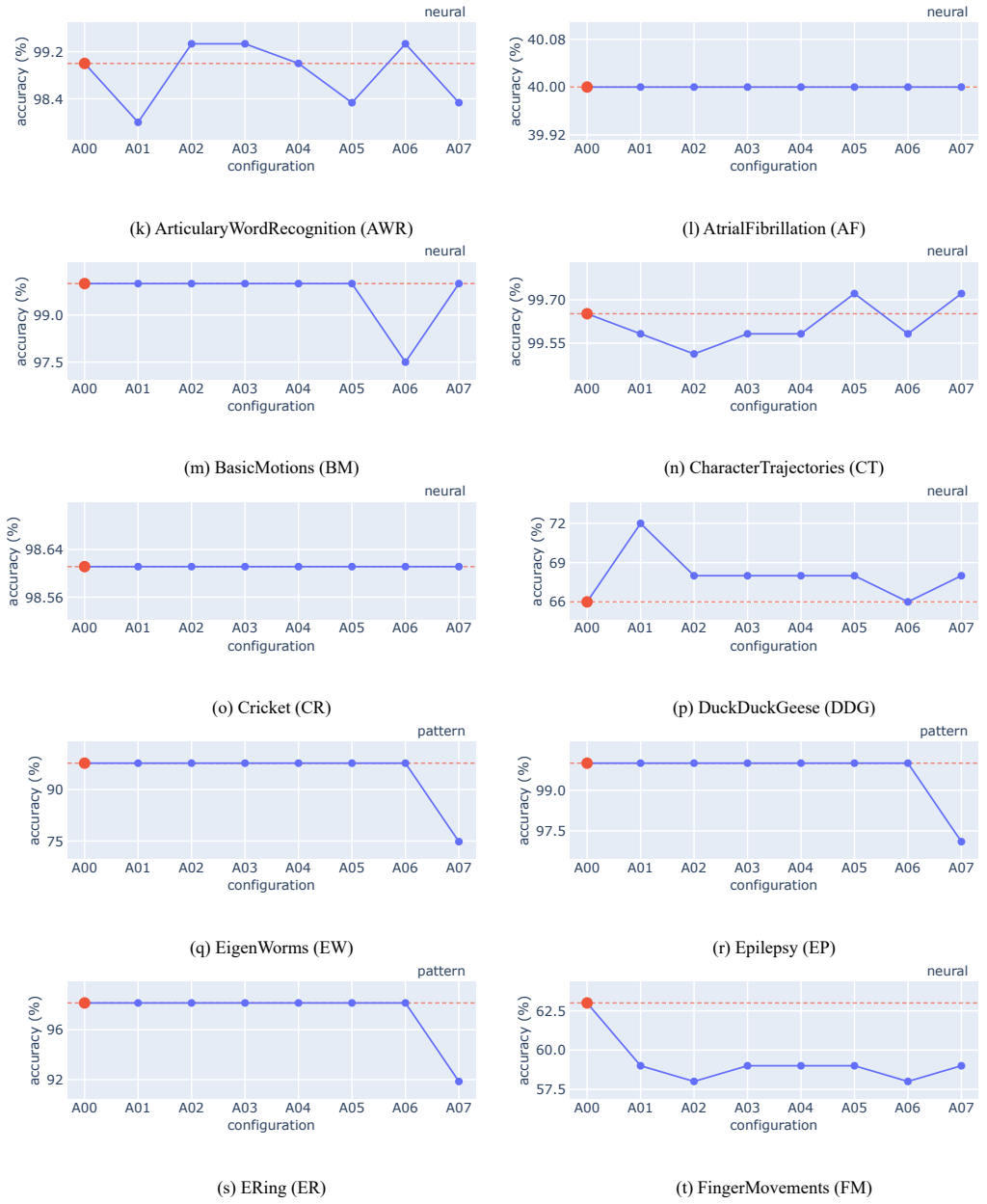


Figure 6: Classification accuracy of MoSiLa and its structural variants on UEA30, panels (a)–(j). Blue connects the eight configurations; the red point and dashed line mark complete A00. Each point is one measured seed-42 result. The annotation gives A00's selected prediction route. Vertical ranges vary by dataset; exact values appear in Table 8.



(k) ArticularyWordRecognition (AWR)

(l) AtrialFibrillation (AF)

(m) BasicMotions (BM)

(n) CharacterTrajectories (CT)

(o) Cricket (CR)

(p) DuckDuckGeese (DDG)

(q) EigenWorms (EW)

(r) Epilepsy (EP)

(s) ERing (ER)

(t) FingerMovements (FM)

Figure 6: (Continued.) Classification accuracy of MoSiLa and its structural variants, panels (k)–(t). Configuration order and colors match panels (a)–(j); every dataset and variant is retained.

