PREFIX-ADAPTIVE AND UTILITY-AWARE STOPPING FOR EEG CLOSED-LOOP RESOURCE-EFFICIENT SLEEP STAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous EEG acquisition dominates the resource budget of wearable sleep staging. Efficiency-oriented models fully observe a time period for decision, which is difficult to apply to resource-constrained actual deployment systems. Early time-series classification (ECTS) would stop once the observed prefix is decisive, but its instance-wise formulation assumes a stopping decision affects only the instance that made it. However, consecutive epochs are dependent, so stopping early truncates the history later epochs rely on, and a model trained on complete histories meets policy-induced covariate shift at inference. We introduce Prefix-Adaptive and Utility-Aware Stopping (PAUSE), a closed-loop framework in which each stopping decision controls both the current acquisition and the memory passed forward. A utility-aware finite-horizon stopper weighs stopping risk against expected continuation risk plus acquisition cost, a causal endpoint-memory backbone conditions each epoch on summaries taken at the policy-determined stopping points of its predecessors, and paired-history training combines complete and randomly truncated histories to reduce sensitivity to policy-induced history truncation. On four public datasets validation, PAUSE raises the macro-F1 versus acquisition frontier by 11.4 to 13.7 points over state-of-the-art ECTS baselines with 5.5 to 7.4 times less EEG. These results demonstrate the effectiveness of PAUSE for evidence-driven EEG acquisition in resource-constrained sleep monitoring.

1 INTRODUCTION

Sleep is closely related to physical and mental health (Scott et al., 2021; Fernandez-Mendoza & Vgontzas, 2013; Irwin, 2019). Sleep staging is used to assess sleep quality and diagnose sleep disorders by dividing electroencephalogram (EEG) recordings into 30-second segments, called sleep epochs, and assigning each epoch to one of five standard sleep stages (Boostani et al., 2017; Memar & Faradji, 2018; Berry et al., 2017). Recent studies in flexible wearable systems combine fabric sensors and miniature amplifiers for less intrusive EEG acquisition with neural networks running on edge servers for automatic sleep staging (Qiu et al., 2024; Chen et al., 2022b;a). As these systems move toward continuous use, reducing sensing requirements has become increasingly important. Existing resource-efficient sleep-staging models primarily reduce computation and memory while retaining complete, fixed-duration EEG collection (Supratak & Guo, 2020; Yang et al., 2023; Liu et al., 2023; Li & Gao, 2023; Ai et al., 2025; Kavoosi et al., 2023). Partial-observation approaches such as MASS instead target sensing efficiency by classifying from only part of the recorded EEG signal (Ai et al., 2026; Follin et al., 2026). Although this reduces acquisition costs, its global prompt and bidirectional inter-epoch encoder jointly process a complete 32-epoch window, so predictions for earlier epochs can depend on later observations. MASS therefore targets fixed-budget, window-level partial observation rather than causal per-epoch stopping. A fixed observation ratio also cannot adapt to epoch difficulty, motivating online acquisition control for latency-sensitive applications (Tăuțan et al., 2022; Esfahani et al., 2023; Nguyen et al., 2023).

These limitations motivate a shift from fixed-budget inference to online acquisition control, in which the classifier repeatedly decides whether the acquired prefix provides sufficient evidence to stop. Early classification of time series (ECTS) provides a natural decision framework using sequen-

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.

tially observed prefixes (Gupta et al., 2020; Renault et al., 2024). Representative ECTS methods implement this trade-off through stochastic halting, confidence-based stopping, or finite-horizon posterior-risk estimation (Hartvigsen et al., 2019; Schäfer & Leser, 2020; Lv et al., 2023; Miyagawa & Ebihara, 2021; Ebihara et al., 2025). Although standard ECTS can model within-series temporal dependencies, the stopping decision of an instance generally does not determine the input visible to subsequent instances (Renault et al., 2024). In contrast, physiological and clinical evidence shows that sleep stages exhibit strong temporal dependencies across consecutive epochs (Zhao et al., 2022; Phan et al., 2023). Ignoring this intrinsic temporal organization can substantially reduce staging accuracy. Simply adding historical context in ECTS does not solve the problem. Since the model is trained with fully observed historical epochs, its training inputs differ from those available during adaptive inference, where previously stopped epochs are only partially observed. Adaptive sleep staging therefore requires a stopping framework that exploits cross-epoch context while remaining compatible with the truncated histories produced by adaptive stopping.

To address this challenge, we propose Prefix-Adaptive and Utility-Aware Stopping for EEG-Based Resource-Efficient Sleep Staging (PAUSE), a closed-loop framework in which each stopping decision controls both current EEG acquisition and the endpoint memory available to subsequent epochs. Once the stopping risk no longer exceeds the continuation risk, the remaining EEG suffix is not acquired and the selected endpoint is retained as context for subsequent epochs. PAUSE utilizes a **utility-aware finite-horizon stopper** to compare immediate classification risk with expected continuation risk plus the cost of acquiring another patch, adaptively allocating the acquisition budget according to epoch difficulty. Its **causal endpoint-memory backbone** conditions each epoch on summaries taken at the policy-determined stopping positions of preceding epochs, thereby capturing sufficient evidence with cross-epoch sleep-stage continuity. Finally, the **paired-history training strategy** trains PAUSE on full and randomly truncated historical endpoints for the same sequence, broadening the range of historical contexts represented during training and reducing sensitivity to policy-induced shifts between training and inference.

The main contributions of the present work are as follows: (1) We formulate resource-efficient sleep staging as a **closed-loop adaptive acquisition** problem, where prefix-wise classification evidence controls how much EEG is acquired for each epoch, advancing the task from fixed-budget partial observation to evidence-driven sensing. (2) We extend ECTS to cross-epoch sleep staging. A causal endpoint-memory backbone and paired-history training exploit preceding epoch information while mitigating the mismatch between training and policy-truncated inference. (3) Experiments on four public sleep datasets demonstrate that PAUSE improves the macro-F1–acquisition frontier over fixed-prefix acquisition and SOTA ECTS systems, while ablations demonstrate the benefit of endpoint-memory backbone and paired-history training for closed-loop inference.

2 Problem Formulation

Let a sleep recording consist of N consecutive epochs $\{(X_n, y_n)\}_{n=1}^{N}$, where $X_n = (x_{n,1}, \ldots, x_{n,P})$ contains P sequential EEG patches and $y_n \in \{1, \ldots, C\}$ is the corresponding sleep-stage label. Acquisition proceeds sequentially within each epoch. At prefix k of epoch n, a system may use only the signals that have actually been acquired:

$$\mathcal{O}_{n,k} = \left(X_{1,1:\tau_1}, \ldots, X_{n-1,1:\tau_{n-1}}, X_{n,1:k}\right), \tag{1}$$

where $\tau_m \in \{1, \ldots, P\}$ is the acquisition length selected for each preceding epoch $m < n$. Based on $\mathcal{O}_{n,k}$, for $k < P$, the system either acquires the next patch $x_{n,k+1}$ or stops and outputs $\widehat{y}_{n,\tau_n}$; at $k = P$, it must stop. Once acquisition stops at τ_n, the remaining signal $X_{n,\tau_n+1:P}$ is not acquired and is therefore unavailable to both the current prediction and all subsequent decisions. The task is to learn an acquisition and classification policy that achieves a favorable trade-off between predictive performance and the expected amount of acquired signal $\mathbb{E}[\tau]$.

