AttentiveGRUAE: An Attention-Based GRU Autoencoder for Temporal Clustering and Behavioral Characterization of Depression from Wearable Data

Nidhi Soley*
Department of Biomedical Engineering,
Institute for Computational Medicine
Johns Hopkins University
Baltimore, MD

Vishal M. Patel
Department of Electrical and Computer Engineering
Johns Hopkins University
Baltimore, MD

Casey O. Taylor
Departments of Medicine and Biomedical Engineering,
Institute for Computational Medicine
Johns Hopkins University
Baltimore, MD

Abstract

In this study, we present AttentiveGRUAE, a novel attention-based gated recurrent unit (GRU) autoencoder designed for temporal clustering and prediction of outcome from longitudinal wearable data. Our model jointly optimizes three objectives: (1) learning a compact latent representation of daily behavioral features via sequence reconstruction, (2) predicting end-of-period depression rate through a binary classification head, and (3) identifying behavioral subtypes through the Gaussian Mixture Model (GMM) based soft clustering of learned embeddings. We evaluate AttentiveGRUAE on longitudinal sleep data from 372 participants (GLOBEM 2018–2019), and it demonstrates superior performance over baseline clustering, domain-aligned self-supervised, and ablated models in both clustering quality (silhouette score = 0.70 vs (0.32–0.70)) and depression classification (AUC = 0.74 vs (0.50–0.67)). Additionally, external validation on cross-year cohorts from 332 participants (GLOBEM 2020–2021) confirms cluster reproducibility (silhouette score = 0.63, AUC = 0.61) and stability. We further perform subtype analysis and visualize temporal attention, which highlights sleep-related differences between clusters and identifies salient time windows that align with changes in sleep regularity, yielding clinically interpretable explanations of risk.

1 Introduction

Wearable sensing provides longitudinal, real-world behavioral signals relevant to depression risk. Yet, many modeling approaches assume dense, high-frequency multimodal data or they function as black-box classifiers with limited clinical interpretability [1–3]. In practice, low-frequency daily summaries are widely available and clinically salient: short or irregular sleep is a modifiable risk factor repeatedly linked to depression [4–7]. Three known challenges for time-series health models are: (i) outcome-agnostic clustering that may not align with clinical screening endpoints [8]; (ii) poor

*Corresponding author: nsoley1@jhu.edu

Workshop on Learning from Time Series for Health, 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

temporal interpretability about when risk meaningfully diverges within a trajectory; and (iii) lack of reproducibility.

To address these gaps, we propose AttentiveGRUAE, a framework for interpretable and reproducible temporal clustering of time series wearable data. AttentiveGRUAE is trained end-to-end to jointly optimize (1) a sequence reconstruction loss to capture temporal dynamics, and (2) a binary outcome loss to ensure that representations are clinically informative. We evaluate this model on the public GLOBEM dataset [2], which comprises four cohorts of wearable data collected across different institutions and years. Our model is benchmarked against several baselines, including various ablations, traditional time series clustering models, and domain-aligned time-series baselines. We further examine cluster reproducibility and stability across multi-year cohorts and analyze behavioral profiles to provide clinically meaningful interpretations.

Contributions. (1) An interpretable, outcome-aware temporal clustering framework that couples attention-guided sequence encoding with soft subtyping for day-level insight; (2) comprehensive benchmarking against both domain-aligned baselines and classical ML methods, with ablative analyses of attention and joint training; and (3) reproducibility and stability demonstrated via cross-cohort (pre/post-COVID) validation and resampling-based Adjusted Rand Index (ARI), yielding consistent subtype structure.

2 Related Work

Passive sensing with wearables and smartphones has linked sleep regularity, mobility, and phone use to depressive symptoms, motivating machine learning approaches for screening from behavioral time series data [9–14]. While these studies established feasibility, many treat depression as a cross-sectional classification problem and do not discover clinically meaningful subtypes [15, 2]. In parallel, deep sequence models (e.g., RNN/CNN/GNN variants) have advanced representation learning in health data [16–20], and unsupervised deep temporal clustering has been explored for latent structure discovery [21–23]; however, outcome-agnostic clustering may misalign with screening endpoints and often lacks temporal interpretability. In the work by Wang et al. [8], outcome-guided subtyping (OG-DTC) addresses alignment but relies on hard assignments with limited interpretability. Attention-augmented GRU autoencoders have been used in high-frequency physiological signals (e.g., ECG) for supervised detection [24], a setting that differs from daily wearable summaries in sampling rate and task design. Self-supervised methods such as time-series representation learning via temporal and contextual contrasting (TS-TCC) [25] and broader time-series foundation models [26] primarily target dense, high-frequency multivariate data.

In contrast to outcome-agnostic clustering and high-frequency waveform settings, the focus here is on low-frequency wearable sleep summaries, with comparisons to domain-aligned self-supervised baselines and evaluation under temporal distribution shifts for reproducibility.

