

Stable and Interpretable Cognitive Status Screening Using Longitudinal Wearable Data

Haeun Kim

Gachon University

Taewan Cho

Gachon University

Andrew Jaeyong Choi

andrewjchoi@gachon.ac.kr

Gachon University

Seon-Heui Lee

Gachon University

Article

Keywords: cognitive screening, dementia, digital biomarkers, interpretable machine learning, wearable sensors

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11111108/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Stable and Interpretable Cognitive Status Screening Using Longitudinal Wearable Data

Haeun Kim¹, Taewan Cho¹, Andrew Jaeyong Choi^{1*},
Seon-Heui Lee^{2*}

¹Department of Artificial Intelligence, Gachon University, Seongnam,
13120, Republic of Korea.

²Research Institute of AI and Nursing Science, College of Nursing,
Gachon University, 191 Hambakmoe-ro, Yeonsu-gu, Incheon, 21936,
Republic of Korea.

*Corresponding author(s). E-mail(s): sunarea87@gachon.ac.kr;
andrewjchoi@gachon.ac.kr;

Abstract

Wearable devices repeatedly measure activity, sleep, and nocturnal physiology, but their value for three-class cognitive screening remains uncertain. We developed an interpretable hierarchical framework to distinguish cognitively normal (CN), mild cognitive impairment (MCI), and dementia (DEM) without the Mini-Mental State Examination or device-generated composite scores. Data from 174 adults aged 55 years or older were represented by seven-day windows encoding level, variability, trend, transitions, bouts, and complexity. Stage 1 separated DEM from non-DEM using the top-3 mean of window probabilities; Stage 2 separated CN from MCI using participant-median slope-and-change features. Participant-level cross-validation and repeated 3- to 6-fold sensitivity analyses preceded evaluation on an internal holdout of 38 participants. Accuracy was 0.711, balanced accuracy was 0.738, and macro F1 was 0.692; recall was 0.682, 0.700, and 0.833 for CN, MCI, and DEM, respectively. TreeSHAP and correlation grouping yielded four preliminary DEM patterns—activity complexity, device-state organization, rest, and light-sleep structure—and three exploratory MCI domains spanning activity persistence, sleep-wake timing, and nocturnal physiology. These findings support risk screening and hypothesis generation rather than diagnosis or validated biomarker claims; external multicenter validation is needed.

Keywords: cognitive screening, dementia, digital biomarkers, interpretable machine learning, wearable sensors

1 Introduction

Mild cognitive impairment (MCI) is characterized by cognitive decline while independence in daily activities is largely preserved [1]. Its course varies: some people remain stable or improve, whereas others progress to dementia. A 2018 American Academy of Neurology guideline, which has since been retired, addressed MCI assessment and management [2]. More recent dementia prevention guidance emphasizes risk-factor management, follow-up, and preparation for treatment and care [3]. Conventional cognitive tests remain essential, but they measure performance at a limited number of time points and may miss gradual changes in daily life.

Digital biomarkers are objective physiologic and behavioral measures obtained from sensors and wearable devices [4, 5]. Wearables repeatedly measure activity, inactivity, sleep-wake rhythms, heart rate, and heart rate variability in free-living settings. This longitudinal view may complement community screening and follow-up [4, 6], and digital screening tools and biomarkers for cognitive impairment have recently been reviewed systematically [7]. However, wearable signals are affected by device error, nonwear, lifestyle, comorbidities, medication use, and adherence. A prediction model must therefore distinguish genuine longitudinal behavior from measurement artifacts.

Most wearable cognitive studies have focused on binary CN–MCI discrimination or cognitive-score prediction [8, 9]. Fewer studies have evaluated CN, MCI, and dementia (DEM) together. This is difficult because the classes overlap, the MCI boundary is subtle, and DEM is often rare. Repeated windows increase the number of observations but not the number of independent participants. High-dimensional correlated features create a second problem: searching many models, thresholds, and feature subsets can overfit development criteria [10, 11]. Subject-level leakage can further inflate performance when repeated observations from the same person cross data partitions [12]. A feature can also improve one fitted model without remaining important under a different participant split.