3 Methods

3.1 Framework Overview

PAUSE instantiates this task within fixed-length windows of $L = 32$ consecutive epochs. In the following, $n, m \in \{1, \ldots, L\}$ index epochs within a window, and endpoint memory is confined

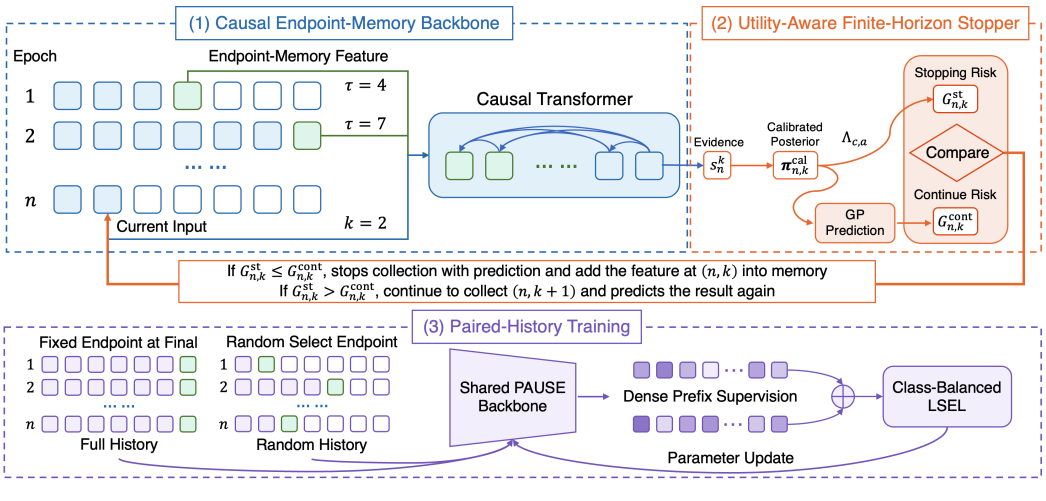


Figure 1: PAUSE overview. Boxes denote 1-s EEG patches; green boxes mark stopping endpoints. (1) The causal backbone combines the acquired prefix with earlier endpoint summaries. (2) The stopper compares classification risk with expected continuation risk plus sensing cost; a stopped endpoint informs later epochs. (3) Full and randomly truncated histories train the backbone with class-balanced LSEL at every prefix.

to that window. As illustrated in Fig. 1, PAUSE realizes this acquisition policy in inference by compressing each previously acquired prefix into an endpoint summary. A causal endpoint-memory backbone combines these historical summaries with the currently acquired prefix to produce prefix-wise class evidence. Per-prefix calibration converts the evidence into a class posterior, on which a utility-aware finite-horizon stopper compares the risk of stopping with the expected risk and cost of continuing. Once the stopper selects τ_n, PAUSE outputs $\widehat{y}_{n,\tau_n}$ and retains $\mathbf{h}_{n,\tau_n}$ as the historical representation available to subsequent epochs.

3.2 Causal Endpoint-Memory Backbone

Causal Endpoint-Memory Backbone first transforms the EEG signals into the frequency domain by computing power spectral density (PSD), producing clearer spectral features. Then, a patch embedding followed by a left-causal convolutional stem is used to map each acquired patch to a d-dimensional token $\mathbf{z}_{n,k} \in \mathbb{R}^d$ to provide local temporal context without violating prefix causality. Since the utility-aware finite-horizon stopper acts on a decision state that must summarize all evidence acquired so far, this imposes two requirements on the backbone. First, it must emit class evidence at every prefix length k with an available class posterior. Second, the evidence at prefix k must summarize the entire acquired prefix rather than a fixed-order window. Therefore, we adopt a causal Transformer with two-dimensional rotary embeddings, which inject the relative positional information into the query and key vectors along the patch and epoch axes separately before the attention scores are computed. From each token $\mathbf{z}_{n,k}$, per-head queries, keys, and values are obtained by learned linear projections. Let $d_h = d/H$ denote the dimension of one attention head, and H is the number of heads. For head a, the query and key vectors are divided into two equal subspaces:

$$\widetilde{\mathbf{q}}^a_{n,k} = \big[\mathcal{R}_p(k;\beta)\mathbf{q}^a_{n,k,p} \,\|\, \mathcal{R}_e(n;\beta)\mathbf{q}^a_{n,k,e}\big], \quad \widetilde{\mathbf{k}}^a_{m,j} = \big[\mathcal{R}_p(j;\beta)\mathbf{k}^a_{m,j,p} \,\|\, \mathcal{R}_e(m;\beta)\mathbf{k}^a_{m,j,e}\big]. \quad (2)$$

where $\mathcal{R}_p$ and $\mathcal{R}_e$ are pairwise rotary transformations with base β for the patch and epoch coordinates, respectively, and $\|$ denotes concatenation. To ensure causality, the Transformer utilizes the following attention mask at position (n,k):

$$\mathcal{A}\big((n,k),(m,j)\big) = \begin{cases} 1, & m = n \text{ and } j \le k, \\ 1, & m < n \text{ and } j = \tau_m, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

A token can therefore attend to its local prefix, and to endpoints from each preceding epoch only. In this process, the hidden state $\mathbf{h}_{m,\tau_m}$ summarizes the acquired prefix $X_{m,1:\tau_m}$ together with the

endpoint context of earlier epochs, without accessing the unacquired suffix $X_{m,\tau_m+1:P}$. PAUSE retains this state as a compact historical summary. This enables PAUSE to model features across sleep epochs within the framework of ECTS, providing compact cross-epoch context which is important for sleep staging.

During training, PAUSE uses candidate key–value banks to evaluate all causal prefixes of epoch n in parallel; at deployment, the same causal model operates incrementally as each patch is acquired. The candidate banks contain P rotated local keys and their unrotated values, together with one rotated endpoint key and its unrotated value from each strictly preceding epoch:

$$\overline{\mathbf{K}}^a_n = \left[\widetilde{\mathbf{k}}^a_{n,1:P}; \widetilde{\mathbf{k}}^a_{1,\tau_1}; \ldots; \widetilde{\mathbf{k}}^a_{n-1,\tau_{n-1}}\right], \quad \overline{\mathbf{V}}^a_n = \left[\mathbf{v}^a_{n,1:P}; \mathbf{v}^a_{1,\tau_1}; \ldots; \mathbf{v}^a_{n-1,\tau_{n-1}}\right]. \tag{4}$$

so that $\overline{\mathbf{K}}^a_n$ has $P + n - 1$ rows. The endpoint of the current epoch is excluded by construction: at query time τ_n is not yet determined, and once determined it is already contained in the local prefix. The output for query position (n, k) is then computed as

$$\mathbf{o}^a_{n,k} = \mathrm{softmax}\left(\frac{\widetilde{\mathbf{q}}^a_{n,k}\overline{\mathbf{K}}^{a\top}_n}{\sqrt{d_h}} + \mathbf{B}_{n,k}\right)\overline{\mathbf{V}}^a_n, \tag{5}$$

where $\mathbf{B}_{n,k}$ is zero on entries permitted by Eq. (3) and $-\infty$ otherwise. Thus, a query at (n, k) attends only to the first k local tokens and the preceding endpoints; local tokens with $j > k$ receive zero attention weight. This parallel computation is causally equivalent to processing the acquired prefix alone. After that, a shared head is utilized to obtain the class evidence at every prefix position: $\mathbf{s}_{n,k} = \mathbf{W}_s\mathbf{h}_{n,k} + \mathbf{b}_s, \mathbf{s}_{n,k} \in \mathbb{R}^C$. Thus, $\mathbf{s}_{n,k}$ contains one evidence score for each sleep stage.

