ERA-U: A CONVENTION-ADAPTIVE EVIDENCE INTERFACE FOR WINDOW-BASED TIME SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Window-based time-series classification (TSC) often faces the multi-class window problem: one observation window contains several states but receives one convention-dependent label. We introduce ERA-U, an explicit, no-bypass factorization of pre-aggregation temporal class scores and a low-capacity positional aggregator. ERA-U does not enlarge the hypothesis class of positional weighted pooling followed by linear classification; instead, it makes the intermediate class-score sequence directly available for diagnosis and defines a restricted interface for supervised convention adaptation while keeping the temporal representation fixed. Across five real-world datasets, ERA-U improves mixed-state-window recognition, reveals that many errors select competing states present in the input, and leads conflict-window Macro-F1 in five of six directed convention transfers while updating only nine aggregation parameters or a 525-parameter lightweight interface. These results establish convention shift as a distinct TSC problem and show that an explicit, restricted prediction interface can support accurate diagnosis and low-capacity adaptation. **Code:** `https://anonymous.4open.science/r/anonymous-code-F1EF/`

1 INTRODUCTION

Window-based time-series classification (TSC) partitions a continuous signal into fixed-length windows and assigns one label to each. It underlies wearable activity recognition, industrial monitoring, and sleep staging (Ordóñez and Roggen, 2016). Because state transitions need not align with window boundaries, one window may contain several classes but receive only one label, creating the multi-class window problem.

Conventional window-level supervision has two drawbacks: one label obscures competing class-specific evidence in mixed-state windows (MSWs), and training under one labeling convention couples that evidence to the rule producing the window label. Figure 1 contrasts single-state windows (SSWs) and MSWs and shows that most evaluated models perform worse on MSWs.

This coupling is easy to overlook because different rules agree on most SSWs. Their difference is concentrated on MSWs, where Last, Center, and Majority can assign different labels to an unchanged signal. A model should therefore distinguish what states the signal supports from which supported state the convention selects.

We formulate convention shift as changing the rule that maps an unchanged multi-state window to one target label. ERA-U implements this view with an explicit, restricted factorization between pre-aggregation temporal class scores and convention-dependent positional aggregation. This is an operational interface rather than a claim of mathematical identifiability or a larger hypothesis class. We evaluate it on five datasets, multiple window scales, and the Last, Center, and Majority rules. Our contributions are:

1. We systematically study the multi-class window problem and show that many MSW errors correspond to competing states present within the input.
2. We define convention shift on unchanged windows and introduce a no-bypass factorization that exposes the pre-aggregation class-score sequence used by the final prediction.

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.

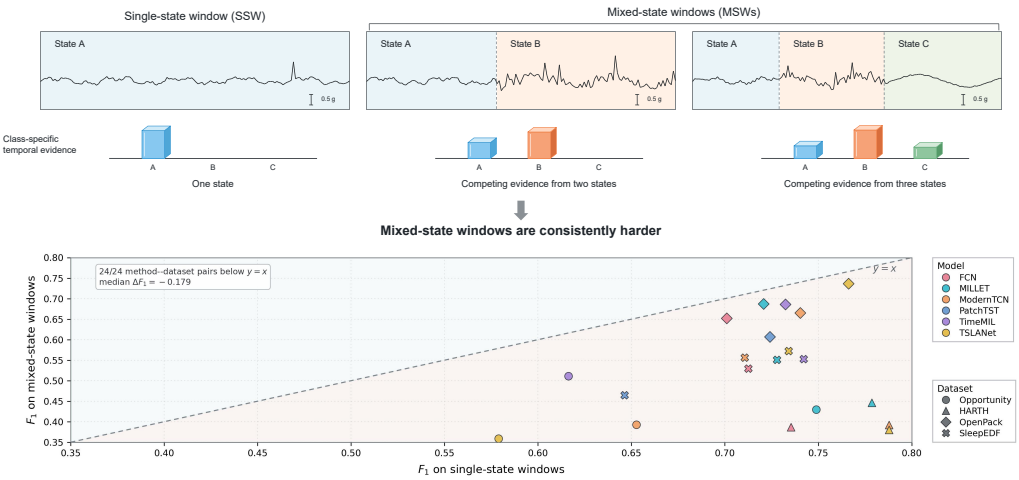


Figure 1: The multi-class window problem. **Top:** SSWs contain one state, whereas MSWs contain competing evidence. **Bottom:** MSW versus SSW Macro-F1; points below $y = x$ perform worse on MSWs. Colors denote methods and markers denote datasets.

3. We introduce a nine-parameter U/F/P aggregator and a 525-parameter lightweight interface for supervised convention adaptation with a frozen temporal representation.
4. We evaluate ERA-U on five datasets and all six directed transfers among Last, Center, and Majority.

