FROM WHICH TO WHEN: MULTIPLE INSTANCE LEARNING FOR TEMPORAL LOCALIZATION IN TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple instance learning (MIL) enables time series classification from annotations that list the classes in a sequence without their temporal locations. We extend MIL for time series from sequence classification to temporal localization, which requires locating each class at the timesteps where it occurs. Knowing which classes occur does not determine when they occur, and we therefore identify two design principles for recovering the latter. Specifically, evidence should be selected sparsely and separately for each class, and instances outside the selected evidence should receive supervision constrained by the bag label. We propose StepMIL, which combines class-wise sparse temporal pooling with Local Soft-Target Propagation (LSTP). The pooling bases each class prediction on a small fraction of timesteps selected for that class. LSTP propagates bag-constrained soft targets between neighboring timesteps, weighted by source reliability and edge affinity, and is used only during training. We evaluate StepMIL on six datasets from human activity recognition, sleep staging and electrocardiography, with bags of up to 90,000 instances, against five MIL methods for bag classification. StepMIL improves temporal localization and sequence classification together, attaining the highest instance macro-F1 and Cohen's κ on every dataset and the highest bag mAP on five of the six datasets.

1 INTRODUCTION

Time series classification (TSC) underlies applications ranging from physiological monitoring to human activity recognition (Ismail Fawaz et al., 2019; Supratak et al., 2017; Hannun et al., 2019; Ordóñez & Roggen, 2016; Ronao & Cho, 2016). In many of these applications, where each class occurs matters more than which classes a recording contains. A sleep study is read through its hypnogram (Phan et al., 2019; Perslev et al., 2019), an arrhythmia through the beats at which it appears (Moody & Mark, 2001), and an activity log through the intervals it spans (Reyes-Ortiz et al., 2016; Gu et al., 2021). Annotating every timestep, however, is labor-intensive and often ambiguous near state transitions (Rosenberg & Van Hout, 2013; Guillot et al., 2020), whereas annotating which classes a recording contains is inexpensive. Multiple instance learning (MIL) addresses this setting by treating a sequence as a bag of temporal instances trained from bag-level labels (Dietterich et al., 1997; Maron & Lozano-Pérez, 1998; Carbonneau et al., 2018), and recent temporal MIL methods couple this formulation with sequence encoders to obtain timestep-level evidence (Early et al., 2024; Chen et al., 2024).

Bag-level accuracy, however, does not guarantee instance-level attribution (Vanwinckelen et al., 2016; Carbonneau et al., 2018). A bag label is satisfied whenever the aggregated score of each present class is high, regardless of which timesteps produce that score. A model can therefore classify a bag correctly while mislabeling most of its timesteps, or even assigning a single class to all of them. Across the models we evaluate, similar bag-level performance coincides with widely different instance-level performance (Section 4.3). When a bag contains several co-occurring classes, recognizing which classes are present and locating them in time are thus distinct problems (Zhou et al., 2012; Liu et al., 2012).

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.

We therefore propose two design principles for locating classes in time from multi-label bag annotations. **(C1) Class-wise sparse evidence selection.** A bag label states only that a class occurs somewhere in the sequence, and the evidence for it often occupies a small part of the bag, such as a single heartbeat in a ten-second recording (Moody & Mark, 2001). Pooling that weights every timestep makes the bag score depend on how long a class lasts. Its bag loss also raises the score of each present class at every timestep, including timesteps where that class is absent, which blurs the assignment of classes to timesteps. Pooling should therefore base each class prediction on a small fraction of informative timesteps, selected separately for each class because co-occurring classes are supported by different portions of the sequence. **(C2) Bag-constrained supervision of unselected instances.** Evidence sufficient for class presence leaves most instances unconstrained, including instances within the same state as the selected evidence. These instances require supervision whose targets are restricted to the classes observed in the bag.

We instantiate these principles in StepMIL (Sparse Temporal Evidence and Propagation for MIL), which combines class-wise sparse pooling with Local Soft-Target Propagation (LSTP). LSTP provides bag-constrained supervision from neighboring predictions and trains the same evidence head as the bag objective. Propagation is used only during training. Controlled ablations show that LSTP improves average instance-level performance while largely preserving bag-level performance. Our contributions are as follows.

- **From sequence classification to temporal localization.** We study MIL for time series as temporal localization under multi-label bag supervision, and articulate two design principles for this setting, class-wise sparse selection (C1) and bag-constrained supervision of unselected instances (C2). We show, formally and empirically, that bag-level performance does not determine instance-level attribution (Proposition 1 and Section 4.3).
- **A method that follows both principles.** In StepMIL, class-wise sparse temporal pooling and LSTP share one evidence head and require only multi-label bag annotations, and LSTP adds no inference cost.
- **A controlled comparison across domains.** We compare against MIL methods for bag classification under a shared encoder and protocol. The evaluation covers six datasets from human activity recognition, sleep staging and electrocardiography, with instance resolutions from 2.8 to 40 ms and bags of 750 to 90,000 instances (Figure 2).
- **Consistent instance-level gains.** Against all five baselines, StepMIL attains the highest instance macro-F1 and κ on every dataset and the highest bag mAP on five of the six datasets.

2 Related Work

Time series classification. Deep learning methods for time series classification (TSC) assign a sequence-level class label to an input time series (Ismail Fawaz et al., 2019; 2020). These methods primarily optimize sequence-level objectives, without explicit supervision of class assignments at individual timesteps. Fully supervised temporal segmentation methods produce timestep-level predictions using densely annotated training sequences (Perslev et al., 2019; Farha & Gall, 2019; Gaugel & Reichert, 2023).

Multiple instance learning. Multiple instance learning (MIL) provides bag-level supervision while individual instance labels remain unobserved (Dietterich et al., 1997; Maron & Lozano-Pérez, 1998; Weidmann et al., 2003; Amores, 2013). Neural MIL aggregates instance representations or scores through max and mean pooling (Wang et al., 2018), attention mechanisms (Ilse et al., 2018), or additive contributions (Javed et al., 2022). Related weak supervision appears in other domains where sequence-level labels supervise temporally localized predictions, including weakly labeled sound event detection (McFee et al., 2018; Wang et al., 2019), weakly supervised temporal action localization (WS-TAL) (Paul et al., 2018; Lee et al., 2020; 2021; Qu et al., 2021; Yun et al., 2024) and set-supervised action segmentation in video (Richard et al., 2018; Li & Todorovic, 2020; Fayyaz & Gall, 2020). These settings share our form of supervision but operate on audio or on features of pretrained video models (Carreira & Zisserman, 2017). Graph-based approaches additionally model dependencies among instances (Zhou et al., 2009; Tu et al., 2019). The distinction between bag-level classification and instance-level prediction is established in the MIL literature (Carbonneau et al.,

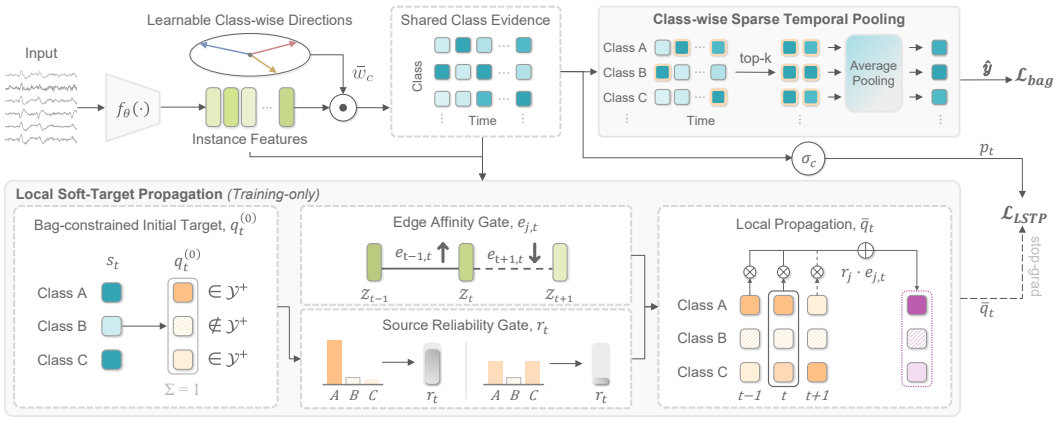

Figure 1: Overall architecture of StepMIL.

2018; Liu et al., 2012). The classical binary formulation concerns the presence of a target class, whereas our setting requires temporal assignments for multiple co-occurring classes.