We address these problems with a participant-level hierarchical framework. Stage 1 prioritizes DEM detection, and Stage 2 separates CN from MCI among participants routed as non-DEM. The framework aggregates evidence differently at each stage, keeps every participant within one cross-validation fold, and separates post hoc model explanation from clinical interpretation. Reporting follows the principles of participant-level prediction-model evaluation and transparent risk-of-bias assessment [13, 14].

The main contributions are as follows:

- We construct longitudinal wearable representations that describe level, variability, trend, temporal transitions, bout structure, and signal complexity.
- We formulate cognitive screening as hierarchical participant-level inference, with DEM-versus-non-DEM classification followed by CN-versus-MCI classification.
- We use stage-specific aggregation: a top-3 mean of window probabilities in Stage 1 and median aggregation of longitudinal features in Stage 2.
- We fix the operating thresholds on development data and examine operating-point sensitivity across repeated cross-validation schemes before internal-holdout evaluation, retaining the full error path in the end-to-end result.

- We distinguish SHAP-ranked model-explanation features from clinically interpretable candidate domains—four preliminary DEM patterns including light-sleep structure and three exploratory MCI domains—and retain explicit split, configuration, seed, and error-routing records in the reproduction audit.

2 Related Work

2.1 Wearable Cognitive Monitoring

Longitudinal fitness-tracker data have been used for MCI status prediction, and wearable physiologic features have been used to predict cognitive scores [8, 9]. Real-world wearable sensing has also identified altered spatial-navigation behavior in people with Alzheimer disease, while multimodal physical and behavioral measures have been used to estimate MCI stage [15, 16]. These studies use different sensors and prediction targets, but they show how daily-life signals can be transformed into machine-learning representations for cognitive assessment. Wearable sensors have also supported cognitive training and stress monitoring in older adults with MCI, while shank-mounted inertial sensors have been evaluated for gait-based MCI prescreening [17, 18]. Movement complexity, including entropy and transition-related measures, has also shown value for detecting cognitive change [19]. Actigraphic sleep and rest-activity rhythms have been associated with preclinical and early Alzheimer disease [20, 21]. These studies support the use of free-living behavior, but differences in cohorts, devices, targets, and validation designs prevent direct comparison of headline metrics.

Recent work has strengthened the case for rhythm and variability features. Jones *et al.* linked later activity onset and sleep-wake fragmentation with poorer cognitive performance in older adults with and without MCI [22]. Actigraphy studies have likewise associated irregular sleep-wake rhythms and circadian misalignment with cognitive function in older adults [23, 24]. Gabb *et al.* showed that remote sleep and cognitive monitoring can be feasible in people with CN, MCI, and dementia [25]. Greater day-to-day sleep-efficiency variability has been associated with lower cognitive scores, while wearable-measured sleep and activity have also been linked with memory performance [26, 27]. Prospective cohorts have connected weakened rest-activity rhythms and accelerometer-derived sleep-wake metrics with later MCI or dementia risk [28–30], and a systematic review summarized sedentary-behavior measurement in people with MCI and dementia [31]. These results support temporal regularity and within-person variability as candidate constructs, but do not establish diagnostic specificity.

Broader digital-health studies have examined cost-effective prediction from functional assessments and reviewed machine-learning approaches for noninvasive cognitive-impairment detection [32, 33]. These studies motivate accessible and reproducible screening, but they do not evaluate the same longitudinal wearable inputs or hierarchical three-class task used here. A recent reproducibility study also addressed MCI classification under extreme class imbalance using behavioral patterns derived from nocturia records [34].

Other studies have used electroencephalography and voice biomarkers for cognitive-impairment detection [35, 36]. These modalities can capture information that is not

Table 1 Participant Distribution by Cognitive Status

Split	CN	MCI	DEM	Total
Full cohort	111	51	12	174
Development	89	41	6	136
Internal holdout	22	10	6	38

available from activity and sleep sensors. In contrast, our study asks whether passive, long-duration wearable signals alone can support screening in daily-life settings.