2 Background and Related Work

Time-series classification. TSC spans convolutional and recurrent networks (Wang et al., 2017; Bai et al., 2018; Karim et al., 2018; Fawaz et al., 2020), random-kernel classifiers (Dempster et al., 2020; 2021; 2023), and general temporal architectures (Wu et al., 2023; Nie et al., 2023; Luo and Wang, 2024; Eldele et al., 2024). Transformer-based models further support long-range sequence modeling (Vaswani et al., 2017; Zerveas et al., 2021). Large benchmark studies compare classical, ensemble, and deep classifiers across diverse archives (Bagnall et al., 2017; Lines et al., 2018; Dau et al., 2019; Ruiz et al., 2021; Middlehurst et al., 2024; Fawaz et al., 2019).

Representation learning and temporal aggregation. Unsupervised and contrastive objectives learn transferable temporal representations without dense labels (van den Oord et al., 2018; Franceschi et al., 2019; Tonekaboni et al., 2021; Eldele et al., 2021; Yue et al., 2022; Zhang et al., 2022; Woo et al., 2022). Multiple-instance learning formalizes prediction from weak bag labels (Dietterich et al., 1997; Maron and Lozano-Pérez, 1998). MILLET produces class-specific temporal evidence through several pooling schemes (Early et al., 2024), TimeMIL models local patterns and temporal dependencies (Chen et al., 2024), and attention-based MIL learns data-dependent instance weights (Ilse et al., 2018). These approaches generally assume fixed bag-label semantics.

Multi-class windows and convention shift. Wearable recognition commonly assigns one label to each fixed window even when activities change within it (Lara and Labrador, 2013; Bulling et al., 2014; Hammerla et al., 2016; Ordóñez and Roggen, 2016). The issue arises across wearable sensing, video activity recognition, industrial operations, and sleep staging (Roggen et al., 2010; Stein and McKenna, 2013; Yoshimura et al., 2024; Logacjov et al., 2021; Kemp et al., 2000). Domain-adaptation research addresses changes in feature distributions, users, devices, or domains (Ganin et al., 2016; Wilson and Cook, 2020; Ragab et al., 2023), whereas ERA-U isolates a change in the rule mapping an unchanged window to one label. It preserves both the input windows and the single-label task while exposing a restricted interface between temporally resolved class scores and window semantics.

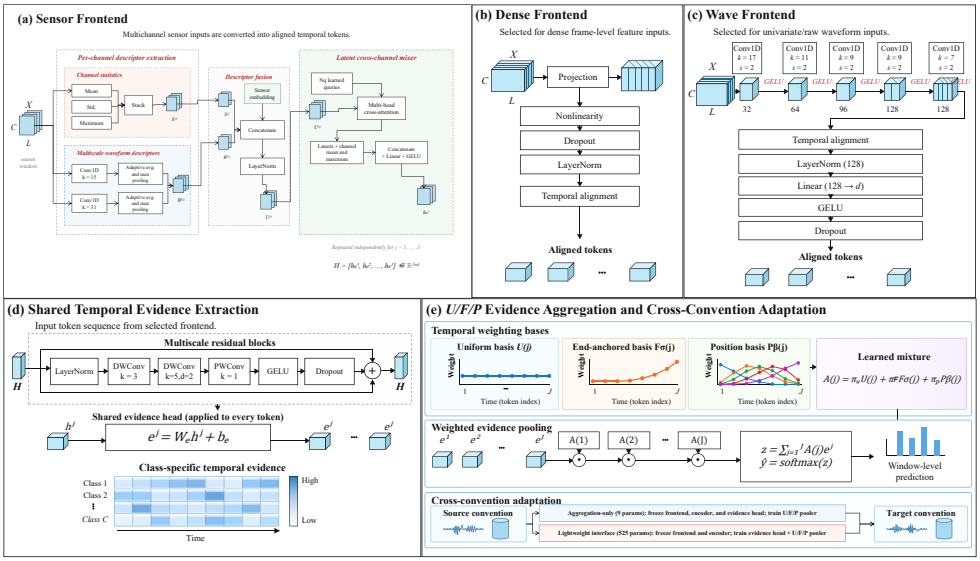


Figure 2: ERA-U overview. A morphology-specific frontend produces aligned tokens; a shared encoder and score head yield temporally resolved class scores; and U/F/P aggregation forms the window prediction. Adaptation preserves the encoder and updates only the prediction interface.

3 METHODOLOGY

ERA-U exposes pre-aggregation temporal class scores and adapts their convention-dependent aggregation through a lightweight interface (Figure 2).

3.1 THE ERA-U FRAMEWORK

ERA-U uses ordered representations and temporally resolved, model-internal class scores. A restricted, class-shared positional aggregator maps the score sequence to one window prediction without a global-classifier bypass. Point-level states define SSW/MSW subsets when available, but never supervise the model.

3.2 ERA-U FOR DEEP TSC: RETHINKING AGGREGATION

A conventional TSC model pools the ordered representation $\mathbf{H}_i = [\mathbf{h}_i^1, \ldots, \mathbf{h}_i^J]$ before classification. ERA-U instead applies one shared linear score interface to every token,

$$\mathbf{e}_i^j = \mathbf{W}_{\mathrm{E}}\mathbf{h}_i^j + \mathbf{b}_{\mathrm{E}}, \qquad \mathbf{E}_i = [\mathbf{e}_i^1, \ldots, \mathbf{e}_i^J] \in \mathbb{R}^{J\times K}, \tag{1}$$

where K is the number of classes and no global classifier bypasses $\mathbf{E}_i$. Each pre-aggregation score therefore remains aligned with a temporal region.