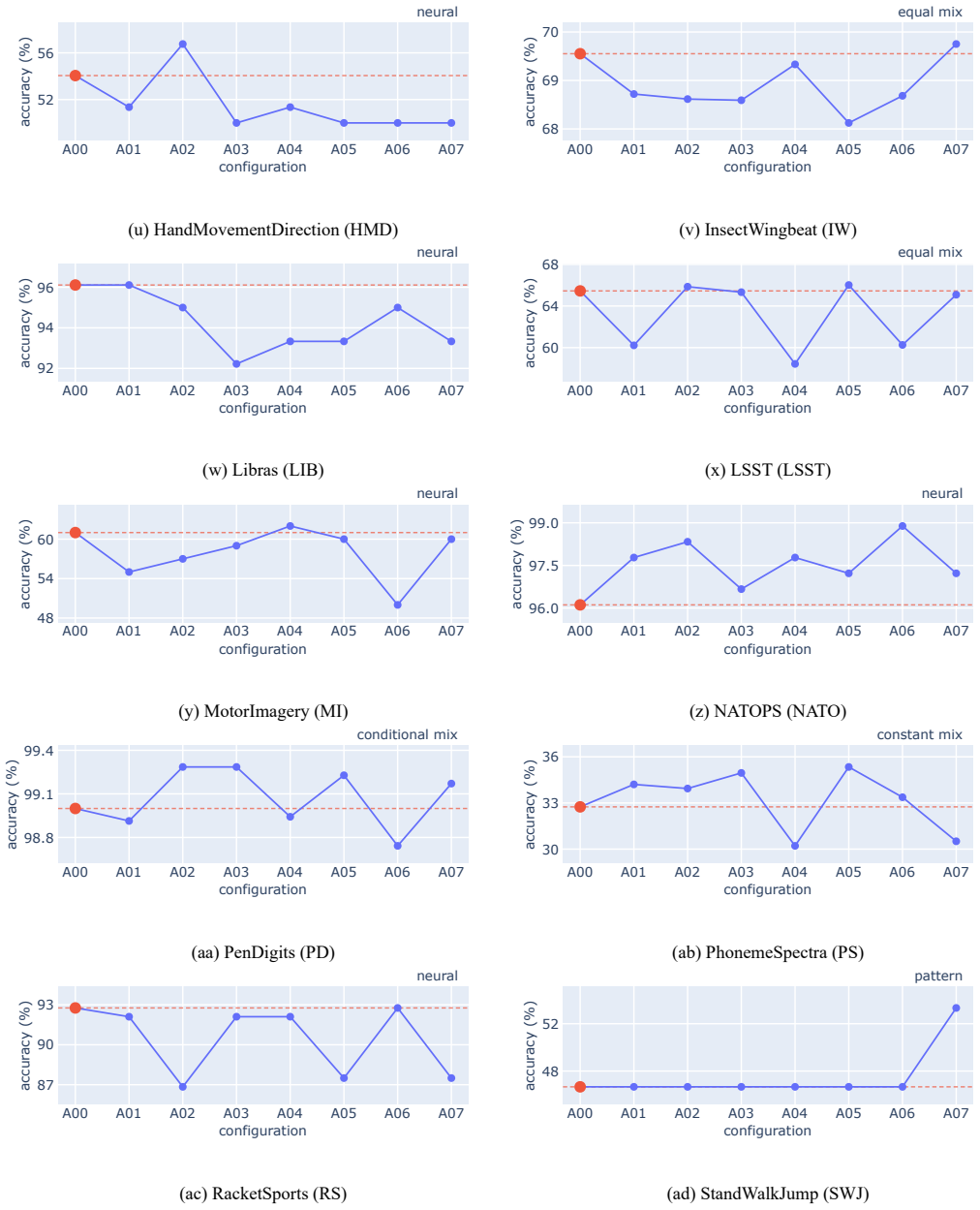


Figure 6: (Continued.) Classification accuracy of MoSiLa and its structural variants, panels (u)–(ad). A00's routes across all 30 datasets are 15 neural-only, seven pattern-only, and eight mixtures (five equal, one constant, two conditional).

C.2 Retuned structural variants

The fixed-recipe ablations above test component changes under the main fitting recipe. Table 9 separately reports the retuned outcomes, actual search counts, and accuracy differences from the full configuration. A00 reaches 79.152%, with all seven ablations numerically lower by 0.408–2.016 percentage points. Search counts differ across variants, so these results describe the recorded retuning outcomes rather than an equal-search-budget comparison.