3.3 Utility-Aware Finite-Horizon Stopper

PAUSE utilizes the Utility-Aware Finite-Horizon Stopper to convert the prefix-wise evidence into a decision state before assigning a sensing budget, thereby achieving resource-efficient sleep staging. Given the trained backbone, we freeze it and fit one positive temperature T_k for each k to prevent deep neural networks from being overconfident. By training to minimize the mean negative log-likelihood, the resulting temperature can help construct the calibrated posterior: $\boldsymbol{\pi}^{\mathrm{cal}}_{n,k} = \mathrm{softmax}\left(\frac{\mathbf{s}_{n,k}}{T_k} + \log\hat{\boldsymbol{\pi}}\right)$, where $T_k > 0$ is the temperature fitted for prefix length k, M_c is the number of training epochs of class c, $M = \sum_c M_c$, $\hat{\boldsymbol{\pi}}$ is the empirical class prior with $\hat{\pi}_c = M_c/M$, and the logarithm is applied element-wise. The temperature scales only the prior-free evidence $\mathbf{s}_{n,k}$, while the class prior is introduced once outside the temperature scaling.

Due to the severe class imbalance in sleep staging, we encode the utility of a prediction by a misclassification-cost matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{C\times C}_+$, whose entry $\Lambda_{c,a}$ is the cost of predicting class a when the true class is c. To prevent frequent stages from dominating the stopping policy, the off-diagonal costs are weighted using the same empirical class prior $\hat{\boldsymbol{\pi}}$:

$$\widetilde{w}_c = \left(\frac{M}{CM_c}\right)^{\alpha}, w_c = \frac{\widetilde{w}_c}{\sum_{r=1}^{C}\hat{\pi}_r\widetilde{w}_r}, \Lambda_{c,a} = \rho w_c \mathbb{I}[c \neq a], \tag{6}$$

where ρ controls the overall penalty for an incorrect decision, and α controls the degree of class balancing. Given the calibrated posterior, the risk of stopping at prefix k and its associated prediction are

$$G^{\mathrm{st}}_{n,k} = \min_{a\in\{1,\ldots,C\}}\sum_{c=1}^{C}\pi^{\mathrm{cal}}_{n,k}(c)\Lambda_{c,a}, \qquad \widehat{y}_{n,k} = \arg\min_{a\in\{1,\ldots,C\}}\sum_{c=1}^{C}\pi^{\mathrm{cal}}_{n,k}(c)\Lambda_{c,a}. \tag{7}$$

This formulation ties the immediate stopping risk to stage-specific misclassification costs, rather than collapsing the decision onto a single confidence threshold. A low stopping risk alone, however, does not justify stopping: what matters is how much further the classification risk could be reduced by continuing to collect data, and whether that reduction outweighs the additional acquisition cost. Because each epoch spans a finite horizon of P patches, the value of acquiring another patch depends on how much continuation risk remains reducible. Let $V_{n,k}$ denote the finite-horizon cost-to-go from prefix k onward, and let $\eta > 0$ be the cost of acquiring one additional patch. The corresponding stop-versus-continue trade-off is expressed by the Bellman recursion:

$$V_{n,P} = G^{\mathrm{st}}_{n,P}, \qquad V_{n,k} = \min\left\{G^{\mathrm{st}}_{n,k},\ \eta + \mathbb{E}\left[V_{n,k+1} \mid \boldsymbol{\pi}^{\mathrm{cal}}_{n,k}\right]\right\}. \tag{8}$$

The second relation is evaluated for $k = P-1, \ldots, 1$. Its conditional expectation is not available analytically. Following the finite-horizon regression strategy of FIRMBOUND (Ebihara et al., 2025), we estimate it with a time-indexed sparse Gaussian process (GP). Starting from the terminal target $V_{n,P} = G^{\mathrm{st}}_{n,P}$, the GP at step k regresses the recursively computed target $V_{n,k+1}$ on the current calibrated posterior $\boldsymbol{\pi}^{\mathrm{cal}}_{n,k}$. Its predictive mean $\mu_k(\boldsymbol{\pi}^{\mathrm{cal}}_{n,k})$ approximates the expected continuation risk, yielding

$$G^{\mathrm{cont}}_{n,k} = \eta + \mu_k\big(\boldsymbol{\pi}^{\mathrm{cal}}_{n,k}\big)\,, \tau_n = \inf\left\{k:\ G^{\mathrm{st}}_{n,k} \le G^{\mathrm{cont}}_{n,k}\right\}. \tag{9}$$

Because the calibrated posterior is used as a compact decision state and the conditional expectation is learned by GP regression, the resulting stopper is a fitted approximation to the finite-horizon value function rather than an exact Bellman-optimal solution. With a forced stop at $k = P$ if the inequality is never met earlier, the first prefix satisfying Eq. (9) determines both the output $\widehat{y}_{n,\tau_n}$ and the endpoint state $\mathbf{h}_{n,\tau_n}$ retained for subsequent epochs. By varying η, we can trace the Speed–Accuracy Tradeoff (SAT): a larger acquisition cost favors earlier decisions, whereas a smaller cost permits additional evidence for ambiguous epochs.

3.4 Paired-History Training

In PAUSE, the historical representation available to epoch n is determined by the preceding endpoints $\{\tau_m\}_{m<n}$. At deployment, these endpoints are generated online by the stopper, but they are not yet available when the backbone is trained. Simply setting $\tau_m = P$ for all preceding epochs exposes the model only to complete histories, whereas deployment presents policy-truncated histories. This discrepancy creates the policy-induced covariate shift described in the introduction. We address this shift by training under a distribution of endpoint configurations. The Paired-History Training strategy constructs two history views from every training window. Let $\boldsymbol{\tau}^{\mathrm{F}}$ and $\boldsymbol{\tau}^{\mathrm{R}}$ denote the endpoint configurations of the full-history and random-history views, respectively. They are defined as

$$\tau^{\mathrm{F}}_{b,m} = P, \qquad \tau^{\mathrm{R}}_{b,m} \sim \mathrm{Unif}\{1, \ldots, P\}, \tag{10}$$

where b indexes a training window, and m indexes an epoch within the window. The random endpoints are sampled independently. For a prediction at epoch n, only endpoints with $m < n$ are accessible under either view following Eq. (3). Therefore, the randomly sampled endpoint of an epoch controls how that epoch is presented as history to later epochs. However, it does not truncate the supervision of the epoch itself. In both views, PAUSE continues to predict the label of every epoch at all prefix positions $k = 1, \ldots, P$. The two views contain identical EEG windows and labels and differ only in the historical endpoints supplied to the backbone. We concatenate them along the minibatch dimension and optimize them in a single update. For a history view $v \in \{\mathrm{F}, \mathrm{R}\}$, we optimize the backbone with a class-balanced, prefix-wise Log-Sum-Exp Loss (LSEL). Let $\mathbf{s}$ be the class-score vector at a single prefix, $s(r)$ its entry for class r, and c the ground-truth class. The instance-wise LSEL is