ERA-U aggregates this evidence with three normalized positional bases. Uniform $U(j) = 1/J$ retains global evidence. The end-anchored basis F_σ is a normalized Gaussian centered at the final token, with learned width σ. The flexible basis $P_{\boldsymbol{\beta}}$ is a normalized fourth-order Bernstein expansion over relative position. These are inductive biases rather than fixed implementations of Last, Center, or Majority.

The bases are mixed and applied directly to the temporal class scores:

$$\begin{aligned} \boldsymbol{\pi} &= \mathrm{softmax}(\alpha_U, \alpha_F, \alpha_P), \\ A(j) &= \pi_U U(j) + \pi_F F_\sigma(j) + \pi_P P_{\boldsymbol{\beta}}(j), \\ \mathbf{z}_i &= \sum_{j=1}^{J} A(j)\mathbf{e}_i^j. \end{aligned} \tag{2}$$

Three mixture logits, one width, and five Bernstein coefficients give nine aggregation parameters. Window labels learn their mixture; no basis is pre-assigned to a labeling rule.

Scope. Because $A(j)$ is normalized and shared across classes, the aggregation is algebraically equivalent to positional weighted pooling followed by a linear classifier. ERA-U therefore does not claim a larger hypothesis class. Its contribution is an explicit no-bypass interface whose nine aggregation parameters can be adapted independently of the temporal representation.

3.3 Temporal Evidence and Convention Adaptation

Interpretation and adaptation. Window labels do not uniquely identify $\mathbf{e}_i^j$; evidence denotes model-internal pre-aggregation class scores, not recovered point-level labels or calibrated probabilities. Weighted terms $A(j)\mathbf{e}_i^j$ are exact logit contributions, enabling temporally resolved diagnosis.

We study aggregation-only adaptation, which freezes the frontend, encoder, and score head and updates nine U/F/P parameters, and lightweight-interface adaptation, which also updates the score head for 525 parameters.

3.4 ERA-U Model Design

Three input frontends produce J vectors of dimension d: the sensor frontend combines channel statistics, shared waveform filters, and cross-channel attention; the dense frontend projects frame features; and the waveform frontend uses a five-layer strided convolutional pyramid.

The shared encoder uses residual depthwise convolutions of widths 3 and 5, pointwise convolution, GELU, dropout, and replicate padding. The evidence interface and U/F/P aggregator then produce the prediction. We use $d = 128$ and two blocks.

4 Experimental Setup

Datasets and windows. We use Opportunity (Roggen et al., 2010), 50Salads (Stein and McKenna, 2013), OpenPack (Yoshimura et al., 2024), HARTH (Logacjov et al., 2021), and SleepEDF (Kemp et al., 2000). Opportunity, OpenPack, and HARTH contain multichannel wearable signals; 50Salads contains dense frame-level video features; and SleepEDF contains physiological waveforms. Opportunity, 50Salads, and OpenPack use window lengths 64, 256, and 512, while HARTH uses 128, 256, and 512. SleepEDF retains its native 30-second epochs and is excluded from analyses requiring point-level states or alternative rules. For the other datasets, Last, Center, and Majority select the final, central, and longest-duration state, respectively. The input windows are unchanged across rules; a rule-conflicting window is one for which a selected rule pair assigns different labels.

Baselines and metrics. We compare ERA-U with FCN (Wang et al., 2017), ModernTCN (Luo and Wang, 2024), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), TSLANet (Eldele et al., 2024), TS2Vec (Yue et al., 2022), TimeMIL (Chen et al., 2024), and MILLET (Early et al., 2024). Macro-F1 is reported on all test windows and on MSWs. For fragmentation analysis, MSWs are grouped by one, two, three, or at least four state switches. Error destinations are categorized as correct, incorrect but present within the window, or incorrect and absent from the window.

Cross-convention protocol. We evaluate all six directed transfers among Last, Center, and Majority. Source models are adapted using target-rule training labels, selected by target-rule validation Macro-F1, and evaluated against target-rule test labels, with conflict-window Macro-F1 as the primary metric. We compare four adaptation settings: (i) the unadapted source model, with no target-rule update; (ii) aggregation-only adaptation, which updates the nine U/F/P parameters; (iii) a frozen-encoder baseline with global-average pooling (GAP) and a newly trained linear classifier; and (iv) lightweight-interface adaptation, which updates the evidence head and aggregator while keeping the temporal encoder frozen (525 trainable parameters).

Implementation and code. ERA-U uses PyTorch, AdamW, cross-entropy loss, validation Macro-F1 selection, $d = 128$, two residual blocks, and dropout 0.2. The anonymous core implementation

and training/evaluation entry points are available at `https://anonymous.4open.science/r/anonymous-code-F1EF/`.