3 Methods

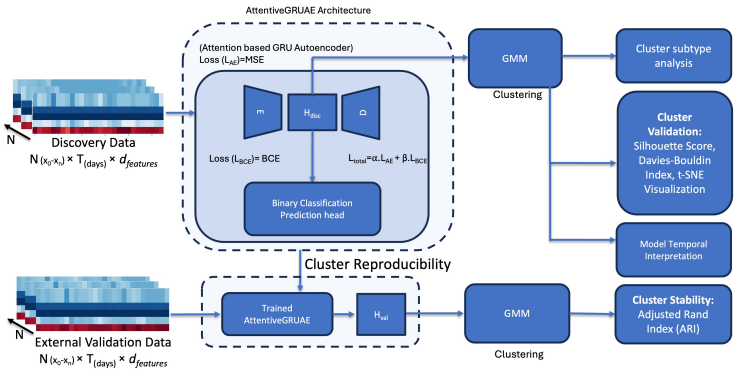

Figure 1: Outcome Guided-GRU encoder–decoder with attention; embeddings clustered by GMM (BIC for K). We report AUC/MSE, Silhouette/DBI, and assess cluster reproducibility via ARI on external validation data.

Overview. We propose *AttentiveGRUAE*, an interpretable framework for temporal clustering of behavioral time series. Given $X = \{x_i\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^{T\times d}$ (a T-day sequence of d wearable-derived features) and a binary outcome $y_i \in \{0, 1\}$, the model jointly learns to (i) reconstruct input sequences and (ii) predict the outcome, producing participant-level embeddings that are subsequently clustered into behavioral subtypes (Fig. 1).

Architecture and objective. AttentiveGRUAE uses a GRU encoder with multi-head attention over encoder states, a GRU decoder for sequence reconstruction, and a small feed-forward head for binary outcome prediction. The encoder produces a participant-level embedding that is pooled across time and shared by both heads: the decoder reconstructs the sequence, and the prediction head estimates the end-of-window outcome label. Training optimizes a weighted sum of a reconstruction loss (mean-squared error over all time steps and features) and a prediction loss (binary cross-entropy of the classification head). The architectural details and mathematical definitions are provided in the Supplementary Material S1.

Training and clustering. We train end-to-end with Adam optimizer [27], gradient clipping, dropout, and L_2 regularization; early stopping monitors validation AUC, and a ReduceLROnPlateau scheduler lowers the learning rate on plateaus (max 50 epochs, batch size 64). To mitigate occasional gradient conflict between tasks, we apply gradient surgery during joint training. After training, we extract latent embeddings h_i per participant and fit a GMM with K selected by BIC (Supplementary Fig. 4), yielding soft subtype assignments (supplement S2). We report MSE/AUC to quantify representation quality, Silhouette and Davies–Bouldin for cluster structure, and t-SNE qualitatively. For external validation, we freeze the encoder, compute H_{val} on a held-out cohort, and assess stability via ARI.

Experiment dataset and features. We use the public GLOBEM dataset [2] (705 participants) with a pre-COVID discovery split (DS1+DS2; 2018–2019; n=373) and a post-COVID validation split (DS3+DS4; 2020–2021; n=332). Inclusion requires $\geq$7 days of wearable data in a trailing 28-day window and a depression label at the end of the 28-day window. Outcomes follow GLOBEM's clinical cut points (PHQ-4 >2 or BDI-II >13). Inputs are six sleep summaries (duration, efficiency, latency, first bedtime/waketime, plus daily average duration), extracted daily, imputed using forward/backward fill, windowed to 28 days, and standardized using training-split statistics.

Baselines and ablations. Most recent foundation models target dense, high-frequency modalities; we therefore report (i) conventional baselines (PCA + GMM, k-means, time-series k-means) and (ii) a domain-aligned self-supervised baseline: TS-TCC [25] re-implemented from scratch on the same sleep-only discovery split, no pretraining, with the identical downstream pipeline (frozen encoder, GMM, same prediction head). We also ablate attention and the prediction head to isolate contributions of attention and multi-task learning. OG-DTC [8] is relevant but lacks public code; our AE-only ablation approximates its latent-space objective.

4 Results

4.1 Performance on discovery data and ablation

Table 1 shows that the full model achieves the best overall trade-off (MSE=0.47, AUC=0.74, Silhouette=0.70, DBI=0.33), and joint training with attention is critical: removing attention reduces AUC to 0.66; removing outcome guidance (AE-only) weakens cluster structure (Silhouette 0.52, DBI 1.70). TS-TCC, a strong self-supervised baseline on the same inputs, attains AUC 0.67 and Silhouette 0.45, but trails AttentiveGRUAE, suggesting that attention and outcome-informed objectives produce more clusterable, task-relevant embeddings for low-frequency sleep time series. Classical clustering methods underperform on both separability and prediction.