MIL for time series. MILLET (Early et al., 2024) and TimeMIL (Chen et al., 2024) adapt MIL to TSC through conjunctive and transformer-based pooling, respectively. Their temporal outputs primarily support the interpretation of sequence-level predictions and are examined through perturbation analyses, synthetic class signatures, or attention visualizations. Our study treats per-timestep class assignment as an explicit prediction objective and evaluates these assignments against dense reference annotations under multi-label bag supervision.

3 Method

StepMIL uses complementary objectives for evidence selection and instance learning. Class-wise sparse temporal pooling constructs bag predictions from selected temporal evidence, while LSTP uses local relationships to supervise instance predictions throughout the sequence.

3.1 Problem Formulation

Let $X = (x_1, \ldots, x_T)$, with $x_t \in \mathbb{R}^{d_x}$, denote a time series whose instances correspond to individual sampling times. The observed label is a multi-hot vector $y \in \{0,1\}^C$, where $y_c = 1$ indicates that class c occurs in the sequence. We assume a single class per timestep, $\ell_t \in \{1, \ldots, C\}$, but these timestep labels are unobserved. The set of bag-present classes is $\mathcal{Y}^+ = \{c \mid y_c = 1\}$. Training uses only y to learn bag-level class presence and instance-level class predictions.

3.2 Shared Class Evidence

A temporal encoder produces contextualized representations $z_t = f_\theta(X)_t \in \mathbb{R}^D$. A shared cosine head maps each representation to class evidence scores:

$$s_{t,c} = \beta \frac{z_t^\top \bar{w}_c}{\|z_t\|_2} + b_c, \qquad \bar{w}_c = \frac{w_c}{\|w_c\|_2}, \tag{1}$$

where $\bar{w}_c$ is the learnable direction of class c, obtained by normalizing a weight vector w_c. Cosine similarity measures the directional alignment between z_t and $\bar{w}_c$, allowing class compatibility to be evaluated independently of feature magnitude. Since cosine similarity has a bounded range, the positive scalar β expands it into discriminative logits and controls the sharpness of class predictions and confidence estimates. The class-specific bias b_c adjusts the evidence offset to account for differences in class prevalence and decision thresholds. The resulting score matrix $S = [s_{t,c}] \in \mathbb{R}^{T \times C}$ is shared by bag aggregation, soft-target construction, and instance prediction. Consequently, bag supervision and LSTP optimize the same evidence head in a common class space.

3.3 CLASS-WISE SPARSE TEMPORAL POOLING

Bag annotations specify class presence without identifying the temporal locations or extent of discriminative evidence. We use class-wise sparse temporal pooling, which averages the k highest scores of each class, to accommodate both class-dependent temporal support and sparse evidence:

$$\mathcal{K}_c = \mathrm{TopK}_t(s_{t,c}, k), \qquad \hat{s}_c = \frac{1}{k} \sum_{t \in \mathcal{K}_c} s_{t,c}, \tag{2}$$

where $\mathcal{K}_c$ contains the indices of the k highest scores for class c, with $1 \leq k \leq T$. Classes select their supporting instances independently, and their selected sets may overlap. Bag logits are supervised with binary cross-entropy:

$$\hat{y}_c = \sigma(\hat{s}_c), \qquad \mathcal{L}_{\mathrm{bag}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{BCE}(\hat{y}_c, y_c)\,, \tag{3}$$

where σ denotes the sigmoid function.

Class-specific evidence allocation. Independent selection allows bag supervision to act on different temporal supports for different classes. Many MIL pooling operators instead share one temporal weight across all classes, including attention-based MIL (Ilse et al., 2018), additive MIL (Javed et al., 2022) and conjunctive pooling (Early et al., 2024). As a representative case of such operators, we consider shared score aggregation $\hat{s}_c = \sum_t a_t s_{t,c}$, where $a_t \geq 0$ is shared across classes. For each pooling rule, let $\delta_c = (\hat{y}_c - y_c)/C$ denote its bag-logit gradient. Holding the shared weights fixed along the direct aggregation pathway gives

$$\left.\frac{\partial \mathcal{L}_{\mathrm{bag}}}{\partial s_{t,c}}\right|_{\mathrm{direct}} = a_t \delta_c. \tag{4}$$

The direct score-gradient vectors across timesteps are therefore nonnegative scalar multiples of the same bag-error vector. When classes co-occur, their negative δ_c values favor increasing all present-class scores at every timestep with $a_t > 0$ through this pathway. For the class-wise operator in Eq. 2, wherever the selected sets remain unchanged,

$$\frac{\partial \mathcal{L}_{\mathrm{bag}}}{\partial s_{t,c}} = \frac{1}{k} \mathbf{1}\{t \in \mathcal{K}_c\} \delta_c. \tag{5}$$

Different selected sets thus permit class-dependent temporal allocation of supervision through the score-aggregation pathway.

Concentrating sparse class evidence. The selected set in Eq. 2 concentrates bag prediction on high-scoring observations for each class. To illustrate the effect of this aggregation, suppose n_c timesteps provide a high score u_c for class c, while the remaining $T - n_c$ timesteps have a lower score $v_c < u_c$. The quantity n_c counts observations with strong class evidence, which may occupy only part of the interval associated with that class. Since the selected set contains $\min\{k, n_c\}$ high-scoring observations, Eq. 2 and mean pooling give

$$\begin{aligned} \hat{s}_c^{\mathrm{mean}} &= v_c + \frac{n_c}{T}(u_c - v_c), \\ \hat{s}_c &= v_c + \frac{\min\{k, n_c\}}{k}(u_c - v_c). \end{aligned} \tag{6}$$

When $n_c \geq k$, the selected set contains only high-scoring observations, and $\hat{s}_c = u_c$ regardless of how long the class persists. When $n_c < k$, the selected set necessarily includes $k - n_c$ lower-scoring observations, and strong evidence receives weight n_c/k. Mean pooling assigns it n_c/T, which is smaller whenever $k < T$. Sparse temporal pooling therefore mitigates rather than eliminates the dilution of classes shorter than the budget, while still combining evidence across the selected set.

3.4 BAG SUPERVISION AND INSTANCE LABELS

By Eq. 5, class-wise sparse temporal pooling updates only the scores in $\mathcal{K}_c$, which directs bag supervision to the strongest evidence for each class. Instances outside the selected sets therefore

receive no bag gradient. The limitation is not specific to this operator: a bag label does not determine the labels of its instances, so no pooling operator can supervise every instance from the bag label alone. Let $m = |\mathcal{Y}^+|$ and call an instance labeling $\ell \in (\mathcal{Y}^+)^T$ *consistent* with y if every class in $\mathcal{Y}^+$ occurs in ℓ at least once.

Proposition 1 *Let $2 \le m \le T$. For every instance prediction $\hat{\ell} \in \{1, \dots, C\}^T$, there exists a labeling ℓ consistent with y such that $\hat{\ell}_t = \ell_t$ holds for at most m of the T timesteps.*

Every consistent labeling yields the same bag label, so bag supervision alone cannot guarantee instance accuracy above m/T, whatever the pooling operator (Appendix A). Instance labels must therefore be recovered from structure beyond the bag label. Temporal continuity provides such structure, since neighboring timesteps within a state tend to share its class. We therefore complement the bag objective with local supervision that exploits this continuity and reaches instances outside the selected sets.

3.5 Local Soft-Target Propagation

Local Soft-Target Propagation (LSTP) implements principle C2 through temporal continuity. Neighboring timesteps within a state tend to share its class, so the bag-constrained prediction at one timestep can serve as a target for its neighbors. Not every prediction, however, is reliable enough to propagate, and not every neighbor belongs to the same state. LSTP therefore propagates targets through two gates. A source reliability gate determines how much each prediction may propagate, and an edge affinity gate weakens propagation between dissimilar neighbors, which are likely to lie on opposite sides of a state transition.

Bag-constrained targets. For $m = |\mathcal{Y}^+| \ge 1$, initial instance-level targets are restricted to the classes observed in the bag,

$$q_{t,c}^{(0)} = \frac{\mathbf{1}\{c \in \mathcal{Y}^+\} \exp(s_{t,c}/\tau_q)}{\sum_{u \in \mathcal{Y}^+} \exp(s_{t,u}/\tau_q)}, \tag{7}$$

where $\tau_q > 0$.

Source reliability gate. We take certainty about a bag-present class as the criterion for a reliable source. A prediction spread over several present classes is uncertain and carries little information for its neighbors. The source reliability gate therefore quantifies the certainty of the initial target through its normalized entropy,

$$r_t = \begin{cases} 1 - \dfrac{H(q_t^{(0)})}{\log m}, & m > 1, \\ 1, & m = 1, \end{cases} \tag{8}$$

where $H(q) = -\sum_{c \in \mathcal{Y}^+} q_c \log q_c$ is the entropy over the bag-present classes.