5 Experimental Results and Analysis

We address four questions: **RQ1:** Can ERA-U classify MSWs while retaining competitive performance on all windows? **RQ2:** Where do MSW errors occur, and how are they related to competing within-window states? **RQ3:** Can temporal evidence learned under one convention be reused under another? **RQ4:** Does ERA-U benefit from class-specific evidence and learned aggregation?

5.1 Mixed-State Window Classification (RQ1)

Method	Rule	Opportunity						50Salads						OpenPack						HARTH						SleepEDF	
		64		256		512		64		256		512		64		256		512		128		256		512		30s	
		All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix	All	Mix
FCN	Last	.581	.479	.487	.376	.445	.357	.512	.273	.427	.329	.380	.309	.702	.669	.707	.726	.608	.616	.739	.373	.773	.442	.629	.338		
	Center	.472	.449	.562	.543	.403	.397	.528	**.409**	.529	**.523**	**.446**	**.414**	.710	.628	.777	.770	.688	.681	.784	**.495**	**.820**	.498	.692	.488	.659	.507
	Majority	.475	.469	.526	.547	.417	.362	.532	**.402**	.537	.536	.400	.355	.718	.634	.766	.761	.589	.581	.760	**.527**	.793	**.589**	.742	.521		
Patch	Last	.518	.379	.446	.361	.546	.563	.461	.273	.480	.376	.442	**.434**	.707	.621	.758	.785	.745	.755	.795	.406	.765	**.499**	.758	**.493**		
	Center	.536	.439	.452	.341	.434	.335	.495	.343	.414	.355	.384	.326	.727	.631	.787	.795	.790	.792	**.818**	.460	.795	.517	.759	.496	.605	.456
	Majority	.546	.458	.573	.491	.475	.428	.495	.332	.433	.362	.354	.330	.724	.623	.753	.749	.620	.618	.814	.484	.780	.435	.794	.534		
TSLA	Last	.641	.459	.558	.468	.497	.405	.499	.319	**.507**	**.426**	.383	.331	**.737**	**.703**	**.832**	**.857**	**.819**	**.825**	**.808**	.429	.773	.454	.768	.454		
	Center	.464	.341	.671	.640	.480	.456	.527	.401	.518	.454	.393	.346	**.743**	**.676**	.820	.821	.819	.822	.784	.462	.802	**.528**	**.836**	**.563**	.724	.580
	Majority	.418	.315	.585	.563	.557	**.521**	.528	.382	.518	.456	**.449**	**.423**	**.746**	.669	**.827**	**.832**	.683	.683	.801	.459	.801	.579	.844	.576		
TMIL	Last	.381	.323	.331	.324	.392	.360	.494	**.346**	.417	.374	.356	.316	.707	.665	.785	.817	.727	.737	.783	.441	.739	.484	.750	.472		
	Center	.398	.469	.448	.421	.319	.233	.501	.384	.438	.421	.400	.381	.719	.644	.787	.795	.722	.722	.789	.441	.760	.463	.779	.542	.717	.551
	Majority	.416	.489	.450	.396	.439	.363	.498	.360	.451	.410	.371	.335	.717	.632	.780	.789	.600	.598	.799	.484	.786	.485	.782	.538		
MTCN	Last	.645	.461	.540	.442	.390	.344	.477	.276	.248	.176	.167	.174	.709	.650	.728	.749	.689	.702	.780	.374	.718	.414	.644	.373		
	Center	.655	.555	.647	.606	**.584**	**.504**	.515	.378	.387	.343	.199	.151	.725	.646	.785	.793	.764	.769	.779	.437	.760	.440	.679	.395	.671	.525
	Majority	.642	.536	.652	.561	**.589**	.517	.510	.383	.330	.260	.142	.128	.741	**.672**	.760	.763	.598	.598	.772	.441	.765	.457	.645	.358		
TNet	Last	.335	.330	.305	.249	.428	.382	.517	.295	.476	.373	.344	.292	.661	.625	.707	.732	.671	.684	.745	.365	.689	.398	.551	.362		
	Center	.474	.425	.482	.378	.296	.342	**.552**	.398	.486	.418	.421	.370	.689	.619	.751	.763	.758	.765	.742	.433	.680	.364	.631	.421	.571	.452
	Majority	.467	.422	.506	.429	.285	.306	**.548**	.387	.483	.407	.379	.335	.680	.608	.728	.743	.595	.600	.775	.442	.661	.437	.626	.450		
TS2Vec	Last	.480	.326	.447	.283	.276	.201	.500	.270	.434	.359	.387	.356	.721	.686	.799	.828	.761	.773	.719	**.466**	.737	.472	.732	.486		
	Center	.388	.412	.411	.481	.363	.381	.518	.365	.523	.444	.435	.385	.726	.673	**.825**	**.842**	.828	.834	.726	.389	.725	.444	.741	.496	.722	.578
	Majority	.463	.417	.534	.444	.570	.496	.515	.365	.520	.443	.408	.367	.733	.667	.810	.826	.676	.676	.752	.425	.772	.462	.804	.570		
MILLET	Last	.823	.574	.579	.469	.412	.356	.493	.329	.462	.343	.363	.325	.715	.701	.797	.816	.717	.725	.791	.435	.763	.434	.675	.430		
	Center	.522	.424	.648	.507	.544	.432	.532	.391	.528	.507	.413	.383	.732	.667	.799	.797	.710	.705	.811	.483	.793	.488	.772	.499	.720	.572
	Majority	.320	.391	.630	.540	.519	.419	.483	.395	**.538**	**.538**	.439	.380	.726	.657	.811	.811	.564	.560	**.826**	.518	.789	.496	**.862**	**.609**		
ERA-U	Last	**.876**	**.655**	**.810**	**.724**	**.621**	**.658**	**.523**	.311	.483	.383	**.459**	.405	.733	.683	.798	.827	.799	.811	.782	.398	**.815**	.478	**.774**	.467		
	Center	**.896**	**.762**	**.772**	**.674**	.540	.432	.540	.380	**.539**	.488	.355	.300	.742	.654	.821	.835	**.838**	**.839**	.817	.457	.801	.468	.834	.553	**.727**	**.588**
	Majority	**.891**	**.622**	**.781**	**.706**	.509	.440	.547	.385	.516	.437	.363	.327	.741	.645	.818	.817	**.719**	**.720**	.822	.452	**.838**	.515	.852	.596		