Table 1: Discovery (DS1+DS2). MSE only for AE-based models (— = n/a). Best in **bold**.

Model	MSE ↓	AUC ↑	Sil. ↑	DBI ↓
AttentiveGRUAE	**0.47**	**0.74**	**0.70**	**0.33**
No Attention	0.57	0.66	**0.70**	0.77
AE-only	**0.47**	—	0.52	1.70
Sequential Training	0.51	0.57	0.58	1.60
TS-TCC (self-sup, frozen)	—	0.67	0.45	—
PCA + GMM	—	0.50	0.32	6.39
k-means	—	0.51	0.47	1.89
Time Series k-means	—	0.50	0.58	0.65
GLOBEM bench. [2]	—	0.51	—	—

Table 2: Generalizability: discovery (DS1+DS2) vs. validation (DS3+DS4).

Model	AUC ↑		Silhouette ↑	
	DS1+DS2	DS3+DS4	DS1+DS2	DS3+DS4
AttentiveGRUAE (full)	**0.74**	**0.61**	**0.70**	**0.63**
No Attention	0.66	0.32	**0.70**	0.42
AE-only	—	—	0.52	0.32
Sequential Training	0.57	0.54	0.58	0.28

4.2 Reproducibility Across Cohorts and Cluster Stability

Transferring the frozen encoder to DS3+DS4 yields modest drops (Table 2): AttentiveGRUAE retains AUC 0.61 and Silhouette 0.63, indicating that embeddings learned on pre-COVID data remain clusterable and predictive post-COVID. The No-Attention variant degrades substantially (AUC 0.32, Silhouette 0.42), underscoring attention's role in robust temporal representation. On DS3+DS4, leave-out resampling (200 trials) yields mean ARI 0.89 (Supplementary Fig. 5), indicating high robustness to cohort perturbations and preservation of subtype structure.

4.3 Cluster subtype analysis and temporal interpretation

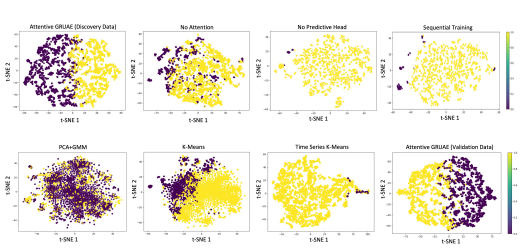

Figure 2: t-SNE of all the model variants and baselines (Cluster 1 = yellow, Cluster 0 = purple).

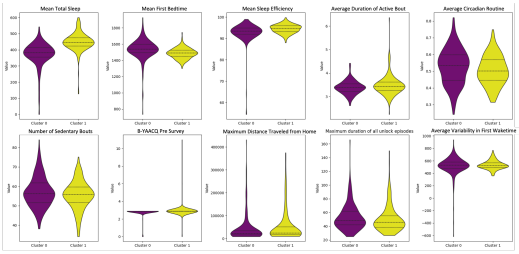

Figure 3: Violin plots for key differentiators.

AttentiveGRUAE yields clearly separated clusters on both discovery and validation cohorts (Fig. 2), indicating robust representations that generalize across cohorts. In contrast, a comparative t-SNE grid for ablations shows weaker and ambiguous structure, highlighting the benefit of outcome-informed attention. To quantify differences between subtypes, we ran per-feature Mann–Whitney U tests with Benjamini–Hochberg FDR correction and reported effect sizes via Cohen's d (Supplementary Figure 6, Table 3, 4). The violin plots (Fig. 3) highlight that the dominant sleep-domain signals, Cluster 1, exhibit longer sleep durations, earlier bedtimes, and higher efficiency. Furthermore, the interpretation of attention weights reveals that attention peaks align with salient trajectory changes: earlier in Cluster 0 (around day 7), near sharp declines in sleep duration, and slightly later in Cluster 1 (around day 9), as sleep stabilizes; see Supplementary S3 (Figs. 7–8).

5 Discussion

AttentiveGRUAE couples outcome-aware learning for time series data with soft latent clustering to yield reproducible, interpretable subtypes from low-frequency wearable sleep data. On discovery, it gives the best trade-off across reconstruction, prediction, and clustering metrics. On validation data, the performance drop may reflect behavioral changes due to the pandemic. Although clustering remained stable with perturbation (ARI = 0.89), suggesting the learned subtypes are reproducible even across distribution shifts. Ablations show both components matter: removing attention sharply degrades generalization, and decoupling reconstruction/prediction weakens cluster structure. This shows that outcome guidance and temporal attention are complementary for learning behaviorally salient representations. Further, the self-supervised baseline (TS-TCC) lags in performance, indicating that attention with task supervision is better matched to low-frequency signals. Cluster subtype analysis revealed that cluster 0 comprised individuals with shorter total sleep duration, greater variability in bedtime and circadian routine, and higher depression rates. In contrast, cluster 1 includes individuals with longer, more stable sleep patterns and a lower depression rate. These findings align with well-established links between sleep and depression. Reduced sleep duration,

sleep inefficiency, and irregular circadian patterns are known risk factors and correlates of depressive episodes [28–32].