Edge affinity gate. Adjacent timesteps near a state transition may belong to different classes, and their representations then tend to differ. Feature similarity therefore indicates whether two neighbors share a state, without requiring state boundaries to be predicted explicitly. The edge affinity gate weights the edge from timestep j to timestep t by

$$e_{j,t} = \frac{1 + \cos(z_j, z_t)}{2}, \qquad e_{t,t} = 1. \tag{9}$$

Propagation. Let $\mathcal{N}_t = \{t-1, t, t+1\} \cap \{1, \dots, T\}$. Each timestep aggregates the initial targets of itself and its neighbors, weighting each source by its reliability and each edge by its affinity,

$$\mu_t = \sum_{j \in \mathcal{N}_t} r_j e_{j,t}, \qquad \bar{q}_t = \mathrm{sg}\left[\frac{\sum_{j \in \mathcal{N}_t} r_j e_{j,t} q_j^{(0)}}{\mu_t + \epsilon} \right], \tag{10}$$

where $\epsilon > 0$ provides numerical stability and sg denotes stop-gradient. The self-term keeps the contribution of the timestep's own prediction, and reliable neighbors with similar representations

add evidence to it. Each target is then weighted by its incoming mass and by its own concentration among the bag-present classes,

$$\omega_t = \mathrm{sg}[\mu_t\, \bar{r}_t]\,, \qquad \bar{r}_t = \begin{cases} 1 - \dfrac{H(\bar{q}_t)}{\log m}, & m > 1, \\ 1, & m = 1, \end{cases} \tag{11}$$

so a target that remains split among the present classes after propagation contributes less, even when both gates are disabled.

Objective. The instance predictor normalizes scores over all classes, $p_{t,c} = \mathrm{softmax}_c(s_{t,c}/\tau_p)$ with $\tau_p > 0$, and LSTP supervises it with the weighted cross-entropy

$$\mathcal{L}_{\mathrm{LSTP}} = -\frac{\sum_{t=1}^{T} \omega_t \sum_{c=1}^{C} \bar{q}_{t,c} \log p_{t,c}}{\sum_{t=1}^{T} \omega_t + \epsilon}. \tag{12}$$

3.6 Training and Inference

The overall objective combines the bag loss with LSTP,

$$\mathcal{L} = \mathcal{L}_{\mathrm{bag}} + \lambda_{\mathrm{LSTP}} \mathcal{L}_{\mathrm{LSTP}}. \tag{13}$$

We use $\lambda_{\mathrm{LSTP}} = 0.5$ and average the objective over training bags. The bag loss trains evidence selection for class presence, while LSTP supplies local instance supervision, including at timesteps outside the selected sets (Appendix A). At inference, bag probabilities are $\hat{y}_c$ and instance labels are $\hat{\ell}_t = \arg\max_c p_{t,c}$, both computed from the evidence head alone.

4 Experiments

We evaluate timestep-level class prediction using only bag-level annotations for training. All methods use the same data splits, training budgets, checkpoint-selection criterion, and encoder architecture, providing a controlled comparison of aggregation and supervision mechanisms.

4.1 Experimental Setup

Datasets. We evaluate six datasets from human activity recognition (Opportunity (Chavarriaga et al., 2013), Harth (Logacjov et al., 2021), Hapt (Reyes-Ortiz et al., 2016)), sleep staging (DOD-H (Guillot et al., 2020), Hmc (Alvarez-Estevez & Rijsman, 2021)), and electrocardiography (MIT-BIH (Moody & Mark, 2001; Goldberger et al., 2000)). The benchmarks span a broad range of temporal scales (Figure 2). For each dataset, bags are fixed-length windows, instances correspond to individual sampling times, and bag labels identify the classes present within each window. Timestep annotations are reserved for evaluation. Appendix B details the datasets, data splits, bag construction, and treatment of background timesteps.

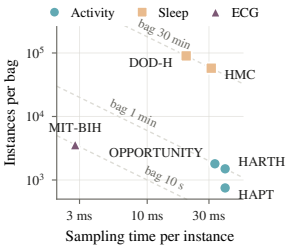

Figure 2: Temporal scales of the six benchmarks.

Baselines. We compare StepMIL with five MIL baselines for bag-level classification: MI-Net (Wang et al., 2018), attention-based MIL (ABMIL) (Ilse et al., 2018), Additive MIL (Javed et al., 2022), MILLET (Early et al., 2024), and TimeMIL (Chen et al., 2024). All methods use the same InceptionTime encoder architecture (Ismail Fawaz et al., 2020). Appendix E specifies how instance-level predictions are obtained for each method.

Evaluation metrics. We report mean average precision (bag mAP) for bag-level classification and accuracy, macro-F1, and Cohen's κ for instance-level predictions $\hat{\ell}_t$ (Section 3.6). Cohen's κ adjusts observed agreement for chance agreement under the predicted and reference class distributions. On MIT-BIH, approximately 90% of timesteps belong to the normal class; consistently predicting this class therefore yields high accuracy but $\kappa = 0$. Macro-F1 and κ consequently serve as the primary instance-level metrics. For Opportunity and Hapt, accuracy (fg acc) and κ are computed over foreground classes (Appendix C).

Table 1: Bag- and instance-level performance against MIL baselines for bag classification. Every cell is the mean ± standard deviation over three seeds, taken over all 12 fold–seed models on OPPORTUNITY and over seed-level averages of the five scorers on DOD-H. **Bold** marks the best value in a column and underline the second-best.

Method	OPPORTUNITY				DOD-H				HARTH			
	bag mAP	fg acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ
MI-Net 2018	.638 ±.143	.310 ±.045	.113 ±.030	.250 ±.044	.894 ±.002	.349 ±.006	.301 ±.006	.166 ±.006	.804 ±.049	.844 ±.013	.567 ±.048	.787 ±.020
ABMIL 2018	.674 ±.100	.293 ±.039	.115 ±.028	.234 ±.042	.897 ±.009	.334 ±.013	.286 ±.007	.145 ±.010	.794 ±.036	.882 ±.016	.597 ±.044	.835 ±.032
Additive MIL 2022	.650 ±.129	.284 ±.077	.111 ±.037	.225 ±.076	.902 ±.005	.331 ±.005	.289 ±.005	.150 ±.006	.767 ±.053	.887 ±.010	.601 ±.056	.842 ±.018
MILLET 2024	.666 ±.119	.286 ±.046	.112 ±.028	.228 ±.047	.894 ±.006	.335 ±.016	.287 ±.008	.148 ±.011	.799 ±.049	.876 ±.011	.586 ±.050	.827 ±.027
TimeMIL 2024	.521 ±.061	.111 ±.068	.031 ±.023	.027 ±.040	.696 ±.010	.267 ±.067	.112 ±.003	.001 ±.002	.803 ±.042	.308 ±.283	.123 ±.106	.169 ±.308
StepMIL (Ours)	**.707** ±.084	**.349** ±.075	**.213** ±.055	**.297** ±.071	**.924** ±.006	**.573** ±.013	**.368** ±.011	**.318** ±.014	**.852** ±.072	**.899** ±.010	**.611** ±.026	**.856** ±.032

Method	HAPT				HMC				MIT-BIH			
	bag mAP	fg acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ
MI-Net 2018	.954 ±.018	.786 ±.012	.548 ±.006	.751 ±.014	.876 ±.032	.367 ±.027	.359 ±.030	.207 ±.039	.459 ±.013	.614 ±.103	.269 ±.022	.062 ±.013
ABMIL 2018	**.965** ±.009	.803 ±.012	.561 ±.008	.771 ±.014	.882 ±.041	.389 ±.058	.374 ±.041	.229 ±.066	.464 ±.008	.365 ±.108	.197 ±.034	.032 ±.010
Additive MIL 2022	.958 ±.005	.793 ±.026	.556 ±.013	.759 ±.030	.870 ±.030	.386 ±.063	.369 ±.043	.225 ±.073	.465 ±.016	.405 ±.007	.214 ±.007	.031 ±.006
MILLET 2024	.961 ±.012	.802 ±.018	.554 ±.004	.770 ±.021	.876 ±.028	.402 ±.054	.378 ±.033	.245 ±.059	.468 ±.036	.450 ±.047	.222 ±.014	.042 ±.004
TimeMIL 2024	.947 ±.011	.325 ±.139	.146 ±.051	.207 ±.153	.767 ±.025	.223 ±.012	.127 ±.059	.020 ±.018	.493 ±.016	.649 ±.435	.183 ±.094	.001 ±.002
StepMIL (Ours)	.955 ±.010	**.880** ±.016	**.655** ±.022	**.860** ±.019	**.916** ±.024	**.494** ±.033	**.449** ±.079	**.327** ±.057	**.497** ±.034	**.864** ±.045	**.362** ±.023	**.269** ±.020

Implementation details. All models are trained for 300 epochs using AdamW with Lookahead. Batch sizes are dataset-specific and identical across methods within each dataset. We retain the checkpoint with the highest validation bag mAP, ensuring that neither optimization nor checkpoint selection uses timestep annotations. For StepMIL, we set $k = \lceil 0.125\,T \rceil$ and $\lambda_{\mathrm{LSTP}} = 0.5$. We report means and standard deviations over three random seeds, except on OPPORTUNITY, where statistics are computed over all 12 fold–seed models. On DOD-H, scores are averaged over the five scorers within each seed before calculating the mean and standard deviation across seeds. Appendix D provides the complete hyperparameter settings.