2.2 Nocturnal Physiology and Interpretable AI

Nocturnal autonomic signals may add information beyond daytime activity. Altered heart rate variability during sleep has been reported in MCI [37], and autonomic function has been associated with later cognitive decline [38]. Sleep-related brain activity has also supported dementia detection, although neural recordings differ from consumer wearable signals [39]. Temperature dynamics may relate to cognitive performance [40]. Consumer devices, however, vary in the accuracy of sleep-stage, heart-rate, and heart-rate-variability estimates [41, 42].

TreeSHAP decomposes a tree-model prediction into feature-specific contributions [43]. It explains how a fitted model uses its input but does not establish causality, clinical importance, or an independent group difference. This distinction is especially important in high-stakes health applications, where post hoc explanations can be mistaken for mechanistic evidence [44, 45]. We therefore use SHAP for model transparency, apply correlation grouping to reduce redundant explanations, and promote only construct-level patterns with interpretable definitions and evidence-consistent directions to candidate domains.

3 Data and Longitudinal Feature Representation

3.1 Data Source and Participants

We used the AI Hub High-Risk Dementia Wearable Lifelog Dataset [46]. The dataset contains daily activity, sleep, related nocturnal physiologic signals, and source cognitive-status labels. The analysis included 174 adults aged 55 years or older: 111 CN, 51 MCI, and 12 DEM. Observation periods ranged from 35 to 121 days.

The development set contained 136 participants (CN, $n = 89$; MCI, $n = 41$; DEM, $n = 6$). The internal holdout contained 38 participants (CN, $n = 22$; MCI, $n = 10$; DEM, $n = 6$). All partitions were defined at the participant level. The available artifacts did not allow the study team to verify the diagnostic procedures or decision criteria used to assign the source labels. Performance therefore measures agreement with those labels, not with diagnoses re-evaluated by the authors. Table 1 summarizes the participant distribution.

3.2 Input Signals and Exclusions

Activity inputs included metabolic equivalent of task (MET), step count, movement, energy expenditure, activity duration, rest, and inactivity. We also used 1-minute MET and 5-minute device-coded activity-class sequences. Because the codebook did not define the clinical intensity represented by activity classes 1 and 3, these classes were retained as raw device states. Sleep inputs included total sleep time, sleep stages, sleep movement, and sleep midpoint. Nocturnal physiologic inputs included heart rate, root mean square of successive differences (RMSSD), respiratory rate, and temperature change. Appendix Table A1 describes the signal families.

The Mini-Mental State Examination (MMSE) was excluded to prevent a cognitive-test result from directly entering the model. Device-generated activity, sleep, and related composite scores were also excluded because their proprietary algorithms were unavailable and could indirectly encode health status. We checked duplicates, date order, nonwear indicators, missingness, and sequence validity. Valid and missing counts were retained as auxiliary quality variables but were not promoted to candidate clinical domains.

3.3 Windowing and Feature Engineering

Daily records were organized into 7-day windows. Overlapping windows with a 3-day stride were used to capture subtle local change, and nonoverlapping windows with a 7-day stride were used as a less redundant representation. Windows with fewer than 3 valid days were removed.

For each window, we calculated the mean, median, standard deviation, range, interquartile range, coefficient of variation, minimum, maximum, and selected percentiles. A linear slope against measurement day represented within-window change. From activity sequences, we calculated Shannon entropy, the number of observed states, transition counts and rates, state proportions, bout counts, and the longest bout. Time-of-day MET proportions described activity distribution. Sleep-stage, heart-rate, and RMSSD sequences were summarized using central tendency, variability, percentiles, transitions, and temporal slopes. We then ordered windows within each participant and calculated longitudinal slopes and changes across the observation period. Appendix Tables A1 and A2 provide the full representation framework.

4 Proposed Hierarchical AI Framework

Fig. 1 summarizes the full pipeline, and Algorithm 1 gives the participant-level inference procedure. The system is designed for risk screening: Stage 1 prioritizes DEM detection, and Stage 2 distinguishes MCI from CN among participants who are not routed to DEM. Random forest was selected through the development comparisons [47]; the methodological contribution is the full hierarchical and longitudinal design rather than the classifier alone. Stability of the operating point was verified before the primary evaluation: repeated 3- to 6-fold cross-validation preserved the development DEM sensitivity of 0.833 across all schemes, and the final threshold was fixed on development data only.