Clinically, these results support *risk stratification* rather than diagnosis: passively measured sleep patterns can separate behaviorally coherent subtypes that align with depression screening endpoints and remain reproducible under temporal shift. Key limitations of our study are that the evaluation is confined to a single public cohort, and sleep-only input, chosen for interpretability, may underuse available signals. Attention weights are not causal explanations; they localize influential windows but do not establish causal treatment relations. Future work will extend to multi-modal inputs and further probe causal and counterfactual relations.

Data & Code Availability

Data. All experiments use the publicly hosted *GLOBEM* datasets (DS1–DS4) on PhysioNet [33]; the dataset was originally introduced in [34]. **Code.** Code, configs, and scripts are available in the AttentiveGRUAE repository [35].

References

[1] Eiko I. Fried and Randolph M. Nesse. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*, 172:96–102, 2015. doi: 10.1016/j.jad.2014.10.010.

[2] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34, 2023.

[3] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. Mobile sensing and support for people with depression: A pilot trial in the wild. *JMIR Mhealth Uhealth*, 4(3):e111, Sep 2016. ISSN 2291-5222. doi: 10.2196/mhealth.5960. URL `http://mhealth.jmir.org/2016/3/e111/`.

[4] Michael J. Murphy and Michael J. Peterson. Sleep disturbances in depression. *Sleep Medicine Clinics*, 10 (1):17–23, 2015. doi: 10.1016/j.jsmc.2014.11.009.

[5] S. Wang, M. E. Rossheim, R. R. Nandy, and U. S. Nguyen. Interaction between sleep duration and trouble sleeping on depressive symptoms among U.S. adults, NHANES 2015–2018. *Journal of Affective Disorders*, 351:285–292, 2024. doi: 10.1016/j.jad.2024.01.260.

[6] J. A. Lim, J. Y. Yun, S. H. Choi, S. Park, H. W. Suk, and J. H. Jang. Greater variability in daily sleep efficiency predicts depression and anxiety in young adults: Estimation of depression severity using the two-week sleep quality records of wearable devices. *Frontiers in Psychiatry*, 13:1041747, 2022. doi: 10.3389/fpsyt.2022.1041747.

[7] L. Li, C. Wu, Y. Gan, et al. Insomnia and the risk of depression: a meta-analysis of prospective cohort studies. *BMC Psychiatry*, 16:375, 2016. doi: 10.1186/s12888-016-1075-3.

[8] Dulin Wang et al. Clinical outcome-guided deep temporal clustering for disease progression subtyping. *Journal of Biomedical Informatics*, 158:104732, 2024. doi: 10.1016/j.jbi.2024.104732.

[9] Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7):1–11, 2015. doi: 10.2196/jmir.4273.

[10] Dror Ben-Zeev, Emily A. Scherer, Rui Wang, Haiyi Xie, and Andrew T. Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38(3):218, 2015. doi: 10.1037/prj0000130.

[11] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1): 1–26, 2018. doi: 10.1145/3191775.

[12] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3): 1–33, 2019. doi: 10.1145/3351274.

[13] Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. Leveraging collaborative-filtering for personalized behavior modeling: A case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–27, 2021. doi: 10.1145/3448107.

[14] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G. Creswell, Jennifer Mankoff, J. David Creswell, Mayank Goel, and Anind K. Dey. Detecting depression and predicting its onset using longitudinal symptoms cap. 2021. doi: 10.1145/3448123.

[15] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE Wireless Health (WH)*, pages 1–8, 2016. doi: 10.1109/WH.2016.7764553.

[16] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[17] Changhee Lee and Mihaela Van Der Schaar. Temporal phenotyping using deep predictive clustering of disease progression. In *International conference on machine learning*, pages 5767–5777. PMLR, 2020.

[18] Finneas JR Catling and Anthony H Wolff. Temporal convolutional networks allow early prediction of events in critical care. *Journal of the American Medical Informatics Association*, 27(3):355–365, 2020.

[19] Xiao Han, Yongjie Huang, Zhisong Pan, Wei Li, Yahao Hu, and Gengyou Lin. Multi-task time series forecasting based on graph neural networks. *Entropy*, 25(8):1136, 2023.

[20] Diego Machado Reyes, Hanqing Chao, Juergen Hahn, Li Shen, Pingkun Yan, and Alzheimer's Disease Neuroimaging Initiative. Identifying progression-specific alzheimer's subtypes using multimodal transformer. *Journal of Personalized Medicine*, 14(4):421, 2024.

[21] Naveen Sai Madiraju, Seid M. Sadat, Dimitry Fisher, and Homa Karimabadi. Deep temporal clustering: Fully unsupervised learning of time-domain features. In *IEEE International Conference on Data Mining (ICDM)*, pages 197–206, 2018. doi: 10.1109/ICDM.2018.00031.