Table 1: Macro-F1 across datasets, window lengths, and labeling rules. Each entry reports *All/Mix*; best values are bold.

Note. SleepEDF is a real-world polysomnography dataset that provides one sleep-stage label per 30-s epoch but no point-level temporal state annotations within an epoch. Consequently, Last, Center, and Majority convention labels cannot be constructed for SleepEDF; we report only its native AASM protocol. *Abbreviations:* Patch = PatchTST, TSLA = TSLANet, TMIL = TimeMIL, MTCN = ModernTCN, TNet = TimesNet.

We compare ERA-U with convolutional, Transformer, representation-learning, and multiple-instance methods on all windows and MSWs (Table 1). Most methods exhibit an SSW–MSW gap, so the difficulty is not architecture-specific.

ERA-U gains most on Opportunity, averaging 0.744 All and 0.630 Mix Macro-F1 and exceeding the strongest baseline averages by 0.150 and 0.127. It also has the highest average All score on HARTH and the best All/Mix scores on SleepEDF. It matches the strongest average All result on 50Salads but trails on Mix, and is within 0.007 of the best OpenPack averages. Thus, its gains are largest when labels must resolve substantial within-window competition.

5.2 Error Analysis on Mixed-State Windows (RQ2)

Increasing window fragmentation. We group MSWs by one, two, three, or at least four state switches. Figure 3 shows the clearest degradation on OpenPack, whereas ERA-U remains comparatively stable on Opportunity and HARTH. Switch count is only a coarse descriptor: it omits state identity, duration imbalance, ordering, and boundary position, all of which may substantially affect class competition and determine which labeling convention dominates the final window-level prediction.

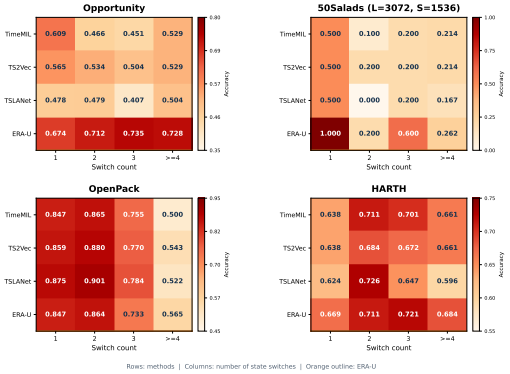


Figure 3: MSW accuracy by state-switch count; the outlined row is ERA-U.

The non-monotone 50Salads profile and the differing trends across datasets show that boundary count alone does not totally order MSW difficulty. Windows with the same count can contain very different competing classes and temporal dominance. ERA-U's relative stability therefore supports robustness to fragmentation in Opportunity and HARTH, rather than a universal monotonic relation between switch count and accuracy.

Where incorrect predictions go. We divide predictions into correct, incorrect but present within the window, and incorrect and absent. ERA-U's present/absent error rates are 23.61/4.17% on Opportunity, 40.68/28.81% on 50Salads, and 23.75/6.70% on HARTH; its OpenPack pattern is comparable to the baselines, indicating that convention-related errors coexist with representation failures whose relative importance varies substantially across datasets and sensing environments.