4.2 COMPARISON WITH BASELINES

Table 1 reports performance at the bag and instance levels. Bag mAP assesses bag-level class recognition, whereas instance accuracy, macro-F1, and Cohen's κ evaluate instance-level class assignments. Macro-F1 weights classes equally regardless of their frequency, while κ adjusts for chance agreement under the predicted and reference class distributions.

StepMIL achieves the highest score in 23 of the 24 dataset–metric comparisons. Its average rank among the six evaluated methods is 1.13, compared with 3.29 for ABMIL, the best-ranked baseline. StepMIL exceeds all five baselines in instance macro-F1 and κ on every dataset, and it also improves bag mAP over the strongest baseline on five of the six datasets, by up to 0.048 on HARTH. Its gains are thus largest in temporal class assignment, while it also strengthens bag-level class recognition.

The improvement is particularly pronounced in Cohen's κ. On MIT-BIH, where approximately 90% of timesteps belong to the normal class, StepMIL achieves $\kappa = 0.269$, compared with 0.062 for the strongest baseline. On DOD-H, κ increases from 0.166 to 0.318. Ventricular and fusion-beat intervals in MIT-BIH can be shorter than the pooling support size k. This parameter specifies the number of observations aggregated for bag prediction and imposes no explicit minimum duration on predicted class intervals. Appendix F presents additional comparisons with WS-TAL methods adapted to time series.

4.3 BAG-LEVEL VERSUS INSTANCE-LEVEL PERFORMANCE

Bag-level supervision can admit multiple instance labelings consistent with the same bag annotation (Proposition 1). Figure 3 shows that models with comparable bag mAP can exhibit substantially different instance-level performance. Among model pairs with an absolute bag mAP difference below 0.01, the maximum instance macro-F1 difference within each dataset ranges from 0.12 to 0.59. This discrepancy supports evaluating instance-level predictions directly, alongside bag-level class recognition.

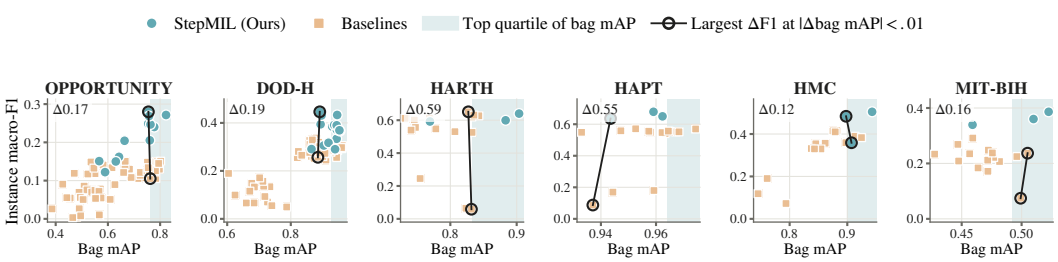

Figure 3: Relationship between bag mAP and instance macro-F1. Each point represents one trained model.

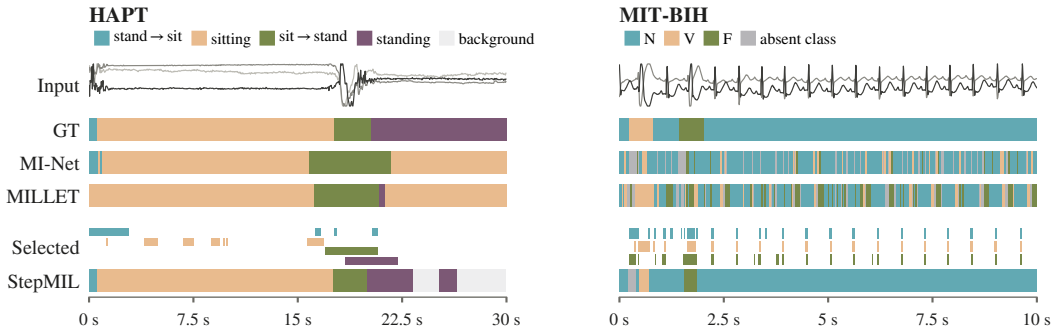

Figure 4: Instance-level predictions for test bags from HAPT and MIT-BIH. GT denotes ground-truth labels, and the narrow bands indicate the timesteps selected for each bag-present class. Additional examples are provided in Appendix J.

StepMIL combines class-specific evidence selection with local instance supervision. Class-wise sparse temporal pooling restricts the direct bag-loss gradient to the selected support of each class (Eq. 5), while LSTP extends auxiliary supervision to timesteps outside these supports (C2). Accordingly, StepMIL achieves the highest mean instance macro-F1 on all six datasets and the highest mean bag mAP on five (Table 1).

4.4 Qualitative Analysis

Figure 4 presents instance predictions alongside the input signals and class-specific temporal selections. Class-wise selection concentrates on characteristic signal patterns, including the orientation change associated with sit-to-stand in HAPT and the widened complexes associated with ventricular and fusion beats in MIT-BIH.

Temporal selection and instance classification involve distinct score comparisons. Top-k pooling ranks timesteps for a given class, whereas instance prediction compares class scores at each timestep. A timestep selected for one class may therefore be assigned to another class with a higher score. On MIT-BIH, the ventricular and fusion intervals contain fewer than k timesteps, so the selected sets extend to other beats. These additional timesteps nevertheless retain their correct instance predictions. LSTP meanwhile supervises these timesteps with targets from neighbors whose predictions favor the normal class. LSTP complements selective pooling with local soft-target supervision across the sequence, including at unselected timesteps. The largely unselected sitting interval in HAPT is recovered as a contiguous segment. The baselines, in contrast, omit brief states or fragment their

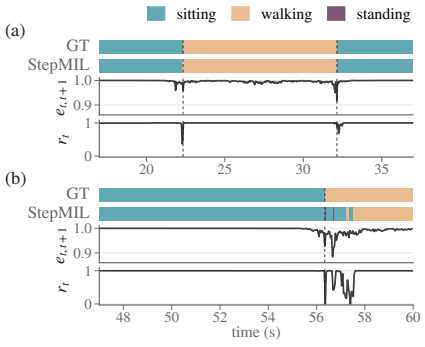

Figure 5: LSTP gates on HARTH test bags. Dashed lines mark transitions.

Table 2: Bag-pooling ablation (C1). LSTP is fixed and only the pooling changes.

Bag pooling	bag mAP	acc	inst mF1
Shared conjunctive (MILLET-style)	.754	.657	.415
Class-wise conjunctive	.727	.641	.333
Class-wise softmax	**.809**	.646	.344
Class-wise sparse (StepMIL)	**.809**	**.677**	**.443**

Table 3: Component ablation (C2). Rows below the first add LSTP to sparse temporal pooling in different configurations.

Variant	bag mAP	acc	inst mF1
Sparse temporal pooling only	.804	.591	.401
+ LSTP without gates	.794	.660	.423
+ LSTP with r, without neighbors	.784	.665	.417
+ LSTP with r	.793	.664	.428
+ LSTP with e	.790	.668	.421
StepMIL (LSTP with r and e)	**.809**	**.677**	**.443**

predictions. Figure 5 shows the LSTP gates on HARTH test bags. The source reliability r_t decreases at annotated transitions, where the prediction is divided between adjacent activities, thereby down-weighting targets propagated across them.

4.5 Ablation Studies

Each ablation modifies a single element of StepMIL under the same protocol. Tables 2 and 3 report averages over the six datasets, and Appendix H defines each variant.

Class-wise sparse selection (C1). Table 2 isolates the aggregation operator while keeping LSTP fixed. Class-wise sparse temporal pooling achieves the best average on all three metrics. Class-wise softmax pooling attains the same bag mAP but a macro-F1 lower by 0.099, because normalized weights distribute each class over timesteps that provide little evidence for it. Both conjunctive variants yield lower values on every metric. These results indicate that class-specific weighting alone is insufficient and that the most effective variant combines class-wise and sparse allocation.