[22] Ying Zhong, Dong Huang, and Chang-Dong Wang. Deep temporal contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9073–9081, 2021.

[23] Absalom E. Ezugwu, Abiodun M. Ikotun, Olaide O. Oyelade, Laith Abualigah, Jeffery O. Agushaka, Christopher I. Eke, and Andronicus A. Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022. doi: 10.1016/j.engappai.2022.104743.

[24] M. Roy, A. Halder, S. Majumder, and U. Biswas. Attentivecgru: Gru based autoencoder with attention mechanism and automated fuzzy thresholding for ecg arrhythmia detection. *Applied Soft Computing*, 167: 112337, 2024. doi: 10.1016/j.asoc.2024.112337.

[25] Ehab Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiao-Li Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021. URL `https://arxiv.org/abs/2106.14112`.

[26] Yuxuan Liang et al. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2024.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Kirstie N Anderson and Andrew J Bradley. Sleep disturbance in mental health problems and neurodegenerative disease. *Nature and science of sleep*, pages 61–75, 2013.

[29] David Nutt, Sue Wilson, and Louise Paterson. Sleep disorders as core symptoms of depression. *Dialogues in clinical neuroscience*, 10(3):329–336, 2008.

[30] Peter L Franzen and Daniel J Buysse. Sleep disturbances and depression: risk relationships for subsequent depression and therapeutic implications. *Dialogues in clinical neuroscience*, 10(4):473–481, 2008.

[31] Andrew G Mayers and David S Baldwin. The relationship between sleep disturbance and depression. *International Journal of Psychiatry in Clinical Practice*, 10(1):2–16, 2006.

[32] Anne Germain and David J Kupfer. Circadian rhythm disturbances in depression. *Human Psychopharmacology: Clinical and Experimental*, 23(7):571–585, 2008.

[33] Xuhai Xu, Ha-Young Zhang, Yasaman Sefidgar, Yun Ren, Xinzhi Liu, Woosuk Seo, Jesse Brown, Kathryn Kuehn, Micah Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Meredith Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. PhysioNet (version 1.1), 2023. RRID:SCR_007345.

[34] Xuhai Xu, Ha-Young Zhang, Yasaman Sefidgar, Yun Ren, Xinzhi Liu, Woosuk Seo, Jesse Brown, Kathryn Kuehn, Micah Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Meredith Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.

[35] Nidhi Soley, Vishal M. Patel, and Casey O. Taylor. Attentivegruae: An attention-based gru autoencoder for temporal clustering and behavioral characterization of depression from wearable data — code. `https://github.com/tirilab/AttentiveGRUAE--An-Attention-Based-GRU-Autoencoder-for-Temporal-Clustering`, 2025. GitHub repository; tag v1.0.0; accessed 2025-11-13.

[36] Norio Tsuno, Alain Besset, and Karen Ritchie. Sleep and depression. *Journal of Clinical Psychiatry*, 66 (10):1254–1269, 2005. doi: 10.4088/jcp.v66n1008.

[37] Yoran J. Toenders, Lianne Schmaal, Ben J. Harrison, Rianne Dinga, Michael Berk, and Christopher G. Davey. Neurovegetative symptom subtypes in young people with major depressive disorder and their structural brain correlates. *Translational Psychiatry*, 10(1):108, 2020. doi: 10.1038/s41398-020-0787-9.

[38] Andrew T. Drysdale, Logan Grosenick, Jonathan Downar, and et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, 23(1):28–38, 2017. doi: 10.1038/nm.4246. Correction: Nat Med. 2017 Feb 7;23(2):264.

[39] Allison G. Harvey. Sleep and circadian rhythms in bipolar disorder: Seeking synchrony, harmony, and regulation. *American Journal of Psychiatry*, 165(7):820–829, 2008. doi: 10.1176/appi.ajp.2008.08010098.

6 Supplementary Material

S1. Model Details and Notation

Notation. Let $\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^{T\times d}$ denote per-participant 28-day sequences of d wearable features; $y_i \in \{0,1\}$ is the end-of-window depression label. The encoder outputs a sequence $\{h_t\}_{t=1}^T$ and a pooled embedding $h_i \in \mathbb{R}^p$.

S1.1 GRU encoder and attention

For day t with input x_t and previous state h_{t-1}:

$$z_t = \sigma(W_z[h_{t-1}, x_t]), \qquad r_t = \sigma(W_r[h_{t-1}, x_t]),$$
$$\tilde{h}_t = \tanh\big(W_h[r_t \odot h_{t-1}, x_t]\big), \qquad h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

Multi-head attention. Two heads (d_k=16) are applied to $H = [h_1; \ldots; h_T]$. Let $H \in \mathbb{R}^{T\times d_h}$. For head $m = 1, \ldots, M$,

$$Q_m = HW_Q^{(m)}, \quad K_m = HW_K^{(m)}, \quad V_m = HW_V^{(m)},$$

with $W_Q^{(m)}, W_K^{(m)}, W_V^{(m)} \in \mathbb{R}^{d_h\times d_k}$ (learned)

$$A_m = \mathrm{softmax}\Big(\frac{Q_m K_m^\top}{\sqrt{d_k}}\Big) V_m$$

Heads are concatenated and time-pooled to yield the participant embedding h_i.