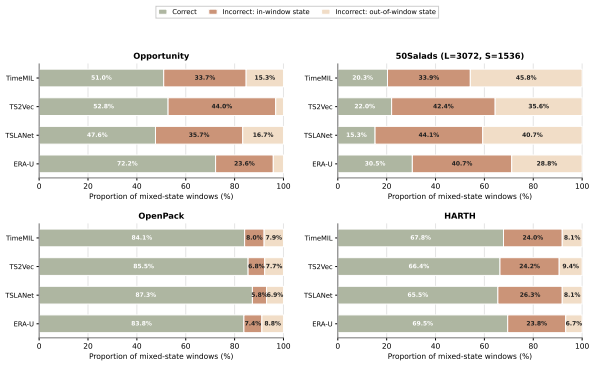


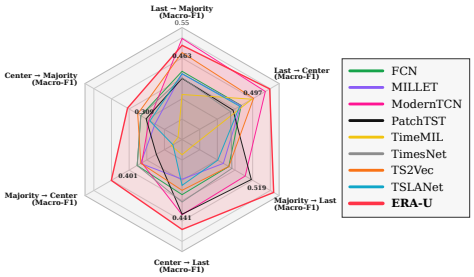
Figure 4: MSW predictions: correct, incorrect but present, and incorrect and absent.

Present-state errors support the central motivation: the model often selects another state that the signal genuinely contains. Absent-state errors instead expose representation or class-confusion failures that a different convention cannot explain. Their larger share on 50Salads marks an important limitation, while the Opportunity and HARTH patterns are dominated by structured disagreement among competing in-window states.

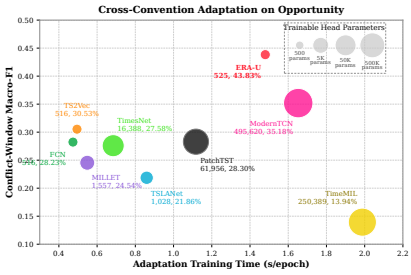
5.3 Cross-Convention Adaptation (RQ3)

We evaluate all six directed transfers among Last, Center, and Majority on unchanged windows. Since most labels agree across rules, evaluation focuses on rule-conflicting windows. Without adaptation, ERA-U retains the source label on 64.73% and 65.68% of Last-to-Majority and Last-to-Center conflicts, but selects the target on only 32.61% and 31.82%. Convention change therefore creates a concentrated decision shift.

We freeze the encoder and adapt with target-convention labels. The strict protocol updates nine U/F/P parameters; the lightweight protocol also updates the evidence head (525 parameters). ERA-U leads conflict-window Macro-F1 in five of six directions, with ModernTCN higher on Last-to-Majority (Figure 5). The same representation can therefore be reinterpreted without end-to-end retraining.



(a) Conflict-window Macro-F1 over six directed transfers.



(b) Macro-F1 versus head-training time; area denotes trainable parameters.

Figure 5: Cross-convention adaptation on Opportunity: **(a)** transfer performance and **(b)** efficiency using the full target-convention training set.

The efficiency panel reports head-training time on the horizontal axis and trainable parameters by bubble area. With 10% target-rule windows, ERA-U reaches 35.55% conflict-window Macro-F1, 2.06 points above PatchTST (Table 2).

Table 2: Cross-convention adaptation with 10% of target-rule training windows. Conflict-window Macro-F1 (%) is averaged over six directed transfers.

Method	FCN	MILLET	ModernTCN	PatchTST	TimeMIL	TimesNet	TS2Vec	TSLANet	ERA-U
Macro-F1	28.33	24.18	26.00	33.49	19.27	24.48	25.45	18.81	**35.55**

5.4 Ablation Studies: Aggregation and Evidence Use (RQ4)

We retrain Opportunity models with uniform aggregation and the same encoder. Learned U/F/P improves All and Mix Macro-F1 under every rule, by 0.102 and 0.053 on average (Table 3). Convention-dependent weighting therefore adds value beyond uniform evidence averaging.

Table 3: Learned versus uniform aggregation on Opportunity (Macro-F1).

Rule	Uniform		Learned U/F/P	
	All	Mix	All	Mix
Last	.719	.628	**.809**	**.724**
Center	.685	.665	**.772**	**.674**
Majority	.651	.653	**.780**	**.706**
Average	.685	.649	**.787**	**.701**

We modify the learned aggregation at inference while fixing the frontend, encoder, and temporal score head. This isolates reliance on the learned positional structure without introducing retraining effects.

Table 4: Inference-only U/F/P ablation. Values are MSW Macro-F1 averaged over Last, Center, and Majority.

Variant	Opportunity	50Salads	OpenPack	HARTH	Mean
Full U/F/P	**.701**	**.416**	**.800**	**.504**	**.605**
Without U	.679	.386	.798	.495	.590
Without F	.635	.399	.783	.495	.578
Without P	.660	.388	.774	.490	.578
Equal U/F/P	.672	.408	.777	.477	.583
Uniform	.644	.398	.775	.485	.576
Reversed weights	.588	.339	.727	.392	.512

Every intervention lowers the four-dataset mean. Removing F or P produces the largest component-wise decrease, while reversing temporal orientation is worst overall. The result supports complementary and oriented positional priors rather than performance arising only from an additional parameter count.

Target-position diagnostic. Excluding target-associated positions sharply reduces performance, while retaining only them is not consistently better: the oracle leads only on 50Salads, whereas ERA-U leads on Opportunity, OpenPack, and HARTH. This indicates that distributed contextual scores and the window-labeling convention jointly determine the decision.