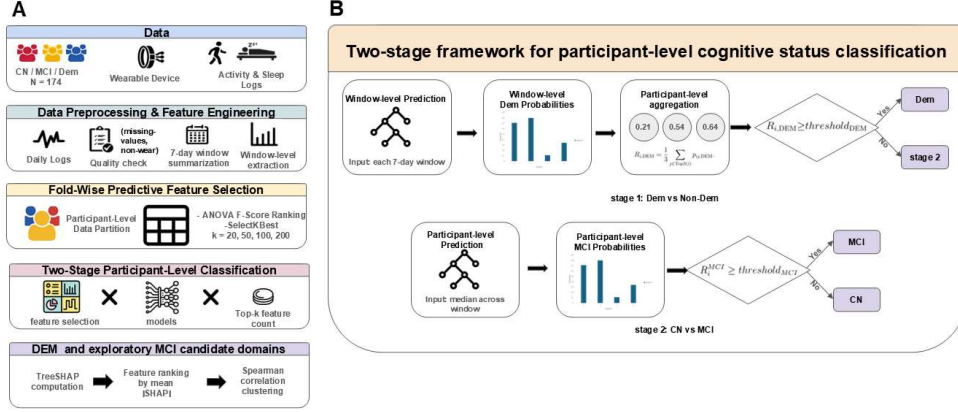


Fig. 1 Hierarchical participant-level framework. The upper lane shows end-to-end inference from longitudinal wearable signals to CN, MCI, or DEM. The lower lane separates participant-level validation and post hoc model interpretation from the prediction path.

4.1 Stage 1: DEM Versus Non-DEM

Stage 1 treats DEM as positive and combines CN and MCI as non-DEM. For participant i , let $p_{ij,DEM}$ be the predicted DEM probability for window j . The participant-level score is

$$R_{i,DEM} = \frac{1}{3} \sum_{j \in \text{Top3}(i)} p_{ij,DEM}. \quad (1)$$

The top-3 mean prevents a limited period with a strong DEM-like pattern from being diluted by many lower-probability windows. The final Stage 1 random forest used balanced class weights and the top 100 features. A participant was routed to DEM when $R_{i,DEM} \geq 0.21$.

4.2 Stage 2: CN Versus MCI

Stage 2 aggregates each selected feature before classification. Let x_{ijk} denote feature k in window j for participant i . The participant-level input is

$$\tilde{x}_{ik} = \text{median}_j(x_{ijk}). \quad (2)$$

The median reduces sensitivity to transient windows and extreme values. The final model used the top 100 slope-and-change features with an MCI class weight of 3. It generated one participant-level MCI probability $R_{i,MCI}$. A participant was classified as MCI when $R_{i,MCI} \geq 0.35$ and as CN otherwise.

Algorithm 1 Two-stage participant-level inference

Require: Participant windows X_i ; fitted models f_{DEM} and f_{MCI} ; thresholds 0.21 and 0.35

Ensure: Prediction $\hat{y}_i \in \{\text{CN}, \text{MCI}, \text{DEM}\}$

```
1: for each window  $x_{ij} \in X_i$  do
2:    $p_{ij,\text{DEM}} \leftarrow f_{\text{DEM}}(x_{ij})$ 
3: end for
4:  $R_{i,\text{DEM}} \leftarrow \text{Top3Mean}_j(p_{ij,\text{DEM}})$ 
5: if  $R_{i,\text{DEM}} \geq 0.21$  then
6:   return DEM
7: end if
8: for each Stage 2 feature  $k$  do
9:    $\tilde{x}_{ik} \leftarrow \text{median}_j(x_{ijk})$ 
10: end for
11:  $R_{i,\text{MCI}} \leftarrow f_{\text{MCI}}(\tilde{\mathbf{x}}_i)$ 
12: if  $R_{i,\text{MCI}} \geq 0.35$  then
13:   return MCI
14: else
15:   return CN
16: end if
```

The final decision rule is

$$\hat{y}_i = \begin{cases} \text{DEM}, & R_{i,\text{DEM}} \geq 0.21, \\ \text{MCI}, & R_{i,\text{DEM}} < 0.21 \text{ and } R_{i,\text{MCI}} \geq 0.35, \\ \text{CN}, & \text{otherwise.} \end{cases} \quad (3)$$

4.3 Model Explanation and Candidate-Domain Construction

After fixing the models and thresholds, we applied TreeSHAP [43]. Global importance was summarized by mean absolute SHAP value. We then grouped the top 100 features using Spearman feature–feature correlation at $|\rho| \geq 0.85$ and retained the feature with the largest mean absolute SHAP value in each group. This process left 34 nonredundant Stage 1 representatives and 66 Stage 2 representatives.