S1.2 Decoder and reconstruction loss

The decoder reconstructs $\hat{X}_i$ from h_i:

$$\mathcal{L}_{\text{AE}} = \frac{1}{NTd} \sum_{i=1}^{N} \left\| X_i - \hat{X}_i \right\|^2$$

S1.3 Outcome head and prediction loss

A two-layer MLP on h_i yields $\hat{y}_i = \sigma(f(h_i))$ with

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \Big]$$

S1.4 Joint objective and (α, β) selection

$$\boxed{\mathcal{L} = \alpha \, \mathcal{L}_{\text{AE}} + \beta \, \mathcal{L}_{\text{BCE}}}$$

We performed a lightweight random search on the discovery split ($\alpha \in \{0.3, 0.5, 0.7, 1.0\}$, $\beta \in \{0.5, 0.7, 1.0\}$; 5 draws), selecting the setting that maximized validation AUC (Silhouette as tiebreaker). Chosen: $(\alpha, \beta) = (0.7, 1.0)$.

S1.5 Gradient "surgery" to reduce task conflict

When $\langle g_{\text{AE}}, g_{\text{BCE}} \rangle < 0$, project away the conflicting component:

$$g_{\text{AE}} \leftarrow g_{\text{AE}} - \frac{\langle g_{\text{AE}}, g_{\text{BCE}} \rangle}{\|g_{\text{BCE}}\|^2} \, g_{\text{BCE}}, \qquad g \leftarrow g_{\text{AE}} + g_{\text{BCE}}.$$

This stabilized joint training without altering the objective.

S1.6 Feature selection for model

We restrict inputs to `sum_duration_asleep`, `avg_duration_asleep`, `avg_efficiency`, `avg_duration_to_fall_asleep`, `first_waketime`, and `first_bedtime` to capture quantity, quality, onset/offset (chronotype), and inertia of sleep over 28 days; sleep is a clinically established correlate/modifier of depression; and limiting d improves interpretability and reduces overfitting in low-frequency data. Additional modalities are used downstream for subtype interpretation.

S2. Training, Clustering, and Validation Details

Optimization. Adam (init LR 10^{-5}, ReduceLROnPlateau), batch size 64, max 50 epochs, gradient clipping (norm 1.0), dropout 0.3–0.5, L_2=10^{-4}; early stopping on validation AUC.

External validation. Freeze the encoder, embed DS3+DS4, reuse the trained GMM, and evaluate AUC/Silhouette. Cluster stability was assessed on external dataset using the ARI through a leave-one-out approach, randomly leaving out a small number of participants $n \in \{1, \ldots, 50\}$ from the external dataset and repeating the process 200 times to measure clustering robustness.

Selection for the number of clusters K. We fit GMMs for $K \in \{1, \ldots, 9\}$ on discovery embeddings and select K by BIC.

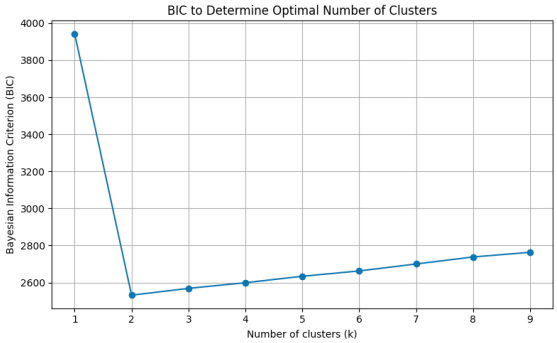

Figure 4: BIC across K. The elbow/minimum occurs at $K=2$. Models with $K>2$ did not improve AUC/Silhouette and produced small, unstable clusters.

Rationale. We retain $K=2$ for all ablations/baselines for comparability and because it aligns with prior outcome-guided temporal subtyping that often separates higher-risk vs. lower-risk disease trajectories [8], and with the two-group structure reported in GLOBEM-style depression analyses [2]. Specifically, peer-reviewed studies have described "hyposomnic" (insomnia/short-sleep) and "hypersomnic" (excessive/long-sleep) depression subtypes [36, 37]. Data-driven studies in youth and adult cohorts likewise discover sleep-driven subtypes with neurovegetative contrasts [38, 39]. Our two clusters align with these well-established categories, strengthening biological plausibility and suggests some concrete first-line intervention (improve sleep regularity) rather than one-size-fits-all pharmacotherapy approach.