5.5 Discussion and Scope (RQ1–RQ4)

Reading the four questions together. RQ1–RQ4 form one chain of evidence. RQ1 establishes competitive classification under within-window competition; RQ2 shows that many errors select states present in the input; RQ3 tests whether frozen temporal representations can be redirected under a new convention; and RQ4 shows that learned, oriented aggregation improves on uniform pooling. Together, the results connect mixed-state difficulty, structured disagreement, low-capacity adaptation, and positional ablations. They support an operational separation between temporal scoring and window semantics without requiring the intermediate scores to be uniquely identifiable local labels.

Functional scope of the factorization. Because the aggregation weights are normalized and shared across classes, ERA-U can be rewritten as

$$\mathbf{z}_i = \sum_j A(j)(\mathbf{W}_{\mathrm{E}}\mathbf{h}_i^j + \mathbf{b}_{\mathrm{E}}) = \mathbf{W}_{\mathrm{E}}\left(\sum_j A(j)\mathbf{h}_i^j\right) + \mathbf{b}_{\mathrm{E}}. \tag{3}$$

The model therefore does not introduce a larger hypothesis class than positional weighted pooling followed by a linear classifier. Its contribution lies instead in the restricted organization of that computation. The score head is applied before aggregation, the same class-shared weighting acts on every score dimension, and no independent global classifier can bypass the interface. This organization exposes the sequence that is actually aggregated and isolates a small set of parameters that can be optimized while the frontend, encoder, and, in the strict protocol, score head remain fixed. The distinction is important: improved results should be attributed to the inductive restriction and adaptation interface, not to an unsupported claim of greater representational capacity.

What temporal evidence means. Window-level supervision cannot uniquely recover point-level semantic states. For any perturbation sequence satisfying $\sum_j A(j)\Delta\mathbf{e}_i^j = \mathbf{0}$, replacing $\mathbf{e}_i^j$ by $\mathbf{e}_i^j + \Delta\mathbf{e}_i^j$ leaves the aggregated logits unchanged. Consequently, $\mathbf{e}_i^j$ is neither a calibrated local-state probability nor a uniquely identified latent label. We use *evidence* in the more limited, model-internal sense of pre-aggregation class scores. Within a trained model, $A(j)e_{i,c}^j$ is nevertheless an exact additive contribution to class-c logit $z_{i,c}$. This makes temporal decomposition useful for model diagnosis while keeping its semantic status explicit. Point-level annotations, intervention-based faithfulness tests, or additional identifiability constraints would be required before making stronger localization claims.

Why low-capacity adaptation is informative. Updating nine aggregation parameters is not intended to solve arbitrary domain shift. It tests whether a representation learned under one labeling rule still contains information that another rule can select. A successful transfer is therefore evidence of reusable representation under a controlled change in window semantics, not evidence that the source and target tasks are identical. The 525-parameter variant relaxes the interface by also updating the score head while keeping the encoder frozen. Reporting both settings separates a strict aggregation test from a still lightweight alternative that can recalibrate class scores. The comparison with newly trained heads further matters because parameter count and observed runtime capture different costs: runtime also includes representation loading, batching, and device utilization, whereas trainable capacity describes the statistical flexibility exposed to target-rule labels.

Failure modes and negative evidence. The results also identify cases in which convention adaptation is insufficient. Switch count alone does not determine difficulty because it omits state identity,

duration, ordering, and the location of boundaries. This helps explain why fragmentation trends are not monotone for every dataset. Likewise, an incorrect prediction that is absent from the input cannot be explained solely as a convention choosing a different supported state. The larger absent-error component on 50Salads indicates that representation or class-confusion errors remain important there. Performance differences across transfer directions also show that convention changes are asymmetric: the information needed to move from one rule to another may be present in the frozen representation but differently accessible through the restricted interface. These negative cases bound the interpretation of the positive transfer results and motivate reporting each directed transfer rather than only an average.

Validity and generalization. The evaluation spans wearable sensors, dense video features, and physiological waveforms, but it does not cover every form of window construction or label semantics. Last, Center, and Majority are deterministic conventions with known target labels, and adaptation uses supervised target-rule windows. The study therefore does not address unknown conventions, unsupervised rule discovery, annotator disagreement, multilabel targets, or continuously varying event boundaries. SleepEDF further illustrates the boundary of the protocol: without point-level temporal annotations inside an epoch, alternative Last, Center, and Majority labels cannot be constructed, so it contributes only to the native-rule classification comparison. Moreover, point-level annotations are used to define diagnostic subsets where available, not as training supervision. Generalization to variable-length windows, streaming decisions, pretrained encoders, and probabilistic conventions remains to be established.