SHAP rank was not treated as biomarker rank. Candidate domains were constructed by jointly reviewing feature definitions, SHAP direction, group medians, Cliff’s delta, consistency among related features, and physiologic or behavioral interpretability. Features dominated by missingness, zero counts, or sequence length were not promoted to clinical domains. The thematic grouping was exploratory and had no prespecified statistical cutoff. This separation is important because feature stability under resampling is a different question from importance in one fitted model [48].

5 Experimental Protocol

5.1 Participant-Level Validation and Model Selection

All windows from a participant remained in one fold. Preprocessing, imputation, feature selection, class-imbalance handling, and threshold selection were restricted to development folds. Stage 1 used repeated stratified 4-fold cross-validation. Operating-point sensitivity was examined using 3- to 6-fold cross-validation repeated five times. Stage 2 used stratified 4-fold cross-validation in the 130 non-DEM development participants. The hierarchy and thresholds were frozen before the primary internal-holdout evaluation. For auditability, candidate-model and feature-representation results on the same holdout are reported post hoc in the Appendix; these descriptive audits are not treated as confirmatory evidence or as stable model rankings.

We compared logistic regression, random forest, extra trees, XGBoost, and LightGBM. Long short-term memory and gated recurrent unit networks were also examined as sequence baselines. Conventional models received derived window features, whereas recurrent models received ordered participant windows. Feature selection used analysis-of-variance F statistics within each training fold. We compared 20, 50, 100, and 200 retained features; mean, median, and top-3 aggregation; and multiple class-weight or synthetic minority oversampling settings [49]. Summary comparisons are provided in Appendix Tables B5–B7.

Stage 1 selected the highest-F1 operating point among thresholds with development DEM sensitivity of at least 0.80; specificity, precision, and balanced accuracy were secondary criteria. Stage 2 compared thresholds by balanced accuracy, macro F1, and MCI F1. The final thresholds were fixed before primary internal-holdout evaluation.

5.2 Evaluation Conditions and Metrics

Stage 2 was evaluated under two conditions. The Oracle condition entered all true CN and MCI holdout participants into Stage 2 and measured CN–MCI discrimination independently of Stage 1 errors. The conditional-pipeline condition included only true CN and MCI participants routed as non-DEM by Stage 1. The final end-to-end evaluation applied the hierarchy to all 38 participants. A DEM participant missed by Stage 1 was allowed to continue through Stage 2, and the resulting error was retained.

We report accuracy, balanced accuracy, macro F1, class recall and precision, specificity, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). Because the holdout was small, results are presented as point estimates and are not used for formal claims of superiority.

5.3 Reproducibility Audit

The selected pipeline was retrained from the source feature table in a clean uv environment using Python 3.11, NumPy 2.3.3, pandas 2.3.3, SciPy 1.16.2, scikit-learn 1.7.2, and joblib 1.5.2. The audit confirmed 174 unique participants, 136 development participants, 38 internal-holdout participants, and no participant overlap. It also confirmed

Table 2 Internal-Holdout Performance of Stage-Specific Models and Three-Class Architectures

Evaluation	n	Acc.	Bal. acc.	Macro F1	AUROC	AUPRC	Sens.	Spec.	Prec.
Stage 1: DEM vs non-DEM	38	0.816	0.823	0.735	0.807	0.512	0.833	0.812	0.455
Stage 2: Oracle CN vs MCI	32	0.844	0.832	0.823	0.805	0.645	0.800	0.864	0.727
Stage 2: Conditional pipeline	26	0.846	0.830	0.830	0.797	0.652	0.778	0.882	0.778
<i>Three-class end-to-end evaluation</i>									
Direct one-stage reference	38	0.289	0.369	0.223	—	—	—	—	—
Final two-stage framework	38	0.711	0.738	0.692	—	—	—	—	—

that none of the 3737 eligible features contained MMSE, cognition, cognitive, convert, or score identifiers. Independent local and remote reruns reproduced 27 correct predictions among 38 participants.