$$\ell_{\mathrm{LSEL}}(\mathbf{s}, c) = \log\left(1 + \sum_{r \neq c} \exp\big(s(r) - s(c)\big)\right), \tag{11}$$

which drives the true-class score $s(c)$ above every competitor and equals the negative log-softmax probability of class c. We apply it at every prefix of every epoch under class balancing. Writing $\mathbf{s}^{v}_{b,n,k}$ for the score vector at prefix k of epoch n in window b under view v and $y_{b,n}$ for its label, the complete loss for view v is $\mathcal{L}_v = \frac{1}{C}\sum_{c=1}^{C} \frac{1}{M_c} \sum_{(b,n):\, y_{b,n}=c} \frac{1}{P} \sum_{k=1}^{P} \ell_{\mathrm{LSEL}}\left(\mathbf{s}^{v}_{b,n,k}, c\right)$, where the inner average runs over prefixes and the outer over classes, so majority stages do not dominate.

The final training objective is $\mathcal{L}_{\mathrm{paired}} = \frac{1}{2}\mathcal{L}_{\mathrm{F}} + \frac{1}{2}\mathcal{L}_{\mathrm{R}}$. Based on the previous LSEL, Paired-History Training is intended to reduce sensitivity to policy-induced history shift by exposing the backbone to both complete and random truncated histories.

4 Experiments

4.1 Datasets and Settings

We evaluated PAUSE performance on four different public sleep staging datasets: **I. DREAMS-SUB** contains 20 overnight sleep records from 20 healthy subjects aged from 20 to 65 years (De-

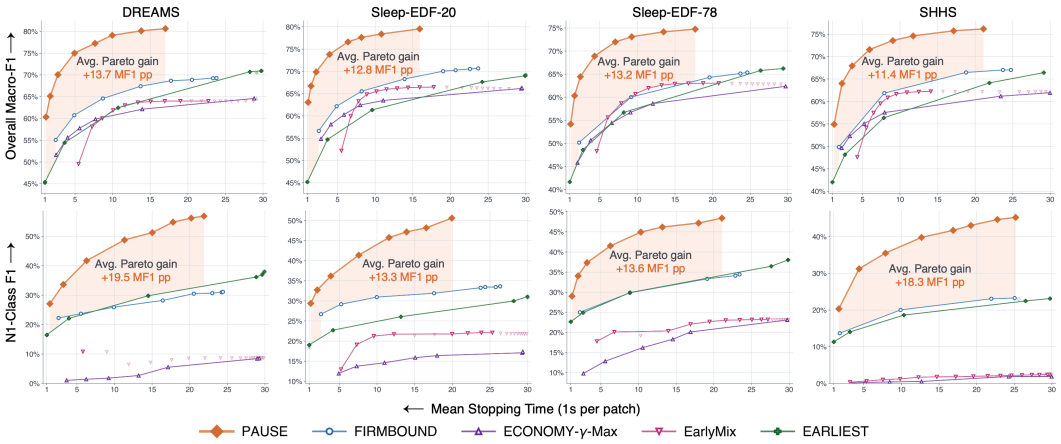


Figure 2: Speed–accuracy frontiers on the four datasets. The shaded region represents matched-budget Pareto gap. Top row: overall macro-F1. Bottom row: F1 of the minority N1 stage. Curves closer to the upper-left corner are better.

vuyst, 2005). **II. Sleep-EDF-20** is from PhysioBank, which contains 39 overnight sleep records from 20 subjects (Kemp et al., 2000). **III. Sleep-EDF-78** is also from PhysioBank, which contains 153 overnight sleep records from 78 subjects (Kemp et al., 2000). **IV. SHHS** contains 329 overnight sleep records from 329 selected subjects (Zhang et al., 2018; Quan et al., 1997; Li et al., 2022). To maintain consistency with previous work, we selected the same channel on these four datasets (Zhang et al., 2023; Eldele et al., 2021). We also performed subject-wise 20-fold cross-validation on all datasets. For each target fold, the per-prefix temperatures and GP stoppers were fitted exclusively from the OOF predictions of the remaining folds. We implemented PAUSE based on NVIDIA RTX 5090 GPU, PyTorch 2.8.0, Python 3.12, CUDA 12.8. Dataset preprocessing and the cross-validation/OOF protocol are detailed in Sections 2.2–2.3 of the Supplementary Material; model and stopper configurations, including hyperparameter search, are detailed in Sections 4.1–4.3.

4.2 System-Level Performance Comparison

We benchmark PAUSE against four representative early time-series classification methods. **I. EARLIEST** jointly trains an LSTM classifier and a reinforcement-learning controller to determine an adaptive stopping time for each sample (Hartvigsen et al., 2019). **II. ECONOMY-γ-Max** uses confidence groups and estimated future costs to make non-myopic stop-or-wait decisions (Zafar et al., 2021). **III. EarlyMix** employs hierarchical sample-level and latent-level mixing to improve classification from information-limited early prefixes (Hu et al., 2025). **IV. FIRMBOUND** formulates finite-horizon early classification as an optimal-stopping problem and learns the stopping boundary through backward induction over posterior risks (Ebihara et al., 2025).

We select baselines that make stop-or-wait decisions using only the currently observed prefix to align the comparison with causal online acquisition task. Each method exposes a single scalar that trades observation length against accuracy. Therefore, we sweep each scalar over its full range and record the corresponding mean stopping time $\mathbb{E}[\tau]$ together with macro-F1 and N1-class F1 under the identical 20-fold protocol. Fig. 2 plots the resulting speed–accuracy frontiers, and we additionally summarize the frontier-wide separation by the average Pareto gain. Each method is evaluated with its classifier and stopper matched to PAUSE in parameter count, so that the frontiers contrast complete acquisition systems at equal capacity rather than model scale.

PAUSE achieves a higher macro-F1–acquisition frontier than the evaluated ECTS systems across the overlapping budget ranges of all four datasets. The annotated average Pareto gain ranges from 11.4 to 13.7 macro-F1 percentage points. The strongest baselines peak at 66.22–70.94 macro-F1 using 23.7–29.8 patches. PAUSE exceeds these performance levels using only 4.9, 4.0, 4.3, and 3.6 patches, respectively, corresponding to 5.5–7.4$\times$ less acquired EEG. Thus, PAUSE discards 83%–88% of each epoch while remaining more accurate than the strongest baseline operating points. With larger acquisition budgets of only 8.17–9.95 patches, PAUSE reaches macro-F1 scores of 79.11,

Table 1: Ablation of PAUSE over the acquisition-cost sweep. Each cell reports macro-F1 (%) with the induced mean acquisition length $\mathbb{E}[\tau]$. Bold marks the best variant at a given cost level.