Table 9: Retuned structural variants. Accuracy is the UEA30 macro-average in percent. Drop is the full A00 accuracy minus each variant, in percentage points. Search counts are reported individually.

Variant	Search count	Avg. acc. (30 datasets)	Drop (pp)
MoSiLa (A00 full)	26	**79.152**	0.000
A01: no path	33	78.744	0.408
A02: no moment	23	78.715	0.437
A03: no relation bias	18	78.535	0.617
A04: mean readout	25	78.247	0.905
A05: max readout	35	78.256	0.896
A06: last readout	28	77.934	1.218
A07: no pattern expert	33	77.136	2.016

D DEPTH COMPARISON

Only relation-block depth varies; the three convolutional units, feature width, token construction, pattern branch, gate fitting, seed, precision, and checkpoint rule remain fixed. The two-block column reuses the primary MoSiLa run. It has the highest macro-average (78.739%) and best mean rank (1.82). Table 10 gives all results; W/T/L counts two-block wins/ties/losses. Both unadjusted paired p values exceed 0.05, supporting a numerical ordering without establishing a significant depth effect.

Table 10: UEA30 relation-block depth comparison, TEST-selected, seed 42. Bold and underline mark the best and second-best distinct scores among these three configurations, including ties.

	Model depth		
Dataset	1	**2 (proposed)**	3
EC	**57.034**	56.274	**57.034**
FD	69.438	69.098	**69.637**
HW	55.176	52.588	**57.412**
HB	**78.537**	**78.537**	**78.537**
JV	99.730	99.459	**100.000**
PEMS	80.347	**89.017**	86.705
SRS1	85.666	**91.468**	85.666
SRS2	55.556	**57.222**	55.556
SAD	99.773	**99.818**	99.773
UW	91.250	**93.125**	91.250
AWR	98.333	99.000	**99.333**
AF	33.333	**40.000**	33.333
BM	**100.000**	**100.000**	**100.000**
CT	**99.721**	99.652	**99.721**
CR	**98.611**	**98.611**	**98.611**
DDG	**68.000**	66.000	58.000
EW	**97.710**	**97.710**	**97.710**
EP	**100.000**	**100.000**	**100.000**
ER	**98.148**	**98.148**	**98.148**
FM	59.000	**63.000**	58.000
HMD	52.703	**54.054**	52.703
IW	68.308	**69.552**	68.132
LIB	**96.111**	**96.111**	95.000
LSST	64.436	**65.450**	64.477
MI	58.000	**61.000**	59.000
NATO	**97.778**	96.111	**97.778**
PD	98.942	98.999	**99.228**
PS	**35.073**	32.747	34.506
RS	92.105	**92.763**	92.105
SWJ	**46.667**	**46.667**	**46.667**
Avg. acc. (30)	77.850	**78.739**	77.801
Avg. rank	2.15	**1.82**	2.03
W/T/L	14/8/8	–	14/7/9
Paired p	0.140	–	0.057

E MEASURED RESULTS VERSUS LITERATURE REPORTS

Table 11 separates measured and published classification results for all 18 baselines. The references for Crossformer, DLinear, FEDformer, HYDRA, MambaSL, ModernTCN, MR+Hydra, PatchTST, TSCMamba, TSLANet, TimeMixer++, TimesNet, and iTransformer come from Table 5 of Jung & Kim (2026). That table is itself a re-evaluation, not necessarily each method's original-paper score. ConvTran, MultiRocket, Non-stationary Transformer, TimeFilter, and TimeMixer use the additional reference summaries supplied with the results. Published numbers retain their source precision and are not included in the primary measured ranking.

Our MambaSL remeasurement is 73.587%, whereas its published Table-5 result is 79.819%. The difference cannot be attributed to one cause from these records alone: the original search, random seed, floating-point execution, and runtime conditions differ. Matching recovered model settings is not a guarantee of reproducing a paper's selected score. The same caveat applies to the other baselines, and it limits claims that MoSiLa improves on their best published performance.

Reference coverage records the dataset count accompanying each summary, while the measured column uses the full UEA30 average. Differences are reported only for references covering all 30 datasets. Non-stationary Transformer and TimeFilter retain their reported means and reference counts of 29 and 28, respectively, alongside the full-UEA30 measurements; no cross-coverage difference is computed for these two rows.

Table 11: Measured versus reported accuracy. Ref. coverage is the reference dataset count; the measured column covers UEA30. Difference is this study minus the report, in percentage points, shown only for full-UEA30 references. A dash denotes differing coverage. These are not paired seed-identical comparisons. Source groups are specified above.