Bag-constrained local supervision (C2). Table 3 examines the contribution of LSTP on top of class-wise sparse temporal pooling. Pooling alone yields a bag mAP comparable to that of StepMIL but the lowest instance-level performance, since unselected instances receive no direct supervision. Bag-constrained targets account for a large part of the improvement. Propagating them without gates ($r_j = e_{j,t} = 1$) increases macro-F1 from 0.401 to 0.423, consistent with C2, but reduces bag mAP below that of pooling alone (0.794 compared with 0.804). With the source reliability gate r enabled, removing the neighbors ($\mathcal{N}_t = \{t\}$) lowers macro-F1 from 0.428 to 0.417, indicating that propagation between adjacent timesteps contributes beyond per-timestep targets. Enabling either gate alone does not restore bag mAP (0.793 and 0.790). With both gates enabled, StepMIL achieves the best average on all three metrics, improving accuracy by 0.086 and macro-F1 by 0.042 over pooling alone and increasing bag mAP to 0.809.

5 Conclusion

We studied temporal MIL, in which per-timestep class assignments are learned from sequence-level class-presence annotations. Bag labels alone do not determine instance-level assignments (Proposition 1), so bag-level performance does not reveal how well a model locates classes in time. StepMIL instantiates two complementary design principles, class-wise sparse evidence selection and bag-constrained supervision of unselected instances. Class-wise sparse temporal pooling bases each class prediction on its own supporting timesteps, and LSTP extends bag-constrained supervision to every timestep without additional inference cost. Across activity, sleep and cardiac recordings, this combination improves temporal class assignment over MIL methods for bag classification while also strengthening bag-level recognition, and the ablations indicate that both principles contribute. These results suggest that locating classes in time from class-presence labels requires identifying where the evidence for each class lies and supervising the instances that this evidence leaves unconstrained.

Limitations. Our evaluation is restricted to time series, and we have not examined whether the approach extends to other modalities. LSTP also propagates targets only between adjacent timesteps. This local design satisfies design principle C2 by supervising every instance with bag-constrained targets, and propagating supervision effectively over wider temporal windows is left for future work.

Reproducibility Statement

Section 4.1 describes the datasets, splits, baselines, evaluation metrics and training protocol, and Appendix D lists every hyperparameter together with the per-dataset window, stride and batch settings. All methods share one encoder, training budget and checkpoint-selection procedure, and no timestep annotation is used for training or checkpoint selection. Appendix E specifies how instance-level predictions are obtained for each baseline.

AI use statement

In this work, we used generative AI tools to assist with baseline adaptations and experimental scripting, and to draft and edit parts of the manuscript. We have reviewed all AI-assisted work, and the authors take full responsibility for the content of this paper.

References

Diego Alvarez-Estevez and Roselyne M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):e0256111, 2021. doi: 10.1371/journal.pone.0256111.

Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353, 2018.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the Kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15): 2033–2042, 2013. doi: 10.1016/j.patrec.2012.12.014.

Xiwen Chen, Peijie Qiu, Wenhui Zhu, Huayu Li, Hao Wang, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7190–7206. PMLR, 21–27 Jul 2024.

Philip de Chazal, Maria O'Dwyer, and Richard B. Reilly. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, 51(7):1196–1206, 2004. doi: 10.1109/TBME.2004.827359.

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, Gavin K. C. Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently interpretable time series classification via multiple instance learning. In *Proceedings of the International Conference on Learning Representations*, 2024.

Yazan Abu Farha and Jürgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3570–3579, 2019.

Mohsen Fayyaz and Juergen Gall. SCT: Set constrained temporal transformer for set supervised action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 498–507, 2020. doi: 10.1109/CVPR42600.2020.00058.

Stefan Gaugel and Manfred Reichert. PrecTime: A deep learning architecture for precise time series segmentation in industrial manufacturing operations. *Engineering Applications of Artificial Intelligence*, 122:106078, 2023.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. A survey on deep learning for human activity recognition. *ACM Computing Surveys (CSUR)*, 54(8): 1–34, 2021.

Antoine Guillot, Fabien Sauvet, Emmanuel H. During, and Valentin Thorey. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(9):1955–1965, 2020. doi: 10.1109/TNSRE.2020.3011181.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1): 65–69, 2019. doi: 10.1038/s41591-018-0268-3.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, 33(4):917–963, July 2019. doi: 10.1007/s10618-019-00619-1.

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(6):1936–1962, 2020.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI Conference on Artificial Intelligence*, 2020.

Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI Conference on Artificial Intelligence*, 2021.

Jun Li and Sinisa Todorovic. Set-constrained Viterbi for set-supervised action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10817–10826, 2020. doi: 10.1109/CVPR42600.2020.01083.

Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. In Steven C. H. Hoi and Wray Buntine (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 253–268, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR.

Aleksej Logacjov, Kerstin Bach, Atle Kongsvold, Hilde Bremseth Bårdstu, and Paul Jarle Mork. HARTH: A human activity recognition dataset for machine learning. *Sensors*, 21(23):7853, 2021. doi: 10.3390/s21237853.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, 10, 1998.

Brian McFee, Justin Salamon, and Juan Pablo Bello. Adaptive pooling operators for weakly labeled sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26 (11):2180–2193, 2018.

George B. Moody and Roger G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. doi: 10.1109/51.932724.

Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. doi: 10.3390/s16010115.

Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. In *European Conference on Computer Vision (ECCV)*, 2018.

Mathias Perslev, Michael Hejselbak Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-Time: A fully convolutional network for time series segmentation applied to sleep staging. In *Advances in Neural Information Processing Systems*, 2019.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019. doi: 10.1109/TNSRE.2019.2896659.

Sanqing Qu, Guang Chen, Zhijun Li, Lijun Zhang, Fan Lu, and Alois Knoll. ACM-Net: Action context modeling network for weakly-supervised temporal action localization. *arXiv preprint arXiv:2104.02967*, 2021.

Jorge-L. Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. Transition-aware human activity recognition using smartphones. *Neurocomputing*, 171:754–767, 2016. doi: 10.1016/j.neucom.2015.07.085.

Alexander Richard, Hilde Kuehne, and Juergen Gall. Action sets: Weakly supervised action segmentation without ordering constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5996, 2018. doi: 10.1109/CVPR.2018.00627.

Charissa Ann Ronao and Sung-Bae Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244, 2016.

Richard S. Rosenberg and Steven Van Hout. The American Academy of Sleep Medicine inter-scorer reliability program: Sleep stage scoring. *Journal of Clinical Sleep Medicine*, 9(1):81–87, 2013. doi: 10.5664/jcsm.2350.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017. doi: 10.1109/TNSRE.2017.2721116.

Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. Multiple instance learning with graph neural networks. In *ICML Workshop on Learning and Reasoning with Graph-Structured Data*, 2019. URL https://arxiv.org/abs/1906.04881.

Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 253–257, 2019. doi: 10.33682/006b-jx26.

Gitte Vanwinckelen, Vinicius Tragante do O, Daan Fierens, and Hendrik Blockeel. Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Mining and Knowledge Discovery*, 30(2):313–341, 2016. doi: 10.1007/s10618-015-0416-z.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern recognition*, 74:15–24, 2018.

Yun Wang, Juncheng Li, and Florian Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, 2019.

Nils Weidmann, Eibe Frank, and Bernhard Pfahringer. A two-level learning method for generalized multi-instance problems. In *Proceedings of the 14th European Conference on Machine Learning*, pp. 468–479, 2003. doi: 10.1007/978-3-540-39857-8_42.

Wulian Yun, Mengshi Qi, Chuanming Wang, and Huadong Ma. Weakly-supervised temporal action localization by inferring salient snippet-feature. In *AAAI Conference on Artificial Intelligence*, 2024.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-I.I.D. samples. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 1249–1256, 2009.

Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.

A Instance Labels Under Bag Supervision

This section proves Proposition 1, discusses its scope, and gives the gradient that LSTP passes to instance scores.

Proof of Proposition 1. Enumerate $\mathcal{Y}^+ = \{c_1, \dots, c_m\}$, choose m distinct timesteps $t_1, \dots, t_m$ and set $\ell_{t_i} = c_i$. Every class in $\mathcal{Y}^+$ then occurs in ℓ, so ℓ is consistent with y, and $\hat{\ell}$ and ℓ can agree on at most these m timesteps. At every other timestep, set ℓ_t to any class in $\mathcal{Y}^+ \setminus \{\hat{\ell}_t\}$, which is nonempty because $m \geq 2$; $\hat{\ell}$ and ℓ disagree there. □

Scope. A learner observes only the input X and the bag label, and both are identical under every consistent labeling. Proposition 1 therefore implies that bag supervision alone cannot guarantee any instance accuracy above m/T, even for a predictor whose bag predictions are perfect. The statement does not depend on the pooling operator. In particular, time-aware pooling, which makes the bag prediction sensitive to the temporal order of the inputs (Chen et al., 2024), improves how order informs the bag label but does not determine which consistent labeling a model recovers. Recovering instance labels thus relies on inductive biases beyond the bag label, such as the encoder and temporal continuity, which LSTP exploits explicitly.