The technical appendix follows the analysis sequence used in the main text. [A](#) documents source signals and feature construction; [B](#) reports architecture, model, aggregation, class-imbalance, and threshold comparisons; [C](#) summarizes representative SHAP evidence; and [D](#) reports error propagation and class-wise discrimination.

5.4 Ethical Considerations

This study was a secondary analysis of de-identified data released through AI Hub. The authors did not recruit or contact participants, perform an intervention, or access directly identifying information. Therefore, additional institutional review board approval and informed consent were not required for this secondary analysis.

6 Results

6.1 Stage-Specific and End-to-End Performance

Table 2 summarizes the primary internal-holdout performance. Stage 1 identified 5 of 6 DEM participants and classified 26 of 32 non-DEM participants as non-DEM. It achieved an accuracy of 0.816, balanced accuracy of 0.823, macro F1 of 0.735, AUROC of 0.807, and AUPRC of 0.512.

In the Stage 2 Oracle evaluation, the random forest correctly classified 8 of 10 MCI participants and 19 of 22 CN participants. Accuracy was 0.844, balanced accuracy was 0.832, macro F1 was 0.823, AUROC was 0.805, and AUPRC was 0.645. Under the conditional-pipeline evaluation, accuracy was 0.846, balanced accuracy was 0.830, macro F1 was 0.830, MCI recall was 0.778, and CN specificity was 0.882.

The DEM-augmented direct three-class reference achieved an accuracy of 0.289, balanced accuracy of 0.369, and macro F1 of 0.223. The final hierarchy achieved 0.711, 0.738, and 0.692, respectively. This direct baseline is one reference configuration rather than an exhaustive representation of all one-stage models. Appendix Tables [B4–B7](#) and Fig. [B1](#) summarize the architecture, model-family, aggregation, imbalance-handling, and Stage 2 feature-representation comparisons.

Fig. 2 shows the final confusion matrix. Recall was 0.682 for CN, 0.700 for MCI, and 0.833 for DEM. Precision was 0.833, 0.778, and 0.455, respectively. Stage 1 routed 5 CN and 1 MCI participant to DEM. One DEM participant was routed as non-DEM

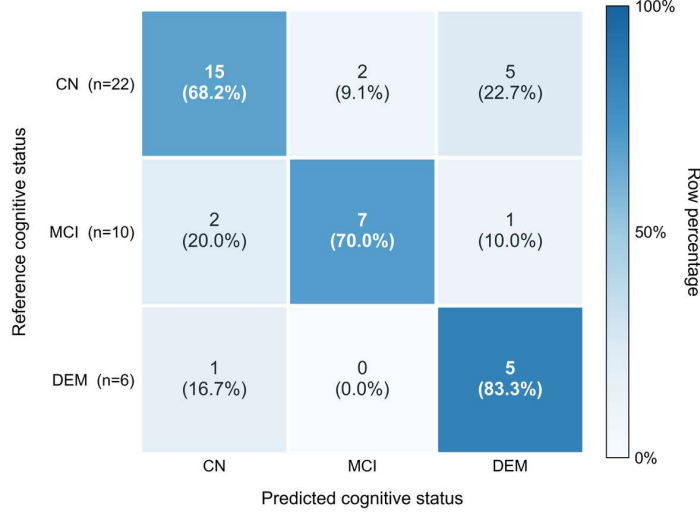


Fig. 2 Confusion matrix for the final three-class hierarchy on the internal holdout. Cells show participant count and row percentage.

and was subsequently classified as CN. Among the conditional Stage 2 cases, 2 CN and 2 MCI participants were confused with the other class. Appendix Tables B8 and B9 provide the full class-wise metrics and counts.