	Reference		This study	
Baseline	Datasets	Accuracy (%)	Accuracy (%)	Difference (pp)
TSLANet	30	76.263	74.949	-1.314
ConvTran	30	79.207	74.208	-5.000
MR+Hydra	30	76.320	73.938	-2.382
MultiRocket	30	76.913	73.766	-3.148
MambaSL	30	79.819	73.587	-6.232
TSCMamba	30	78.405	72.834	-5.571
ModernTCN	30	76.938	71.663	-5.275
Crossformer	30	77.427	71.475	-5.952
Non-stationary Tr.	29	75.346	71.457	–
Hydra	30	77.124	71.368	-5.756
TimesNet	30	76.921	71.073	-5.848
TimeMixer	30	72.835	70.839	-1.996
TimeMixer++	30	77.279	70.547	-6.732
PatchTST	30	74.729	69.494	-5.235
iTransformer	30	74.797	69.491	-5.307
FEDformer	30	74.128	68.851	-5.278
TimeFilter	28	69.633	66.933	–
DLinear	30	68.247	66.342	-1.905

Table 12: Measured (M) versus reported (R) accuracy (%) on UEA30. R uses MambaSL Table 5; M and R are distinct evaluation records. Each panel ends with the macro-average across all 30 datasets.

Dataset	Crossformer		DLinear		FEDformer		HYDRA		MambaSL		ModernTCN		MR+Hydra	
	M	R	M	R	M	R	M	R	M	R	M	R	M	R
EC	36.502	43.346	27.757	31.179	27.757	33.840	52.091	54.373	31.939	42.586	31.559	32.319	52.091	60.076
FD	68.275	69.637	67.963	69.495	69.268	69.892	52.866	60.982	66.941	69.296	65.579	66.998	59.024	61.294
HW	29.059	30.353	23.176	23.765	30.353	33.412	47.529	56.353	55.176	60.824	26.353	30.118	52.118	53.529
HB	78.049	79.024	76.098	76.585	74.146	78.049	78.049	76.585	78.537	80.488	75.122	78.049	77.561	77.073
JV	97.297	98.378	96.486	96.486	95.405	97.568	98.378	97.838	96.486	98.649	97.838	98.649	98.919	98.378
PEMS	79.191	89.595	80.347	82.659	87.861	89.017	80.347	98.266	83.237	85.549	82.659	86.705	80.347	82.659
SRS1	91.468	93.857	93.515	92.833	69.625	75.085	85.324	86.689	92.150	92.491	92.833	93.515	89.078	94.881
SRS2	55.556	61.667	53.333	57.222	55.556	59.444	52.778	58.889	55.556	65.000	51.667	60.000	55.000	55.000
SAD	98.317	98.863	96.407	96.953	98.545	99.454	98.863	99.136	99.545	99.955	98.272	99.500	99.409	99.045
UW	84.375	88.438	81.875	82.813	78.438	79.688	90.000	94.375	90.938	93.438	87.188	88.438	94.062	94.063
AWR	98.667	98.667	97.333	97.667	90.333	92.333	99.333	99.333	97.667	99.333	98.333	98.667	99.333	99.333
AF	33.333	73.333	40.000	53.333	40.000	73.333	6.667	26.667	33.333	53.333	26.667	73.333	6.667	6.667
BM	100.000	100.000	85.000	87.500	92.500	95.000	100.000	100.000	82.500	100.000	97.500	100.000	100.000	100.000
CT	98.398	98.955	97.772	98.120	68.593	97.284	98.816	99.164	99.304	99.721	98.538	99.304	99.304	99.373
CR	95.833	97.222	91.667	91.667	83.333	90.278	98.611	100.000	94.444	100.000	97.222	97.222	98.611	98.611
DDG	56.000	72.000	66.000	66.000	30.000	42.000	56.000	68.000	60.000	70.000	50.000	66.000	46.000	60.000
EW	54.198	58.015	42.748	42.748	61.069	63.359	64.885	98.473	78.626	83.969	56.489	64.122	83.969	98.473
EP	90.580	95.652	46.377	65.217	94.928	92.754	100.000	100.000	97.101	97.826	95.652	96.377	100.000	100.000
ER	94.815	97.407	90.000	91.852	95.926	95.926	98.519	98.889	91.852	93.704	92.222	97.407	99.630	99.259
FM	56.000	64.000	60.000	62.000	61.000	72.000	56.000	55.000	60.000	71.000	66.000	67.000	61.000	60.000
HMD	59.459	66.216	63.514	66.216	71.622	71.622	22.973	51.351	67.568	70.270	59.459	66.216	36.486	36.486
IW	71.948	73.712	19.188	18.996	65.100	65.244	59.472	67.052	65.392	66.304	75.200	75.588	61.928	65.744
LIB	86.667	90.556	73.333	70.556	90.000	93.333	92.778	93.333	86.111	91.667	88.333	90.000	92.778	93.333
LSST	37.429	41.768	33.090	34.023	36.861	40.998	54.866	66.504	41.281	45.580	40.998	43.593	47.324	64.274
MI	62.000	68.000	59.000	64.000	58.000	65.000	50.000	54.000	55.000	69.000	64.000	67.000	49.000	51.000
NATO	88.889	93.889	95.556	96.111	87.778	96.111	86.111	90.000	95.000	98.889	94.444	95.556	93.333	93.333
PD	98.285	98.513	91.995	92.481	98.914	98.599	96.398	97.856	98.628	99.257	97.885	98.513	96.884	97.284
PS	5.458	14.256	6.472	6.651	11.035	11.691	33.254	33.164	29.824	30.331	13.630	13.749	32.806	33.642
RS	84.868	87.500	80.921	78.947	81.579	84.868	90.132	91.447	90.132	92.763	81.579	84.211	88.816	90.132
SWJ	53.333	80.000	53.333	53.333	60.000	66.667	40.000	40.000	33.333	73.333	46.667	80.000	66.667	66.667
Avg. acc.	71.475	77.427	66.342	68.247	68.851	74.128	71.368	77.124	73.587	79.819	71.663	76.938	73.938	76.320