Gradient of the LSTP objective. Eq. 5 gives the bag gradient, which is zero outside $\mathcal{K}_c$. Since $\bar{q}_t$ and ω_t are held fixed, the LSTP objective in Eq. 12 gives

$$\frac{\partial \mathcal{L}_{\mathrm{LSTP}}}{\partial s_{t,c}} = \frac{\omega_t}{\tau_p\big(\sum_u \omega_u + \epsilon\big)}\big(p_{t,c} - \bar{q}_{t,c}\big), \tag{14}$$

which is nonzero at every timestep with $\omega_t > 0$ whose prediction differs from its target, whether or not the timestep is selected. Bag supervision thus locates the evidence for each class, and LSTP extends the resulting predictions to the remaining instances.

B Datasets

Opportunity (gestures track) comprises 17 gesture classes and unannotated background (NULL) recorded from 113 body-worn sensor channels at 30 Hz, with bags spanning 60 s (1,800 timesteps) and four leave-one-subject-out folds. Harth (12 activities, two tri-axial accelerometers) and Hapt (six basic activities, six postural transitions, with unlabeled segments treated as background) are downsampled from 50 to 25 Hz, with bags of 60 s and 30 s, respectively. Dod-H is a sleep-staging benchmark of 25 polysomnography recordings scored independently by five experts. Following its evaluation protocol, we train on the annotations of a single scorer and evaluate against the consensus of the remaining four, repeating this procedure for each scorer. Hmc is a second sleep-staging dataset annotated by a single scorer. For both sleep datasets, a bag covers 60 scoring epochs (30 min) and an instance is a single signal sample rather than a 30 s epoch, yielding 90,000 instances per bag for Dod-H at 50 Hz and 57,600 for Hmc at 32 Hz. The model therefore receives no epoch

grid indicating where stage transitions may occur. MIT-BIH contains the four AAMI heartbeat classes (N, S, V, F) in 10 s bags of raw 360 Hz electrocardiogram signal under the inter-patient DS1/DS2 split (de Chazal et al., 2004). The database annotates each beat at its R-peak, and we label every sample with the class of the beat whose segment contains it, placing segment boundaries at the midpoints between consecutive R-peaks. Its ventricular and fusion beats typically span fewer timesteps than the selection budget k, so this dataset also tests the regime in which a class is shorter than k and sparse temporal pooling necessarily averages in lower-scoring timesteps (Section 3.3). As Figure 2 shows, the datasets cover instance resolutions from 2.8 to 40 ms, bags of 750 to 90,000 instances, and bag durations from 10 s to 30 min. They also differ substantially in temporal structure. A 30 min sleep bag contains 6.6 (DOD-H) to 9.3 (HMC) stage transitions on average, whereas the median MIT-BIH bag contains a single class and its ventricular and fusion beats last less than one second. Table 5 lists the sampling rate after decimation, the bag length, the training stride and the test split of each dataset.

Background timesteps. On OPPORTUNITY (NULL) and HAPT (unlabeled segments), background is not a class. It is excluded from the label space, so a bag label lists only the foreground classes occurring in the window, $m = |\mathcal{Y}^+|$ counts foreground classes, and a window that contains only background has an empty label. Background covers 72% of the test timesteps on OPPORTUNITY and 24% on HAPT, and 19% and 4% of the training bags contain no foreground class. During training, a bag with $m = 0$ contributes only the bag loss, which lowers every class score, and no LSTP term, since the bag-constrained target (Eq. 7) is undefined without a present class. In a bag with $m \geq 1$, LSTP has no background target: background timesteps receive soft targets over the present classes, weighted down when the prediction is split among them. At inference, a timestep is predicted as background when no class probability $\sigma(s_{t,c})$ reaches 0.5 and as $\arg\max_c s_{t,c}$ otherwise. The same fixed threshold is applied to every method without an explicit background output; BaS-Net, ACM-Net and ISSF, which predict a background class, use their own background decision instead. The foreground accuracy and κ (Appendix C) do not depend on this threshold.

C EVALUATION METRICS

Instance-level metrics are computed for each trained model over all test timesteps, pooled across bags, and then aggregated across seeds, folds and scorers as described in Section 4.1. Let N be the number of evaluated timesteps and n_{ab} the number of timesteps whose true label is a and predicted label is b, with row sums $n_{a\cdot}$ and column sums $n_{\cdot b}$. Cohen's κ is

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \frac{1}{N}\sum_c n_{cc}, \qquad p_e = \frac{1}{N^2}\sum_c n_{c\cdot}\, n_{\cdot c}. \tag{15}$$

Here p_o is the accuracy, and p_e is the accuracy expected if predictions were drawn independently of the true labels with the same class frequencies. Perfect agreement gives $\kappa = 1$, chance-level agreement gives $\kappa = 0$, and $\kappa < 0$ indicates agreement below chance. A model that assigns one class to every timestep has $p_o = p_e$ and therefore $\kappa = 0$, however frequent that class is. Macro-F1 is the unweighted mean of the per-class F1 scores over all classes, including the background class on OPPORTUNITY and HAPT. The foreground-class accuracy and κ on these two datasets are computed on the timesteps whose true label is a foreground class, with each prediction taken as the highest-scoring foreground class. Bag mAP is the mean over classes of the average precision of the bag scores.

D IMPLEMENTATION DETAILS

Table 4 lists the settings shared by all datasets, and Table 5 the per-dataset bag construction. Every method in Section 4 uses the same training budget, checkpoint-selection rule and batch size on a given dataset. All methods use the encoder of Table 4, and all methods except the WS-TAL baselines use its optimizer. Test bags never overlap, and no class reweighting or resampling is applied. Each model is trained on a single NVIDIA RTX A6000 GPU (48 GB).

Instance-level read-out. Every method, including ours, labels timestep t with the argmax over classes of its per-timestep class score (Appendix E). For every method except TimeMIL, whose

bag score comes from a class token, this score is the per-timestep term that the method's own bag score aggregates over time. Background on OPPORTUNITY and HAPT is decided as described in Appendix B. We apply this rule to all methods for three reasons. First, the WS-TAL methods natively output temporal proposals, obtained by thresholding their class activation sequences, in most methods at several levels merged by non-maximum suppression, and evaluated by detection mAP at temporal IoU thresholds. These proposals may overlap and assign several classes to one timestep, so converting them into one label per timestep would require choices that the original methods do not specify. Second, their thresholds were set on video benchmarks, and adapting them to our datasets would require timestep annotations, which our protocol uses for neither training nor model selection. Third, a single read-out makes the comparison depend only on the learned class scores. No threshold enters the foreground accuracy and κ, nor any metric on the datasets without background.

Table 4: Hyperparameters shared by all datasets.

Setting	Value
Encoder	InceptionTime, two Inception blocks, 32 filters per branch ($D = 128$)
Dropout	0
Cosine head scale β	learnable (softplus), initialized to 10
Class bias b_c	learnable
Temperatures τ_q, τ_p	1 (the scale is carried by β)
Stability constant ϵ	10^{-8}
Evidence budget k	$\lceil 0.125\,T \rceil$
LSTP weight λ_{LSTP}	0.5
Optimizer	AdamW with Lookahead
Learning rate / weight decay	5×10^{-3} / 10^{-4}
Gradient clipping	2.0 (global norm)
Training epochs	300
Checkpoint selection	every 5 epochs from epoch 20; validation bag mAP, ties by bag macro-F1

Table 5: Per-dataset settings. Bag length is given in instances and in time; instances are single sampling times after decimation.