Dataset	Variant	Acquisition cost η (decreasing → larger acquisition budget)							
		0.5	0.233	0.133	0.067	0.033	0.02	0.0067	0.0005
DREAMS-SUB	**PAUSE (ours)**	**60.34**$_{1.1}$	**65.14**$_{1.7}$	**70.05**$_{2.7}$	**75.02**$_{4.9}$	**77.27**$_{7.7}$	**79.11**$_{9.9}$	**80.12**$_{13.8}$	**80.63**$_{17.0}$
	w/o Paired-History	50.65$_{1.1}$	55.29$_{1.4}$	60.30$_{2.2}$	67.33$_{5.0}$	72.44$_{10.6}$	75.84$_{13.7}$	74.58$_{17.6}$	75.38$_{19.7}$
	w/o Endpoint Memory	55.85$_{2.4}$	64.01$_{5.4}$	69.75$_{9.4}$	73.00$_{14.8}$	74.15$_{18.2}$	74.34$_{19.8}$	74.60$_{21.0}$	74.68$_{21.6}$
Sleep-EDF-20	**PAUSE (ours)**	**63.06**$_{1.1}$	**66.72**$_{1.5}$	**69.88**$_{2.2}$	**73.84**$_{3.9}$	**76.62**$_{6.4}$	**77.63**$_{8.2}$	**78.40**$_{10.8}$	**79.56**$_{15.9}$
	w/o Paired-History	57.69$_{1.1}$	61.54$_{1.4}$	63.69$_{1.9}$	69.14$_{3.6}$	73.24$_{6.9}$	75.42$_{9.8}$	76.41$_{14.1}$	77.41$_{17.2}$
	w/o Endpoint Memory	58.99$_{2.5}$	65.42$_{4.7}$	69.28$_{7.5}$	72.27$_{11.6}$	73.52$_{14.9}$	73.85$_{16.7}$	74.12$_{19.1}$	74.21$_{20.6}$
Sleep-EDF-78	**PAUSE (ours)**	**54.18**$_{1.1}$	**60.32**$_{1.6}$	**64.44**$_{2.4}$	**68.91**$_{4.3}$	**71.97**$_{7.1}$	**73.10**$_{9.2}$	**74.15**$_{13.5}$	**74.74**$_{17.7}$
	w/o Paired-History	50.92$_{1.1}$	58.28$_{1.7}$	62.18$_{2.4}$	66.41$_{4.3}$	69.50$_{7.2}$	70.79$_{9.3}$	72.18$_{14.2}$	72.81$_{17.3}$
	w/o Endpoint Memory	50.77$_{1.9}$	57.28$_{3.9}$	62.02$_{6.7}$	66.30$_{11.5}$	68.16$_{15.6}$	68.73$_{17.8}$	69.08$_{21.2}$	69.15$_{22.7}$
SHHS	**PAUSE (ours)**	**54.90**$_{1.2}$	**64.01**$_{2.3}$	**67.93**$_{3.6}$	**71.55**$_{5.9}$	**73.58**$_{9.0}$	**74.59**$_{11.7}$	**75.73**$_{17.1}$	**76.16**$_{21.1}$
	w/o Paired-History	50.23$_{1.2}$	59.68$_{2.1}$	64.95$_{3.5}$	69.40$_{6.1}$	71.64$_{9.2}$	72.81$_{11.6}$	74.08$_{15.9}$	74.83$_{19.8}$
	w/o Endpoint Memory	51.24$_{1.9}$	58.40$_{4.5}$	62.88$_{7.5}$	66.62$_{12.4}$	68.05$_{16.4}$	68.57$_{19.0}$	68.86$_{22.2}$	68.94$_{24.3}$

77.63, 73.10, and 73.58 on the four datasets, respectively, exceeding the best baseline systems at comparable budgets by 11.74–14.52 points. Even at its loosest setting, PAUSE uses at most 21.1 patches and remains 8.52–9.69 points more accurate than baselines using 23.7–29.8 patches. These results demonstrate the effectiveness of PAUSE as an integrated system for acquisition-efficient sleep staging, where adaptive stopping, policy-compatible cross-epoch context, and class-balanced evidence learning jointly determine the observed frontier.

The lower row of Fig. 2 further shows PAUSE's performance on difficult N1 class. N1 accounts for only 3.2% to 12.5% of epochs across the four datasets with the least separable features. Therefore, a cost-insensitive stopper may sacrifice its performance under tight budgets in exchange for better overall classification results. We observed this phenomenon in the EarlyMix and ECONOMY-γ-Max methods, which largely fail to recognize N1. On the contrary, PAUSE at its cheapest setting, using only 1.1–1.2 patches, already exceeds their best F1 scores at any budget. The corresponding average Pareto gain for N1 ranges from 13.3 to 19.5 F1 percentage points. PAUSE also matches the best F1 of EARLIEST and FIRMBOUND with only 2.2–4.3 patches and ultimately exceeds them by 10.41–21.96 points. These results show that PAUSE reduces signal acquisition without sacrificing the most difficult minority stage. Its improved macro-F1–acquisition frontier reflects balanced early classification rather than gains driven primarily by the majority stages.

4.3 Ablation Study

Table 1 isolates the two components that adapt PAUSE to cross-epoch staging. **w/o Paired-History** trains on the full-history view only, fixing $\tau_m = P$ in Eq. (10) instead of pairing it with the randomly truncated view. **w/o Endpoint Memory** further removes the endpoint keys and values of Eq. (4), so each epoch is classified from its own prefix alone, reducing PAUSE to a single-instance ECTS system on a matched backbone. Because a single stopping rule traces a frontier rather than a point, we sweep the acquisition cost η over eight levels and report both macro-F1 and the mean acquisition length $\mathbb{E}[\tau]$. It can be found that the two components fail in opposite directions as the budget grows, and each failure matches the mechanism it is meant to address.

Without endpoint memory, more of the current epoch does not close the performance gap. At the loosest cost setting, the memory-free variant uses between 20.6 and 24.3 patches but reaches only 68.94 to 74.68 macro-F1. PAUSE exceeds these scores after 4.9 to 7.1 patches, requiring three to fourfold fewer observations. At budgets matched to the operating points, this variant trails PAUSE by 5.8 to 12.5 macro-F1, accounting for most of the margin over the instance-wise baselines. Furthermore, the performance gap between the methods also increases as η decreases, from 3.41 to 4.49 points at $\eta = 0.5$ to 5.35 to 7.22 points at $\eta = 0.0005$. These results strongly suggest that preceding epochs can provide richer information about the current sleep stage that cannot be recovered simply by observing more of the current epoch. This is consistent with the results compared to the baseline of SOTA ECTS and also aligns with the conclusions of studies related to sleep (Phan et al., 2023; Zhao et al., 2022).

Removing paired-history training produces a different pattern. It has the largest performance penalty under aggressive truncation. At the lowest acquisition cost ($\eta = 0.0005$), PAUSE w/o Paired-

History can exceed PAUSE w/o Endpoint Memory by 0.70 to 5.89 macro-F1 points while using 1.9 to 5.4 fewer patches on average. However, as stopping becomes more aggressive, its gap to PAUSE widens from 1.33–5.25 points at $\eta = 0.0005$ to 3.26–9.69 points at $\eta = 0.5$; it also falls below w/o Endpoint Memory on three datasets at $\eta = 0.5$. This pattern is consistent with policy-induced history shift affecting the early stopping performance of cross-epoch ECTS. By introducing Paired-History Training, PAUSE improves consistently across the whole budget range on all four datasets. These results support that the Endpoint-Memory Backbone and Paired-History Training are useful for mitigating history shift and improving overall performance.