Dataset	PatchTST		TSCMamba		TSLANet		TimeMixer++		TimesNet		iTransformer	
	M	R	M	R	M	R	M	R	M	R	M	R
EC	27.757	32.700	26.236	30.798	28.897	31.939	31.939	34.601	28.897	33.460	27.757	31.939
FD	68.473	68.076	67.764	70.204	62.940	66.969	67.310	69.665	68.530	70.148	66.175	68.417
HW	29.765	29.529	41.176	45.059	55.529	62.000	32.588	33.647	35.529	37.529	30.000	31.294
HB	69.756	73.659	77.073	78.537	80.976	79.512	78.537	80.488	75.610	83.902	77.073	78.049
JV	94.865	97.027	97.568	98.378	98.919	98.919	97.568	99.189	97.297	98.649	97.297	98.108
PEMS	83.815	89.017	84.971	90.173	82.081	89.017	86.127	89.017	82.081	87.861	86.127	91.329
SRS1	84.983	89.078	92.150	92.833	87.713	89.761	90.102	92.150	89.078	92.150	91.809	93.857
SRS2	53.333	61.667	54.444	62.222	60.556	63.889	53.333	60.556	57.778	60.000	55.000	58.889
SAD	98.090	98.818	70.987	95.816	99.955	98.045	98.954	99.545	99.136	99.545	98.454	99.318
UW	85.625	89.375	95.000	95.313	93.125	94.375	91.562	90.625	88.125	88.438	87.188	90.000
AWR	98.000	98.667	98.000	99.333	99.000	99.333	98.667	99.000	98.000	98.667	98.333	99.000
AF	60.000	80.000	46.667	66.667	46.667	66.667	26.667	66.667	33.333	66.667	33.333	66.667
BM	77.500	85.000	100.000	100.000	100.000	100.000	95.000	100.000	95.000	100.000	92.500	97.500
CT	98.259	98.398	99.164	99.513	99.373	99.164	98.816	99.373	97.632	98.747	97.841	98.886
CR	91.667	97.222	98.611	100.000	98.611	91.667	93.056	98.611	91.667	97.222	93.056	97.222
DDG	26.000	40.000	26.000	50.000	68.000	74.000	54.000	74.000	62.000	68.000	56.000	60.000
EW	51.908	57.252	80.916	88.550	75.573	81.679	58.779	61.069	60.305	63.359	54.962	59.542
EP	93.478	98.551	97.101	99.275	97.826	100.000	83.333	94.928	89.130	96.377	78.986	80.435
ER	94.074	96.667	92.593	97.037	86.667	85.926	93.704	96.296	94.074	95.556	93.704	94.815
FM	57.000	63.000	57.000	66.000	59.000	68.000	60.000	67.000	59.000	66.000	56.000	64.000
HMD	59.459	63.514	52.703	60.811	62.162	48.649	66.216	72.973	59.459	72.973	48.649	58.108
IW	59.768	60.936	71.024	68.244	55.860	10.012	63.864	64.192	63.512	64.024	73.124	73.436
LIB	78.333	81.667	86.667	91.667	96.111	95.556	84.444	88.333	85.000	88.333	89.444	91.667
LSST	55.637	56.732	58.110	60.057	38.281	51.014	29.521	44.444	40.065	44.120	37.794	40.673
MI	61.000	69.000	59.000	67.000	63.000	71.000	60.000	67.000	59.000	68.000	59.000	67.000
NATO	79.444	80.556	85.556	92.222	96.111	99.444	91.667	96.111	94.444	98.333	86.111	88.889
PD	97.570	98.485	96.827	98.513	98.628	99.028	97.970	98.599	98.599	98.914	98.199	98.428
PS	15.001	15.091	22.905	23.800	24.068	4.921	15.151	15.508	14.345	14.524	10.498	11.542
RS	80.921	85.526	88.816	90.789	86.184	94.079	84.211	91.447	82.237	89.474	76.974	81.579
SWJ	53.333	86.667	60.000	73.333	46.667	73.333	33.333	73.333	33.333	66.667	33.333	73.333
Avg. acc.	69.494	74.729	72.834	78.405	74.949	76.263	70.547	77.279	71.073	76.921	69.491	74.797