Relation to alternative pooling mechanisms. Global average pooling treats every position symmetrically and provides no convention-specific control. Data-dependent attention and multiple-instance pooling can be more expressive, but their flexibility makes it harder to isolate whether adaptation changes representation, instance scoring, or aggregation. ERA-U instead uses a class-shared positional sequence formed from three normalized bases. This restriction is not universally optimal; it exchanges expressive freedom for a compact intervention point whose parameters directly control the role of relative position. The ablations therefore show that this restricted structure is useful in the studied setting, not that fixed positional aggregation generally dominates attention-based pooling.

Practical implication. ERA-U is most appropriate when the temporal representation is expensive to retrain, the target labeling convention is specified, and a small labeled target set is available. In that setting, the explicit interface provides a controlled place to adapt and inspect the decision. When the shift changes sensing conditions, class definitions, or the underlying input distribution, a richer domain-adaptation method may be necessary. The central outcome is thus not an identifiable decomposition of latent states, but a transparent and deliberately limited boundary between reusable temporal scores and the convention that converts them into one window label.

6 Conclusion

We studied convention shift in multi-class windows with unchanged inputs. ERA-U exposes a no-bypass score–aggregation interface without claiming greater expressivity than positional pooling or unique recovery of point-level states. Across five datasets, MSWs are difficult, many errors select competing states present in the input, and ERA-U leads five of six directed transfers. Ablations support complementary temporal priors, while nine- and 525-parameter updates reuse a frozen temporal representation. Future work will study local supervision, faithfulness, probabilistic conventions, and variable-length or streaming models.

AI Use Statement

Generative AI tools were used to assist with research ideation and execution, critical feedback on methodological exposition and result interpretation, literature search and organization, drafting and language editing, and LaTeX and table formatting. All AI-assisted suggestions, citations, equations, code changes, experimental results, and scientific claims were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This study uses existing benchmark datasets and does not collect new human-participant data. We use the datasets only for aggregate time-series classification experiments, follow their original access and usage conditions, and make no attempt to identify individual participants. The principal ethical considerations are responsible handling of wearable and physiological signals and faithful reporting of experimental results.

Reproducibility statement

Section 4 specifies the datasets, window protocols, labeling conventions, baselines, metrics, adaptation settings, optimization procedure, and model dimensions used in the experiments. Section 3 defines the complete ERA-U computation and its trainable adaptation interfaces. The anonymous repository linked in the abstract provides the core implementation and training and evaluation entry points needed to reproduce the reported analyses.

References

Anthony Bagnall et al. The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31:606–660, 2017.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271, 2018.

Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3):33, 2014.

Xiwen Chen et al. TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning. In *ICML*, 2024.

Hoang Anh Dau et al. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34:1454–1495, 2020.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MiniRocket: A very fast (almost) deterministic transform for time series classification. In *KDD*, pages 248–257, 2021.

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.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997.

Joseph Early et al. Inherently interpretable time series classification via multiple instance learning. In *ICLR*, 2024.

Emadeldeen Eldele et al. Time-series representation learning via temporal and contextual contrasting. In *IJCAI*, pages 2352–2359, 2021.

Emadeldeen Eldele et al. TSLANet: Rethinking transformers for time series representation learning. In *ICML*, 2024.

Hassan Ismail Fawaz et al. Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, 33:917–963, 2019.

Hassan Ismail Fawaz et al. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34:1936–1962, 2020.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *NeurIPS*, 2019.

Yaroslav Ganin et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *IJCAI*, 2016.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

Fazle Karim et al. LSTM fully convolutional networks for time series classification. *IEEE Access*, 6:1662–1669, 2018.

Bob Kemp et al. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Oscar D. Lara and Miguel A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013.

Jason Lines, Sarah Taylor, and Anthony Bagnall. Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles. *Machine Learning*, 107:1771–1812, 2018.

Aleksej Logacjov et al. HARTH: A human activity recognition dataset for machine learning. *Sensors*, 21(23):7853, 2021.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *ICLR*, 2024.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *NeurIPS*, 1998.

Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. Bake off redux: A review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38:1958–2031, 2024.

Yuqi Nie et al. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

Francisco J. Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

Mohamed Ragab et al. ADATIME: A benchmarking suite for domain adaptation on time series data. *ACM Transactions on Knowledge Discovery from Data*, 17(8):1–18, 2023.

Daniel Roggen et al. Collecting complex activity datasets in highly rich networked sensor environments. In *INSS*, pages 233–240, 2010.

Alejandro Pasos Ruiz et al. The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35:401–449, 2021.

Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *UbiComp*, pages 729–738, 2013.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *ICLR*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

Ashish Vaswani et al. Attention is all you need. In *NeurIPS*, 2017.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *IJCNN*, 2017.

Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):51, 2020.

Gerald Woo et al. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *ICLR*, 2022.

Haixu Wu et al. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *ICLR*, 2023.

Naoya Yoshimura et al. OpenPack: A large-scale dataset for recognizing packaging works in IoT-enabled logistic environments. In *IEEE PerCom*, 2024.

Zhihan Yue et al. TS2Vec: Towards universal representation of time series. In *AAAI*, 2022.

George Zerveas et al. A transformer-based framework for multivariate time series representation learning. In *KDD*, pages 2114–2124, 2021.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *NeurIPS*, 2022.