4.4 Disentangling the Backbone and the Stopper

PAUSE couples a causal endpoint-memory backbone with a utility-aware finite-horizon stopper, so the frontiers in Fig. 2 are traced by the two components jointly. We further examine them separately. We first disable the stopper and evaluate the backbone against full-signal sleep staging models, and then fix the backbone and replace only the stopping rule.

4.4.1 Sleep Staging Backbone.

Table 2: Comparison of overall performance (%) with state-of-the-art sleep staging models on four datasets.

Method	DREAMS-SUB			Sleep-EDF-20			Sleep-EDF-78			SHHS		
	ACC	MF1	MGm	ACC	MF1	MGm	ACC	MF1	MGm	ACC	MF1	MGm
AttnSleep	82.39	74.47	82.86	84.21	77.91	**86.20**	78.47	73.90	82.76	81.72	72.93	82.86
LGSleepNet	82.46	74.31	82.62	83.90	77.36	85.59	78.03	73.57	**82.81**	82.53	73.20	82.60
MSA-CNN	81.58	71.94	79.72	82.43	70.27	77.93	78.00	72.31	80.76	82.74	66.80	71.54
ULW-SleepNet	82.89	73.07	80.80	83.95	75.79	82.94	76.54	70.71	79.59	78.10	63.28	70.21
PAUSE (Ours)	**84.84**	**80.27**	**86.36**	**85.53**	**79.79**	85.97	**79.55**	**74.44**	82.13	**83.62**	**75.96**	**83.36**

We compare the backbone with four widely used automatic sleep staging models: **I. AttnSleep** (Eldele et al., 2021); **II. LGSleepNet** (Shen et al., 2023); **III. MSA-CNN** (Goerttler et al., 2026); and **IV. ULW-SleepNet** (Wang et al., 2026). All baselines were evaluated under the same subject-wise 20-fold protocol. For this comparison, the stopper in PAUSE is disabled and thus acquires all $P = 30$ patches. Table 2 shows that the PAUSE backbone is a strong and consistent full-observation sleep-staging model. It achieves the highest ACC and macro-F1 on all four datasets, exceeding the strongest baseline by 0.54 to 5.80 macro-F1 points. It also achieves the highest MGm on DREAMS-SUB and SHHS, while remaining within 0.23 and 0.68 points of the best MGm on Sleep-EDF-20 and Sleep-EDF-78, respectively. Overall, these results demonstrate competitive staging performance across all four datasets and establish a reliable classification foundation for adaptive acquisition. Detailed per-stage results are presented in Section 6.4 of the Supplementary Material.

4.4.2 Stopping Rule.

We freeze the PAUSE backbone and utilize different stoppers on all four datasets: **I. Fixed** stops at the same, fixed prefix length. **II. Threshold** applies a time-invariant threshold to the estimated log-likelihood ratios (Miyagawa & Ebihara, 2021). **III. AAGC** calculates a separate threshold for each prefix length (Ringel et al., 2024). **IV. FIRMBOUND** uses traditional finite-horizon Bellman recursion with GP regression (Ebihara et al., 2025). Each rule has its own trade-off parameter, and we sweep each parameter over its full range so that the rules are compared as frontiers. On a shared backbone and calibrated posterior, the gains can reflect the quality of complete decision policies.

Fig. 3 shows the overall macro-F1 frontiers on all four datasets. PAUSE remains above the joint upper envelope of baselines over their shared acquisition-budget range. Its average matched-budget Pareto gains are 2.21, 1.78, 3.55, and 3.92 macro-F1 points on DREAMS-SUB, Sleep-EDF-20, Sleep-EDF-78, and SHHS, respectively. The corresponding PAUSE peaks exceed the best alternative peaks by 0.71, 0.89, 1.72, and 2.68 points while using 29.1%–46.6% fewer patches. With the backbone, calibration, and evaluation folds fixed, these comparisons support the effectiveness of the complete utility-aware stopping policy. Full class-wise frontiers and matched-budget values are reported in Section 6.2 of the Supplementary Material.

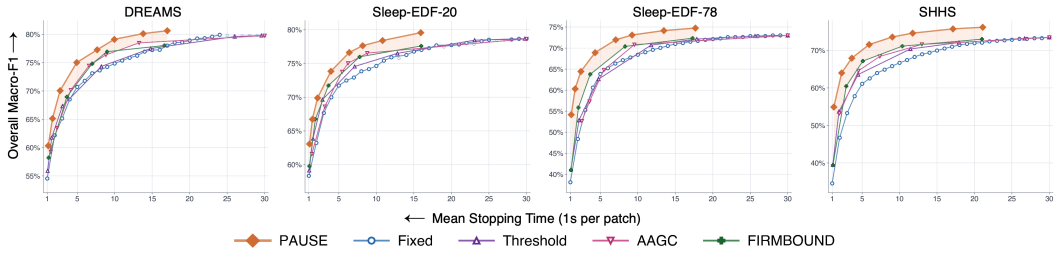


Figure 3: Overall macro-F1 versus mean stopping time for five stopping rules on DREAMS-SUB, Sleep-EDF-20, Sleep-EDF-78, and SHHS. Upper-left is better.

4.5 Estimated Front-End Power Comparison

Under conventional full acquisition, the analog front end (AFE) that amplifies and digitizes EEG stays active throughout recording, regardless of where inference runs. Following MASS (Ai et al., 2026), we estimate the AFE power of three amplifiers common in wearable EEG systems: **I. ADS1299-4**, **II. ADS131A04**, and **III. ADS1294** (Texas Instruments, 2012; 2016; 2010). The AFE runs in normal mode for the τ acquired patches and in standby for the rest of the epoch, giving $\overline{P}_{\mathrm{AFE}} = P_{\mathrm{S}} + (\mathbb{E}[\tau]/P)(P_{\mathrm{N}} - P_{\mathrm{S}})$ with datasheet powers $(P_{\mathrm{N}}, P_{\mathrm{S}}) = (22.0, 5.1)$, $(15.8, 2.6)$, and $(10.1, 4.0)$ mW, respectively. Because PAUSE acquires a contiguous prefix, each epoch incurs at most one standby entry and one wake-up, whose energy we neglect. As $\overline{P}_{\mathrm{AFE}}$ depends only on $\mathbb{E}[\tau]$, we compare at matched accuracy rather than budget: each ECTS baseline at its peak-macro-F1 operating point, and PAUSE, conservatively, at its first operating point exceeding all these peaks.

Table 3: Estimated AFE power (mW). I: ADS1299-4; II: ADS131A04; III: ADS1294.