F TRAIN-LOSS SELECTION AND REPEATED RUNS

F.1 COMPARISON UNDER TRAIN-BASED SELECTION

Table 13 separates seed-42 point comparisons from three-seed summaries. Classical methods use TRAIN-based ridge selection, while MoSiLa uses the neural TRAIN-loss checkpoint rule. All four methods are summarized over seeds 42–44. W/T/L and paired p compare MoSiLa with the listed baseline using their respective seed-42 results on the same 30 datasets. The three-seed SD is calculated across seed-specific macro-averages, not across pooled dataset–seed observations.

Table 13: TRAIN-based-selection comparison. Accuracy is in percent; all three-seed summaries use seeds 42–44 and sample SD. W/T/L counts MoSiLa wins, ties, and losses at seed 42, with unadjusted exploratory paired p. Existing baseline preprocessing differences in Appendix B remain applicable.

Model	Seed-42 accuracy	Three-seed mean ± SD	W/T/L	Paired p
MoSiLa	**76.469**	**76.570** ±0.439	n/a	n/a
MultiRocket	73.766	73.557 ±0.445	18/3/9	0.020
Hydra	71.368	71.436 ±0.119	23/4/3	4.08×10^{-6}
MR+Hydra	<u>73.938</u>	<u>73.755</u> ±0.287	18/5/7	0.018

Table 14: Classical-baseline TEST accuracy for all 270 dataset–model–seed records. Model names group the columns; subheaders identify seeds 42, 43, and 44. The final row is the UEA30 macro-average for each run.

	MultiRocket			HYDRA			MR+Hydra		
Dataset	42	43	44	42	43	44	42	43	44
EC	53.232	50.570	52.471	52.091	54.753	52.091	52.091	48.289	50.570
FD	60.726	59.648	61.010	52.866	52.667	53.746	59.024	58.428	58.002
HW	52.000	50.941	52.588	47.529	47.765	47.412	52.118	51.882	53.176
HB	76.098	77.073	76.585	78.049	77.561	78.049	77.561	79.024	75.610
JV	98.919	98.378	98.919	98.378	98.919	98.649	98.919	99.189	99.189
PEMS	79.769	78.035	81.503	80.347	81.503	82.081	80.347	79.191	80.925
SRS1	91.468	91.468	93.174	85.324	84.300	84.300	89.078	89.078	89.078
SRS2	55.556	54.444	54.444	52.778	52.778	54.444	55.000	53.333	53.889
SAD	99.454	98.863	99.682	98.863	99.045	98.545	99.409	99.090	99.500
UW	94.375	92.188	93.438	90.000	89.062	87.812	94.062	91.250	93.750
AWR	99.333	98.667	100.000	99.333	100.000	99.333	99.333	99.000	99.667
AF	6.667	13.333	20.000	6.667	13.333	6.667	6.667	20.000	0.000
BM	100.000	100.000	97.500	100.000	95.000	100.000	100.000	100.000	100.000
CT	99.304	98.747	99.721	98.816	98.747	98.677	99.304	99.164	99.861
CR	97.222	94.444	98.611	98.611	97.222	97.222	98.611	97.222	97.222
DDG	50.000	46.000	42.000	56.000	56.000	50.000	46.000	42.000	54.000
EW	77.863	77.099	77.863	64.885	63.359	63.359	83.969	84.733	85.496
EP	100.000	99.275	100.000	100.000	100.000	98.551	100.000	100.000	100.000
ER	98.519	97.778	98.519	98.519	97.407	97.407	99.630	99.630	100.000
FM	66.000	59.000	63.000	56.000	52.000	57.000	61.000	65.000	63.000
HMD	35.135	39.189	29.730	22.973	21.622	28.378	36.486	33.784	36.486
IW	59.548	58.812	59.828	59.472	59.260	58.864	61.928	61.848	61.944
LIB	92.778	91.667	92.222	92.778	92.222	93.333	92.778	91.667	93.889
LSST	46.715	45.661	46.796	54.866	54.461	55.718	47.324	48.013	47.689
MI	51.000	51.000	46.000	50.000	50.000	55.000	49.000	45.000	46.000
NATO	93.889	92.778	96.111	86.111	86.111	86.667	93.333	95.000	91.667
PD	96.827	95.740	97.284	96.398	96.913	95.912	96.884	96.770	97.227
PS	32.419	31.792	33.284	33.254	32.210	33.164	32.806	32.985	33.075
RS	88.158	88.816	86.842	90.132	90.132	88.158	88.816	88.816	89.474
SWJ	60.000	60.000	66.667	40.000	46.667	46.667	66.667	53.333	66.667
Avg. acc. (30)	73.766	73.047	73.860	71.368	71.367	71.574	73.938	73.424	73.902