S3. Additional Results and Plots

S3.1 ARI stability under resampling

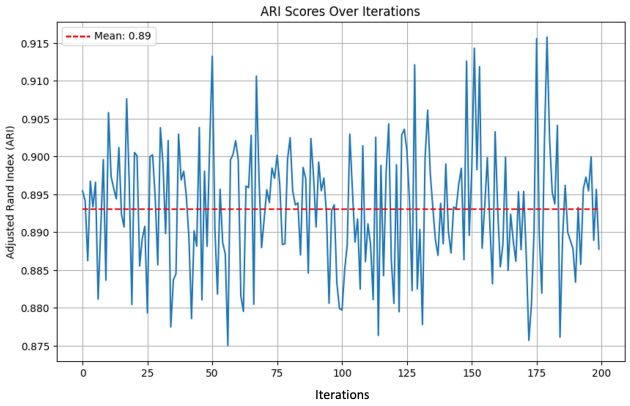

Figure 5: ARI over 200 resamples on DS3+DS4 (mean ≈ 0.89, dashed line).

S3.2 Top differentiating features (Cohen's d)

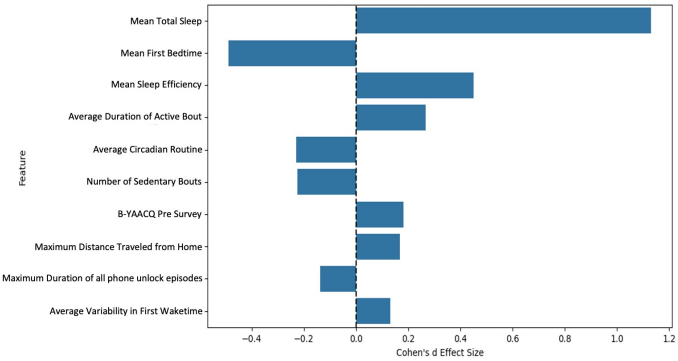

Figure 6: Top 10 features ranked by Cohen's d. Sleep features dominate.

S3.3 Attention-based temporal interpretation

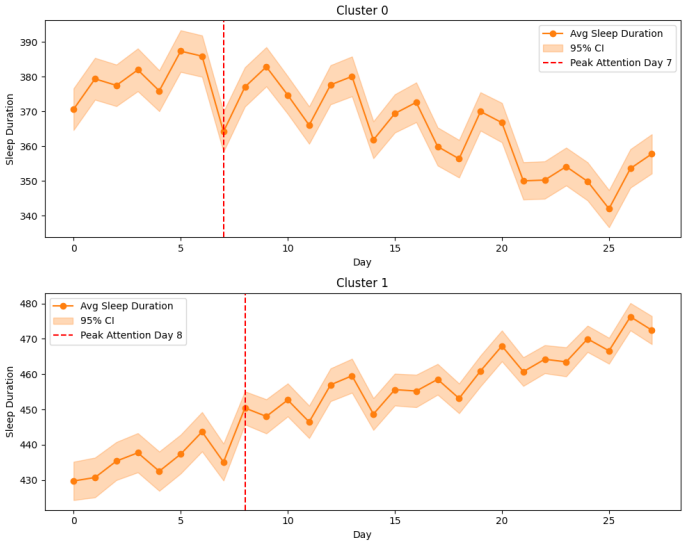

Figure 7: Cluster-level average sleep trajectories with peak attention day (red dashed). Cluster 0 peaks earlier (around day 7) near sharp declines; Cluster 1 peaks later (around day 8) as sleep stabilizes.

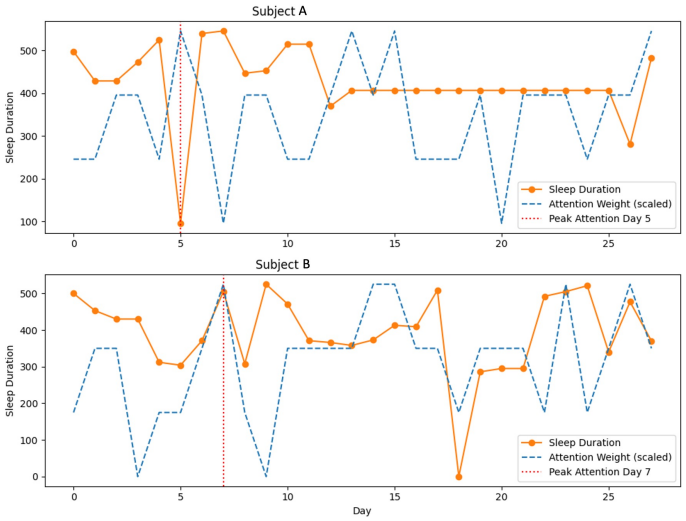

Figure 8: Attention weights over time (dashed blue line) for two participants A (cluster 0), B (cluster 1). Peak attention (vertical dashed red line) aligns with behavioral deviation (e.g., low, high sleep (solid orange line))

S3.4 Subtype summary (prevalence and peak attention)

Table 3: Subtype summary on discovery and validation: prevalence and average peak-attention day.