Method	DREAMS-SUB			Sleep-EDF-20			Sleep-EDF-78			SHHS		
	I	II	III	I	II	III	I	II	III	I	II	III
EARLIEST	21.87	15.70	10.05	21.98	15.79	10.09	21.65	15.53	9.97	21.50	15.41	9.92
ECONOMY-γ-Max	21.33	15.28	9.86	21.71	15.58	10.00	21.83	15.66	10.04	21.97	15.78	10.09
EarlyMix	17.91	12.60	8.62	15.12	10.42	7.61	16.82	11.76	8.23	13.04	8.80	6.87
FIRMBOUND	18.51	13.08	8.84	18.44	13.02	8.81	18.96	13.43	9.00	19.04	13.49	9.03
PAUSE (Ours)	**7.88**	**4.77**	**5.00**	**7.32**	**4.34**	**4.80**	**7.52**	**4.49**	**4.87**	**7.14**	**4.19**	**4.74**

The baselines fall into two groups (Table 3). EARLIEST and ECONOMY-γ-Max peak only after 28.8 to nearly 30 patches, retaining 96.7%–99.9% of full-acquisition power, so their stopping yields almost no hardware saving at their best accuracy. FIRMBOUND and EarlyMix stop earlier, at 14.1–24.8 patches, yet still retain 55.7%–89.4%. PAUSE surpasses all of these peaks with only 3.6–4.9 patches, lowering the estimated AFE power relative to the baselines by 45.2%–67.5%, 52.3%–73.4%, and 31.0%–53.1% for amplifiers I–III, respectively. Standby power already accounts for 54.5%–84.5% of the remaining AFE power of PAUSE, most on ADS1294, where the saving is smallest; further gains therefore depend mainly on lower-standby hardware. These component-level estimates exclude processors, wireless transmission, storage, and power conversion.

5 Conclusion

This paper introduced PAUSE, a closed-loop framework that replaces fixed observation budgets with evidence-driven EEG acquisition. It couples utility-aware stopping with policy-compatible cross-epoch context, while paired-history training mitigates the resulting training–inference shift. Across four public datasets, PAUSE improved the macro-F1–acquisition frontier over the evaluated ECTS systems while acquiring 5.5–7.4$\times$ less EEG at operating points above their peak macro-F1. With the backbone held fixed, its stopper gained 1.78–3.92 macro-F1 points at matched budgets; at the system-level operating points, estimated analog front-end power was 31.0%–73.4% lower than the baselines, depending on amplifier. These retrospective results support acquisition-efficient sleep staging, while prospective hardware and clinical validation remain future work.

AI USE STATEMENT

We used generative AI tools to assist with writing and polishing the manuscript and to adjust the LaTeX formatting of figures and tables. We did not use generative AI to design experiments or to process the data used to create figures. The authors reviewed all AI-assisted work, including the associated LaTeX source and the correctness of the affected figures and tables. The authors take responsibility for the final text, code, figures, tables, results, and claims in this submission.

ETHICS STATEMENT

This study is a secondary analysis of previously collected sleep EEG datasets. We did not recruit participants, collect new physiological recordings, or conduct interventions. The datasets were used in accordance with their respective access and use conditions. Our results are based on retrospective algorithmic evaluation and do not establish clinical safety or effectiveness for diagnosis or closed-loop stimulation.

REPRODUCIBILITY STATEMENT

Sections 2–4 of the main paper define the causal observation setting, PAUSE model, training objective, stopping rule, and evaluation setup. In the Supplementary Material, Section 2.1 reports dataset composition and channel selection, Section 2.2 specifies EEG preprocessing, and Section 2.3 gives the subject-wise splits and cross-fitted out-of-fold protocol for calibration and stopper fitting. Sections 3.1–3.7 define the classification, prefix-wise, acquisition, speed–accuracy frontier, fold-aggregation, and deployment-cost metrics. Section 4.1 documents the backbone architecture, optimization, checkpoint selection, and random seed; Section 4.2 describes hyperparameter search; and Section 4.3 specifies the GP stopper and its fitting settings. Section 5.1 documents the system-level ECTS baselines, Section 5.2 the frozen-backbone decision policies, and Section 5.3 the full-observation sleep-staging baselines. Sections 6.1–6.4 provide the class-wise, stopper, ablation, and full-observation analyses underlying the main results, respectively. Sections 7.1–7.4 report the online-staging comparison, per-prefix calibration analyses, stop-versus-continue risk trajectories, and edge-deployment measurements, respectively. The software and hardware environment is stated in Section 4.1 of the main paper.

REFERENCES

Lejun Ai, He Chen, Yu Qiu, Yixue Hao, Xiaoli Li, Min Chen, and Xiao-kun Wu. Joint fine-grained representation learning and masked relational modeling for eeg-based automatic sleep staging in fabric space. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Lejun Ai, Yulong Li, Haodong Yi, Jixuan Xie, Yue Wang, Jia Liu, Min Chen, and Rui Wang. Resource efficient sleep staging via multi-level masking and prompt learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):3–11, Mar. 2026. doi: 10.1609/aaai.v40i1.36958. URL https://ojs.aaai.org/index.php/AAAI/article/view/36958.

Richard Berry, Rita Brooks, Susan Harding, Robin Lloyd, Stuart Quan, Matthew Troester, and Brad Vaughn. Aasm scoring manual updates for 2017 (version 2.4). *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, 13, 04 2017. doi: 10.5664/jcsm.6576.

R. Boostani, Foroozan Karimzadeh, and M. Nami. A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer methods and programs in biomedicine*, 140:77–91, 2017. doi: 10.1016/j.cmpb.2016.12.004.

Min Chen, Pan Li, Rui Wang, Yuanzhuo Xiang, Zhiheng Huang, Qiao Yu, Muyao He, Jia Liu, Jiaxi Wang, Minyu Su, et al. Multifunctional fiber-enabled intelligent health agents. *Advanced Materials*, 34(52):2200985, 2022a.

Min Chen, Jia Liu, Pan Li, Hamid Gharavi, Yixue Hao, Jingyu Ouyang, Jiayu Hu, Long Hu, Chong Hou, Iztok Humar, et al. Fabric computing: concepts, opportunities, and challenges. *The Innovation*, 3(6), 2022b.

Stephanie Devuyst. The dreams databases and assessment algorithm. Data set, 2005. URL https://doi.org/10.5281/zenodo.2650142.

Akinori F Ebihara, Taiki Miyagawa, Kazuyuki Sakurai, and Hitoshi Imaoka. Learning the optimal stopping for early classification within finite horizons via sequential probability ratio test. *arXiv preprint arXiv:2501.18059*, 2025.

Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

Mahdad Jafarzadeh Esfahani, Soha Farboud, Hong-Viet V Ngo, Jules Schneider, Frederik D Weber, Lucia M Talamini, and Martin Dresler. Closed-loop auditory stimulation of sleep slow oscillations: Basic principles and best practices. *Neuroscience & Biobehavioral Reviews*, 153:105379, 2023.

J. Fernandez-Mendoza and A. Vgontzas. Insomnia and its impact on physical and mental health. *Current Psychiatry Reports*, 15:1–8, 2013. doi: 10.1007/s11920-013-0418-8.

Louise Frøstrup Follin, Julie Anja Engelhard Christensen, Janita Vevelstad, Hilde T Juvodden, Rannveig Viste, Berit Hjelde Hansen, Mathias Perslev, Tobias Kaufmann, Alexander Neergaard Zahid, and Stine Knudsen-Heier. Optimizing automated sleep stage scoring of 5-s mini-epochs: a transfer learning study. *SLEEPJ*, 49(4):zsaf393, 2026.