Dataset	Rate (Hz)	Classes	Bag length	Train stride	Batch	Test split
OPPORTUNITY	30	17 + bg	1,800 (60 s)	1,800	8	one subject per fold (4 folds)
DOD-H	$250 \to 50$	5	90,000 (30 min)	90,000	4	5 recordings
HARTH	$50 \to 25$	12	1,500 (60 s)	750	48	4 subjects
HAPT	$50 \to 25$	12 + bg	750 (30 s)	375	96	official 9 test subjects
HMC	$256 \to 32$	5	57,600 (30 min)	57,600	8	8 recordings
MIT-BIH	360	4	3,600 (10 s)	1,800	32	DS2 (inter-patient)

E BASELINE DETAILS

The MIL and WS-TAL baselines share the encoder of Table 4. For every baseline, the instance-level label at timestep t is the argmax over classes of a per-timestep class score, as for the proposed model. **MI-Net** applies an instance classifier to every timestep and averages the resulting logits into the bag logit; the instance logits serve as instance scores. **Additive MIL** scales each timestep's representation by its attention weight before the instance classifier and averages the resulting logits; these per-timestep logits serve as instance scores. **MILLET** classifies unscaled representations and weights the logits by attention during pooling; we use the logits prior to this weighting. **ABMIL** computes one attention weight a_t per timestep and applies a linear classifier (w_c, b_c) to the average of the attention-scaled representations. Its bag logit for class c therefore decomposes over time as $\frac{1}{T}\sum_t (a_t\, w_c^\top \tilde{z}_t + b_c)$, where $\tilde{z}_t$ is the representation after positional encoding, and we use the per-timestep term $a_t\, w_c^\top \tilde{z}_t + b_c$ as its instance score. The attention weight alone would assign every class the same score. **TimeMIL** uses its released head, a class token followed by wavelet positional encoding and two Nyström transformer layers, trained with a 20-epoch warm-up; instance scores are obtained by applying its classification head to the output instance tokens.

Table 6: Additional comparison with WS-TAL methods adapted to time series (Appendix F). StepMIL is repeated from Table 1. Every cell is the mean ± standard deviation, aggregated as in Table 1. **Bold** marks the best value in a column and underline the second-best. ISSF is not evaluated on DOD-H, HMC and MIT-BIH (—), where the cost of its quadratic attention is prohibitive.

Method	OPPORTUNITY				DOD-H				HARTH			
	bag mAP	fg acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ
WS-TAL, adapted to time series												
W-TALC 2018	.487 ±.121	.287 ±.081	.142 ±.063	.223 ±.080	.888 ±.005	.417 ±.041	.286 ±.022	.144 ±.030	.765 ±.012	.820 ±.015	.491 ±.029	.756 ±.018
BaS-Net 2020	.417 ±.058	.295 ±.079	.161 ±.061	.239 ±.081	.850 ±.011	.498 ±.000	.133 ±.000	.000 ±.000	.708 ±.021	.557 ±.034	.289 ±.023	.336 ±.066
UM 2021	.415 ±.071	.158 ±.053	.050 ±.013	.084 ±.038	.850 ±.007	.245 ±.008	.197 ±.011	.050 ±.004	.651 ±.013	.324 ±.049	.204 ±.013	.210 ±.041
ACM-Net 2021	.440 ±.067	.150 ±.045	.052 ±.005	.076 ±.053	.759 ±.012	.368 ±.158	.104 ±.035	.000 ±.000	.635 ±.084	.797 ±.048	.411 ±.095	.721 ±.071
ISSF 2024	.519 ±.062	.211 ±.060	.099 ±.027	.144 ±.062	—	—	—	—	.677 ±.041	.775 ±.053	.458 ±.012	.701 ±.062
StepMIL (Ours)	**.707** ±.084	**.349** ±.075	**.213** ±.055	**.297** ±.071	**.924** ±.006	**.573** ±.013	**.368** ±.011	**.318** ±.014	**.852** ±.072	**.899** ±.010	**.611** ±.026	**.856** ±.032

Method	HAPT				HMC				MIT-BIH			
	bag mAP	fg acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ	bag mAP	inst acc	inst mF1	κ
WS-TAL, adapted to time series												
W-TALC 2018	.901 ±.019	.639 ±.090	.494 ±.059	.578 ±.105	.869 ±.006	.313 ±.029	.288 ±.031	.109 ±.030	.493 ±.008	.824 ±.025	.327 ±.009	.191 ±.018
BaS-Net 2020	.881 ±.012	.606 ±.041	.323 ±.029	.539 ±.049	.860 ±.015	.342 ±.000	.102 ±.000	.000 ±.000	.436 ±.039	**.883** ±.004	.265 ±.007	.047 ±.012
UM 2021	.885 ±.015	.256 ±.043	.187 ±.018	.164 ±.047	.875 ±.006	.249 ±.032	.211 ±.010	.060 ±.008	.451 ±.033	.177 ±.002	.135 ±.005	.013 ±.005
ACM-Net 2021	.891 ±.017	.694 ±.016	.342 ±.039	.643 ±.018	.734 ±.007	.164 ±.032	.056 ±.010	.000 ±.000	.448 ±.009	.590 ±.441	.219 ±.116	.061 ±.047
ISSF 2024	.916 ±.006	.717 ±.007	.481 ±.031	.672 ±.008	—	—	—	—	—	—	—	—
StepMIL (Ours)	**.955** ±.010	**.880** ±.016	**.655** ±.022	**.860** ±.019	**.916** ±.024	**.494** ±.033	**.449** ±.079	**.327** ±.057	**.497** ±.034	.864 ±.045	**.362** ±.023	**.269** ±.020

F ADDITIONAL COMPARISON WITH WS-TAL METHODS

Weakly supervised temporal action localization (WS-TAL) methods localize actions in untrimmed videos from video-level labels and operate on snippet features of pretrained video backbones. Applying them to time series requires modifications that the published methods do not include. We replace the video features by the features of our shared encoder, reduce the width of their hidden layers, and read per-timestep class scores from their class activation sequences (Appendix D) instead of the temporal proposals they natively output. Because these modifications depart from the published methods, we report this comparison as an additional experiment rather than as a main baseline. Table 6 reports the results. Among these five methods and StepMIL, StepMIL attains the best value in 23 of the 24 columns and an average rank of 1.04 against 2.54 for the strongest WS-TAL method (W-TALC). On MIT-BIH, W-TALC attains an instance macro-F1 of .327, above the best MIL method for bag classification (.269, MI-Net), and StepMIL attains .362.

Setup. W-TALC, BaS-Net, UM, ACM-Net and ISSF are ported from the authors' public code and keep their released layers, losses, top-k ratios and loss weights. Only the input changes, from I3D snippet features ($D = 2048$) to the features of the shared encoder ($D = 128$), and the 2048-wide hidden layers of the heads are reduced to 512 so that the longest bags fit in memory. Each method is applied at every sampling time, and its class activation sequence over the C foreground classes serves as instance scores. BaS-Net, ACM-Net and ISSF also model an explicit background class, whose decision is used on datasets with a background label. BaS-Net trains this class in two branches with opposite targets. Its suppression branch, used at test time, is trained to give background zero probability, so on these datasets we read both the class and the background decision of BaS-Net from its base branch, where background is a target in every bag. These methods keep their released optimizers (Adam with learning rate 10^{-4}, or 5×10^{-5} for ISSF, and their released weight decay), because under the optimizer of Table 4 BaS-Net stops improving and UM diverges. The training budget, batch size, splits, seeds and checkpoint selection are those of all other methods. ISSF forms an attention matrix between every timestep and its salient snippets, whose cost grows quadratically with the bag length and is prohibitive on DOD-H, HMC and MIT-BIH. ISSF is therefore not evaluated on these three datasets.

Behavior of the WS-TAL heads. These heads were designed for untrimmed videos in which a few action instances are separated by background. Their video-level objectives normalize the label over the classes present and fit a softmax over classes to the pooled activations, a target that class scores varying little over time can meet. Such scores assign every timestep of a bag the same class, which is correct only for single-class bags. The failures therefore concentrate on OPPORTUNITY,

DOD-H and HMC, whose bags contain 4.8, 3.1 and 3.2 classes on average, against 1.4 to 2.3 on the other datasets. **BaS-Net** predicts a single class at every timestep on DOD-H and HMC ($\kappa = 0$). **ACM-Net** leaves 13 of 17 classes unpredicted on OPPORTUNITY, 3 of 5 on DOD-H and 4 of 5 on HMC, and predicts background at 97% of OPPORTUNITY and 80% of HAPT timesteps, whose true background rates are 72% and 24%. **UM** treats the $T/4$ timesteps with the smallest feature magnitude as background and trains their class distribution toward uniform. On the four datasets without a background label, these timesteps belong to real classes, and UM's κ stays at or below .21 on every dataset. **W-TALC** has no background model and adds a co-activity loss between bags that share a class. It attains the highest instance macro-F1 among the WS-TAL heads on five of the six datasets, and BaS-Net does on OPPORTUNITY. **ISSF** ranks second on HARTH and HAPT and third on OPPORTUNITY.