F.2 MoSiLa and max-readout repeated runs

No TRAIN validation split is introduced in this supplement. The checkpoint objective is neural cross-entropy evaluated on all TRAIN examples with the current fixed checkpoint in evaluation mode. It is not the minibatch training-loss average accumulated while weights change, nor is it the combined predictor's cross-entropy. Equal loss can update the checkpoint; the maximum epoch count and patience follow the historical MoSiLa recipe. The fixed TRAIN-fitted gate is used for the subsequent TEST predictions. Table 15 reports three actual runs; standard deviations use the sample denominator $n - 1$.

The macro-average standard deviation is computed by first averaging each seed over all 30 datasets and then taking the standard deviation of those three averages. It is not the mean of the 30 dataset-level standard deviations. Earlier exploratory access to these benchmark TEST labels remains part of the development history even though these particular checkpoints are not selected on TEST. The exploratory A05 max-readout follow-up uses the same seeds. Its macro-averages are 75.276%, 75.234%, and 75.229%, with mean 75.247% and sample SD 0.026 percentage points. Separate column groups retain the distinction between complete MoSiLa/A00 and this follow-up.

Table 15: TRAIN-loss-selected TEST accuracy (%) for seeds 42, 43, and 44: complete MoSiLa/A00 and the exploratory A05 max-readout follow-up. Mean and sample SD are per dataset; the final row summarizes seed-specific UEA30 macro-averages.

	MoSiLa / A00				A05: max readout (follow-up)			
Dataset	42	43	44	Mean ± SD	42	43	44	Mean ± SD
EC	55.133	53.232	51.711	53.359 ±1.715	55.133	52.471	51.711	53.105 ±1.797
FD	67.849	66.544	65.607	66.667 ±1.126	66.515	67.423	67.650	67.196 ±0.601
HW	49.412	53.765	52.235	51.804 ±2.208	56.118	58.824	57.294	57.412 ±1.357
HB	78.537	75.610	76.585	76.911 ±1.490	78.537	75.610	76.585	76.911 ±1.490
JV	98.919	98.649	99.459	99.009 ±0.413	98.378	98.649	98.919	98.649 ±0.270
PEMS	83.815	82.659	87.283	84.586 ±2.407	84.393	79.191	82.659	82.081 ±2.649
SRS1	90.785	92.150	92.833	91.923 ±1.043	85.666	91.468	89.420	88.851 ±2.943
SRS2	56.667	52.778	50.000	53.148 ±3.349	55.000	53.889	50.000	52.963 ±2.625
SAD	99.409	99.363	99.363	99.379 ±0.026	99.500	99.591	99.773	99.621 ±0.139
UW	92.500	92.812	89.688	91.667 ±1.721	90.625	91.562	91.562	91.250 ±0.541
AWR	99.000	99.000	98.333	98.778 ±0.385	98.000	98.667	97.667	98.111 ±0.509
AF	13.333	26.667	26.667	22.222 ±7.698	13.333	20.000	26.667	20.000 ±6.667
BM	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
CT	99.513	99.652	99.582	99.582 ±0.070	99.234	99.373	99.373	99.327 ±0.080
CR	98.611	98.611	98.611	98.611 ±0.000	98.611	98.611	100.000	99.074 ±0.802
DDG	58.000	56.000	72.000	62.000 ±8.718	52.000	48.000	46.000	48.667 ±3.055
EW	97.710	98.473	98.473	98.219 ±0.441	97.710	98.473	98.473	98.219 ±0.441
EP	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	97.101	99.034 ±1.673
ER	98.148	98.889	98.889	98.642 ±0.428	98.148	98.889	98.889	98.642 ±0.428
FM	57.000	49.000	58.000	54.667 ±4.933	45.000	51.000	54.000	50.000 ±4.583
HMD	52.703	48.649	52.703	51.351 ±2.341	43.243	44.595	50.000	45.946 ±3.575
IW	67.568	68.508	68.568	68.215 ±0.561	67.932	67.628	67.980	67.847 ±0.191
LIB	95.000	95.000	91.667	93.889 ±1.925	93.333	95.000	92.778	93.704 ±1.156
LSST	66.748	66.545	64.517	65.937 ±1.233	65.247	64.071	62.571	63.963 ±1.341
MI	53.000	50.000	55.000	52.667 ±2.517	55.000	44.000	50.000	49.667 ±5.508
NATO	98.333	96.111	97.222	97.222 ±1.111	96.667	93.333	95.000	95.000 ±1.667
PD	98.771	98.714	98.714	98.733 ±0.033	98.628	97.913	98.571	98.370 ±0.397
PS	30.808	32.210	31.673	31.564 ±0.707	33.492	32.001	32.687	32.727 ±0.746
RS	90.132	89.474	89.474	89.693 ±0.380	86.184	90.132	83.553	86.623 ±3.311
SWJ	46.667	46.667	46.667	46.667 ±0.000	46.667	46.667	40.000	44.444 ±3.849
Avg. acc. (30)	76.469	76.191	77.051	76.570 ±0.439	75.276	75.234	75.229	75.247 ±0.026