	Cluster 0	**Cluster 1**
Depressed fraction (disc.)	55%	45%
Peak attention (days)	5–7	8–10
Qualitative profile	lower/variable sleep	higher/stable sleep

Cluster 0 exhibits shorter total sleep, later bedtimes, and lower sleep efficiency; Cluster 1 shows longer, more stable sleep. Activity/survey signals are directionally consistent (longer active bouts, fewer sedentary bouts, and healthier BYAACQ patterns in Cluster 1). Attention peaks highlight inflection windows, which are earlier in Cluster 0 near declines; slightly later in Cluster 1 as stability returns.

S3.5. Cluster-wise comparison of behavioral and survey features (Cohen's d, FDR)

We ran Mann–Whitney U tests per feature with Benjamini–Hochberg Adjusted-FDR correction and computed Cohen's d.

Table 4: Cluster-wise comparison of behavioral/survey features. Means ± SD. Effect sizes are Cohen's d; q is FDR- p (top rows remain significant at $q<0.05$).

Feature	**Cluster 0 (N=195)**	**Cluster 1 (N=177)**	**Cohen's d / q**
Average Total Sleep Duration	371.65 ± 77.09	449.15 ± 58.73	1.13 / < 0.001
Average First Bedtime	1535.89 ± 125.75	1487.09 ± 64.62	-0.49 / < 0.001
Average Sleep Efficiency	92.85 ± 4.89	94.53 ± 2.05	0.45 / < 0.001
Average Active Bout Duration	3.38 ± 0.28	3.48 ± 0.42	0.27 / < 0.05
Average Daily Circadian Routine	0.53 ± 0.11	0.50 ± 0.09	-0.23 / < 0.05
Number of Sedentary Bouts	57.06 ± 7.86	55.35 ± 7.38	-0.22 / < 0.05
Average Variability in First Bedtime	91.55 ± 78.73	78.03 ± 67.67	-0.18 / < 0.05
BYAACQ (Pre)	2.81 ± 0.46	2.92 ± 0.72	0.18 / < 0.05
Max Distance from Home	42301.00 ± 58967.92	52913.61 ± 66576.54	0.17 / < 0.05
Average Variability in First Waketime	90.86 ± 77.76	78.88 ± 65.46	-0.17 / < 0.05
Max Duration of Phone Unlocks	53.32 ± 19.14	50.66 ± 19.51	-0.14 / < 0.05
PHQ-4 (weekly)	3.12 ± 0.15	3.10 ± 0.10	-0.12 / < 0.05
BFI-10 (Openness, Pre)	7.17 ± 0.50	7.05 ± 0.43	-0.12 / < 0.05
PSS-10 (Pre)	20.62 ± 1.33	19.76 ± 1.88	-0.09 / 0.227
ERQ-Suppression (Pre)	4.32 ± 0.25	4.28 ± 0.25	-0.14 / 0.323
2-Way SSS (Receiving Emotional, Pre)	25.00 ± 1.15	29.08 ± 1.70	0.05 / 0.193
Average First Waketime	518.63 ± 152.74	534.33 ± 71.13	0.13 / 0.976
Time Spent at Home	829.31 ± 149.77	813.28 ± 125.53	-0.12 / 0.129
Average Sleep Latency	0.15 ± 0.90	0.07 ± 0.43	-0.12 / 0.951
STAI (Pre)	45.87 ± 2.11	40.00 ± 0.00	-0.10 / 0.404
CES-D-9 (Pre)	8.37 ± 0.86	7.26 ± 1.34	-0.10 / 0.621
SocialFit (Pre)	74.17 ± 2.30	75.37 ± 1.72	0.10 / 0.387
EDS (Pre)	9.98 ± 1.81	10.16 ± 2.17	0.09 / 0.685
Phone Unlock Episodes (count)	102.22 ± 38.21	98.99 ± 34.32	-0.09 / 0.526
Avg Time in Active Bout (min)	186.75 ± 29.64	189.27 ± 31.74	0.08 / 0.842
Total Sedentary Bout Duration (min)	1250.74 ± 31.75	1253.25 ± 29.64	0.08 / 0.837
Avg Sedentary Bout Duration (min)	30.84 ± 7.04	31.28 ± 7.29	0.06 / 0.713
CHIPS (Pre)	20.04 ± 4.50	19.80 ± 3.39	-0.06 / 0.268
BDI-II (Pre)	14.34 ± 1.74	12.42 ± 2.11	-0.04 / 0.970
Avg Step Count	7380 ± 10.27	7500 ± 21.31	0.06 / 0.629
FSPWB (Pre)	43.09 ± 1.32	45.28 ± 1.42	0.14 / 0.494
Avg Duration of Unlock Episodes (min)	4.33 ± 2.61	4.29 ± 3.61	-0.01 / 0.156