G RELATED WEAKLY SUPERVISED SETTINGS

Weakly labeled sound event detection. Weakly labeled sound event detection (SED) also learns frame-level predictions from clip-level multi-labels (McFee et al., 2018; Wang et al., 2019; Turpault et al., 2019). SED models aggregate frame-level probabilities with MIL pooling functions such as max, average, linear softmax, softmax and attention pooling (Wang et al., 2019), and the class-wise softmax pooling of Table 2 belongs to this family. Complete SED systems also contain components designed for audio, such as spectrogram front-ends and event-level post-processing, which we do not evaluate.

Set-supervised action segmentation. Set-supervised action segmentation learns frame-level action labels from the unordered set of actions in each video (Richard et al., 2018; Li & Todorovic, 2020; Fayyaz & Gall, 2020), the same form of supervision as our bag labels. These methods were developed for procedural video represented by features of pretrained video models, whereas we predict a class at every sampling time of sensor and physiological signals. We do not evaluate them in this study, and a comparison under our protocol is left for future work.

H ABLATION DETAILS

Both ablations keep every setting of the main experiments and report the plain average over the six datasets of the per-dataset means in Table 1. The accuracy column is fg acc on OPPORTUNITY and HAPT and inst acc elsewhere.

Bag pooling (Table 2). LSTP with both gates and $\lambda_{\mathrm{LSTP}} = 0.5$ is fixed, and only the operator that aggregates the class scores $s_{t,c}$ into $\hat{s}_c$ changes. Shared conjunctive pooling computes $\hat{s}_c = \mathrm{mean}_t(a_t s_{t,c})$ with one sigmoid gate $a_t = \sigma(\mathrm{MLP}(z_t))$ shared by all classes, as in MILLET (Early et al., 2024). Class-wise conjunctive pooling replaces this gate by a per-class gate $a_{t,c} = \sigma(U z_t)$. Class-wise softmax pooling computes $\hat{s}_c = \sum_t \mathrm{softmax}_t(s_{t,c})\, s_{t,c}$. Class-wise sparse pooling is Eq. 2.

Components (Table 3). The first row trains class-wise sparse temporal pooling with the bag loss alone. LSTP without gates sets $r_j = e_{j,t} = 1$, so each target averages the soft distributions of a timestep and its two neighbors, while the loss retains the target-confidence weight $\omega_t = \mu_t \bar{r}_t$ (Eq. 11). The row without neighbors enables the gate r but sets $\mathcal{N}_t = \{t\}$, so each timestep is supervised only by its own bag-constrained target. The next two rows enable only the source reliability gate r or only the edge affinity gate e, and the last row enables both.

I SINGLE- AND MULTI-CLASS BAGS

Table 7 splits the test bags of each dataset by the number m of foreground classes they contain and recomputes the instance-level metrics of Table 1 on each part, with the same evaluation code and all trained models. In a bag with $m = 1$, the bag-constrained target of LSTP is one-hot on the bag class, whereas in a bag with $m \geq 2$ it has to separate co-occurring classes. Single-class bags account for 5% (HMC) to 62% (MIT-BIH) of the test bags. On bags with $m \geq 2$, StepMIL attains

the highest macro-F1 on four datasets, ties on HARTH, and attains the highest κ on all five, so the gains in Table 1 do not come from single-class bags alone. Bags that contain only background are excluded from both parts. DOD-H is omitted, since 96% of its bags contain at least two sleep stages and the split would leave its results essentially unchanged.

Table 7: Instance-level performance on single-class ($m = 1$) and multi-class ($m \geq 2$) test bags. "Base" is the highest value among MI-Net, ABMIL, Additive MIL, MILLET and TimeMIL in each cell, and **bold** marks the higher of the two. κ is the foreground variant on OPPORTUNITY and HAPT.

		mF1, $m = 1$		mF1, $m \geq 2$		κ, $m \geq 2$	
Dataset	$m = 1$ bags	StepMIL	Base	StepMIL	Base	StepMIL	Base
OPPORTUNITY	8%	**.079**	.049	**.216**	.127	**.300**	.246
HARTH	55%	**.430**	.424	**.604**	**.604**	**.734**	.724
HAPT	21%	**.245**	.176	**.643**	.572	**.837**	.762
HMC	5%	**.331**	.326	**.443**	.372	**.315**	.234
MIT-BIH	62%	**.241**	.195	**.389**	.324	**.337**	.134

J ADDITIONAL QUALITATIVE RESULTS

Figure 6 extends Figure 4 to five datasets, including sleep staging on HMC, where predictions at the resolution of individual samples remain partly fragmented for every method.

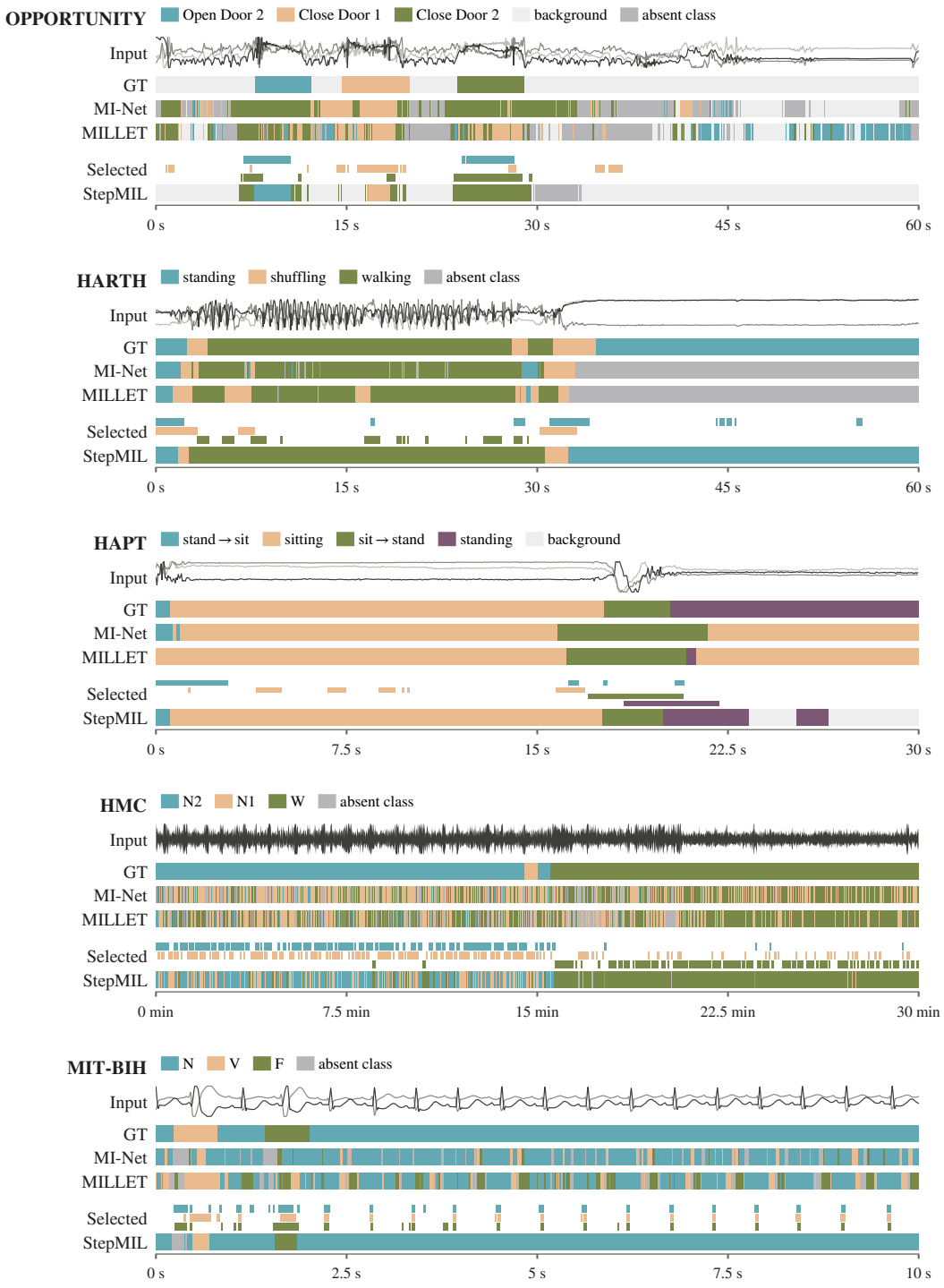

Figure 6: Instance predictions on test bags from five datasets. The input row overlays the forearm accelerometer axes (OPPORTUNITY), the lower-back accelerometer axes (HARTH), the smartphone accelerometer axes (HAPT), two EEG derivations (HMC) and both ECG leads (MIT-BIH). Long signals are drawn as min–max envelopes. Background marks timesteps assigned to no foreground class, and gray marks predicted classes that do not occur in the bag.