Stephan Goerttler, Yucheng Wang, Emadeldeen Eldele, Fei He, and Min Wu. Msa-cnn: A lightweight multi-scale cnn with attention for sleep stage classification. *Biomedical Signal Processing and Control*, 120:110141, 2026. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2026.110141. URL https://www.sciencedirect.com/science/article/pii/S1746809426006956.

Ashish Gupta, Hari Prabhat Gupta, Bhaskar Biswas, and Tanima Dutta. Approaches and applications of early classification of time series: A review. *IEEE Transactions on Artificial Intelligence*, 1(1):47–61, 2020.

Thomas Hartvigsen, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner. Adaptive-halting policy network for early classification. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 101–110, 2019.

Shuguo Hu, Jun Hu, Junwei Lv, and Huaiwen Zhang. Earlymix: Hierarchical mixing for early time series classification. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209452.

Michael R. Irwin. Sleep and inflammation: partners in sickness and in health. *Nature Reviews Immunology*, 19, 07 2019. doi: 10.1038/s41577-019-0190-z.

Ali Kavoosi, Morgan P Mitchell, Raveen Kariyawasam, John E Fleming, Penny Lewis, Heidi Johansen-Berg, Hayriye Cagnan, and Timothy Denison. Morpheusnet: Resource efficient sleep stage classifier for embedded on-line systems. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2315–2320. IEEE, 2023.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. 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.

Chengfan Li, Yueyu Qi, Xuehai Ding, Junjuan Zhao, Tian Sang, and Matthew Lee. A deep learning method approach for sleep stage classification with eeg spectrogram. *International journal of environmental research and public health*, 19(10):6322, 2022.

Weiming Li and Junhui Gao. Automatic sleep staging by a hybrid model based on deep 1d-resnet-se and lstm with single-channel raw eeg signals. *PeerJ Computer Science*, 9:e1561, 2023.

Guisong Liu, Guoliang Wei, Shuqing Sun, Dandan Mao, Jiansong Zhang, Dechun Zhao, Xuelong Tian, Xing Wang, and Nanxi Chen. Micro sleepnet: efficient deep learning model for mobile terminal real-time sleep staging. *Frontiers in Neuroscience*, 17:1218072, 2023.

Junwei Lv, Yuqi Chu, Jun Hu, Peipei Li, and Xuegang Hu. Second-order confidence network for early classification of time series. *ACM Transactions on Intelligent Systems and Technology*, 15 (1):1–28, 2023.

Pejman Memar and Farhad Faradji. A novel multi-class eeg-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26:84–95, 2018. doi: 10.1109/TNSRE.2017.2776149.

Taiki Miyagawa and Akinori F Ebihara. The power of log-sum-exp: Sequential density ratio matrix estimation for speed-accuracy optimization. In *International Conference on Machine Learning*, pp. 7792–7804. PMLR, 2021.

Anh Nguyen, Galen Pogoncheff, Ban Xuan Dong, Nam Bui, Hoang Truong, Nhat Pham, Linh Nguyen, Hoang Nguyen-Huu, Khue Bui-Diem, Quan Vu-Tran-Thien, et al. A comprehensive study on the efficacy of a wearable sleep aid device featuring closed-loop real-time acoustic stimulation. *Scientific Reports*, 13(1):17515, 2023.

Huy Phan, Kristian P Lorenzen, Elisabeth Heremans, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, Mathias Baumert, Kaare B Mikkelsen, and Maarten De Vos. L-seqsleepnet: Whole-cycle long sequence modeling for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4748–4757, 2023.

Yu Qiu, Min Chen, Weifa Liang, Dusit Niyato, Yue Wang, Yizhe Li, Victor CM Leung, Yixue Hao, Long Hu, and Yin Zhang. Reliable or green? continual individualized inference provisioning in fabric metaverse via multi-exit acceleration. *IEEE Transactions on Mobile Computing*, 2024.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Aurélien Renault, Alexis Bondu, Antoine Cornuéjols, and Vincent Lemaire. Early classification of time series: A survey and benchmark. *arXiv preprint arXiv:2406.18332*, 2024.

Liran Ringel, Regev Cohen, Daniel Freedman, Michael Elad, and Yaniv Romano. Early time classification with accumulated accuracy gap control. *arXiv preprint arXiv:2402.00857*, 2024.

Patrick Schäfer and Ulf Leser. Teaser: early and accurate time series classification. *Data mining and knowledge discovery*, 34(5):1336–1362, 2020.

Alexander J Scott, Thomas L Webb, Marrissa Martyn-St James, Georgina Rowse, and Scott Weich. Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. *Sleep medicine reviews*, 60:101556, 2021.

Qi Shen, Junchang Xin, Xinyao Liu, Zhongyang Wang, Chuangang Li, Zhihong Huang, and Zhiqiong Wang. Lgsleepnet: an automatic sleep staging model based on local and global representation learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023.

Akara Supratak and Yike Guo. Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 641–644. IEEE, 2020.

Alexandra-Maria Tăuțan, Alessandro C Rossi, and Bogdan Ionescu. Automatic sleep scoring with lstm networks: impact of time granularity and input signals. *Biomedical Engineering/Biomedizinische Technik*, 67(4):267–281, 2022.

Texas Instruments. *ADS129x Low-Power, 8-Channel, 24-Bit Analog Front-End for Biopotential Measurements*, 2010. `https://www.ti.com/lit/ds/symlink/ads1294.pdf`.

Texas Instruments. *ADS1299-x Low-Noise, 4-, 6-, 8-Channel, 24-Bit, Analog-to-Digital Converter for EEG and Biopotential Measurements*, 2012. https://www.ti.com/lit/ds/symlink/ads1299.pdf.

Texas Instruments. *ADS131A0x 2- or 4-Channel, 24-Bit, 128-kSPS, Simultaneous-Sampling, Delta-Sigma ADC*, 2016. https://www.ti.com/lit/ds/symlink/ads131a04.pdf.

Zhaowen Wang, Dongdong Zhou, Qi Xu, Fengyu Cong, Mohammad Al-Sa'd, and Jenni Raitoharju. Ulw-sleepnet: An ultra-lightweight network for multimodal sleep stage scoring. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6596–6600, 2026. doi: 10.1109/ICASSP55912.2026.11462444.

Chenguang Yang, Baozhu Li, Yamei Li, Yixuan He, and Yuan Zhang. Lwsleepnet: A lightweight attention-based deep learning model for sleep staging with singlechannel eeg. *Digital Health*, 9: 20552076231188206, 2023.

Paul-Emile Zafar, Youssef Achenchabe, Alexis Bondu, Antoine Cornuéjols, and Vincent Lemaire. Early classification of time series: Cost-based multiclass algorithms. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. IEEE, 2021.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Yongqing Zhang, Wenpeng Cao, Lixiao Feng, Manqing Wang, Tianyu Geng, Jiliu Zhou, and Dongrui Gao. Shnn: A single-channel eeg sleep staging model based on semi-supervised learning. *Expert Systems with Applications*, 213:119288, 2023.

Caihong Zhao, Jinbao Li, and Yahong Guo. Sleepcontextnet: A temporal context network for automatic sleep staging based single-channel eeg. *Computer Methods and Programs in Biomedicine*, 220:106806, 2022.