G EXTERNAL DATASETS

ADFTD contains scalp EEG recordings from Alzheimer's disease, frontotemporal dementia, and healthy groups (Miltiadous et al., 2023). We use the existing preprocessed 256-timestep, 19-channel windows and three classes. The author loader splits subject IDs within each class by 60/20/20 boundaries before loading their windows; this yields 40,446 TRAIN and 14,648 TEST windows in the records used here. ADFTD inputs are also normalized within each window by the author loader. Our fitting pipeline subsequently applies its training-fitted channel normalization. The subject-defined external split does not imply that the gate's internal sample-level folds are subject-independent.

FLAAP is a human-activity dataset (Kumar & Suresh, 2022). The preprocessed input used here has 100 timesteps, six channels, and ten class labels. It contains 13,123 windows, split by the author loader into 7,873 TRAIN, 2,625 validation, and 2,625 TEST examples with fixed split random state 42. This is a window-level random split, not an unseen-subject protocol. The external loader's splits remain fixed across the five model seeds; repeated runs measure training randomness, not variation across five independently sampled partitions. Channel/class counts in the reference's dataset summary are not copied where they conflict with the actual data interface.

For both datasets the five base-model folds retain 100-epoch RAdam fits at learning rate 0.001 and batch 16. Full refitting uses Adam at 0.001, batch 32, at most 100 epochs, no cosine scheduler, and SWA-style parameter averaging updated every epoch. The existing implementation uses the averaged model's current batch-normalization buffers; it does not perform an additional final batch-normalization recalibration. Validation macro-F1 selects the combined predictor with improvement threshold 10^{-5} and patience ten. TEST is evaluated after selection. Five seeds 42–46 are reported without choosing the best seed.

Table 16: All ten actual external runs. Accuracy and macro-F1 are percentages.

			TEST performance (%)	
Dataset	Seed	Correct / total	Accuracy	Macro-F1
ADFTD	42	8446 / 14648	57.660	55.487
ADFTD	43	7758 / 14648	52.963	51.727
ADFTD	44	7536 / 14648	51.447	48.867
ADFTD	45	6585 / 14648	44.955	41.464
ADFTD	46	7374 / 14648	50.341	45.690
FLAAP	42	1838 / 2625	70.019	69.667
FLAAP	43	1799 / 2625	68.533	68.587
FLAAP	44	1825 / 2625	69.524	69.387
FLAAP	45	1822 / 2625	69.410	69.375
FLAAP	46	1805 / 2625	68.762	68.290

G.1 LITERATURE CONTEXT

Table 17: Literature context from MambaSL Table 12. Medformer and MambaSL retain their reported means and SDs in percent, separately from the measured systems in Table 2. Bold and underline mark the higher and lower mean within these two reports.

	ADFTD		FLAAP	
Reported method	Accuracy	Macro-F1	Accuracy	Macro-F1
Medformer	**53.27** ±1.54	**50.65** ±1.51	<u>76.44</u> ±0.64	<u>76.25</u> ±0.65
MambaSL	<u>51.68</u> ±0.89	<u>48.73</u> ±3.05	**77.47** ±1.03	**77.59** ±1.15

MoSiLa has numerically higher mean accuracy and macro-F1 than the additional measured comparators wherever their metrics are available. However, searches and seeds differ from the literature reference, which searched configurations before repeated external runs. The unpaired reports do not establish matched-compute superiority, and the FLAAP gap to Medformer and MambaSL remains a negative result.