6.2 Threshold and Robustness Analysis

Fig. 3 summarizes the operating-point analyses. Across repeated 3- to 6-fold Stage 1 cross-validation, the sensitivity-prioritized rule selected thresholds between 0.18 and 0.21 while maintaining a development DEM sensitivity of 0.833 in every scheme. The 6-fold, five-repeat setting selected 0.21 because it had the highest development specificity among the tied-sensitivity schemes. When applied to the internal holdout, each selected threshold identified 5 of 6 DEM participants, corresponding to a recall of 0.833; this discrete result should be interpreted in light of the small DEM count. Post hoc Stage 2 holdout sensitivity was highest near the prespecified threshold of 0.35. Internal-holdout metrics were not used in the Stage 1 development tie-break. Full values are reported in Appendix Table B10.

6.3 DEM-Related Model Explanations

Fig. 4 shows the eight leading Stage 1 representatives after correlation grouping. Lower 1-minute MET entropy, lower 5-minute activity-class entropy, fewer device-coded class 3 bouts, greater occupancy of device-coded class 1, more rest, and higher device-recorded light-sleep measures contributed toward DEM. Table 3 groups these features into four preliminary behavioral and sleep patterns. Classes 1 and 3 are raw device codes and were not relabeled as low- or high-intensity states. The light-sleep measures were retained because they were selected among the top 100 Stage 1 features and showed a consistent DEM-higher group difference (Cliff's delta 0.627–0.688). Stage 1

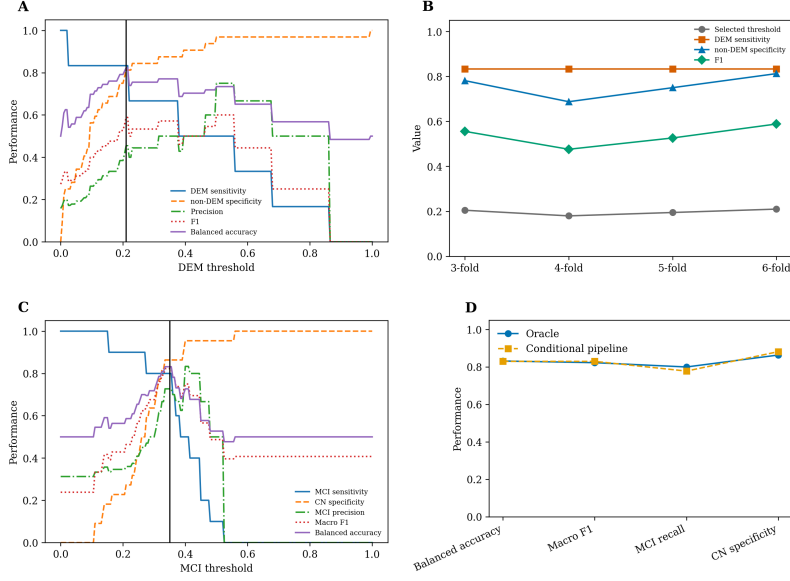


Fig. 3 Threshold and robustness analyses. Panels summarize Stage 1 DEM-threshold stability, Stage 2 MCI-threshold behavior, and Oracle-to-pipeline error propagation.

development included only six independent DEM participants, so all four patterns are preliminary. Appendix Table C11 lists representative DEM-related model-explanation features.

6.4 MCI-Related Model Explanations

Fig. 5 shows the eight leading Stage 2 representatives. The longitudinal slope of evening MET proportion ranked first by mean absolute SHAP, but the median was 0.000 in both CN and MCI, Cliff’s delta was -0.190 , and the SHAP direction did not align with the univariate distribution. We retained the feature for transparent model reporting but excluded it from the interpreted domains.

The remaining features supported three exploratory domains in Table 4: changes in activity persistence and fragmentation, instability in sleep-wake timing, and longitudinal changes in nocturnal autonomic and temperature dynamics. The full list contained 45 sleep-related representatives among 66 Stage 2 features. Each domain describes a joint model pattern rather than independent separation by one feature. Appendix Tables C12–C13 provide representative Stage 2 features and filtering diagnostics.

7 Discussion

7.1 Principal Findings

This study produced four main findings that organize the discussion. First, the hierarchical framework achieved a DEM recall of 0.833 and an MCI recall of 0.700 in

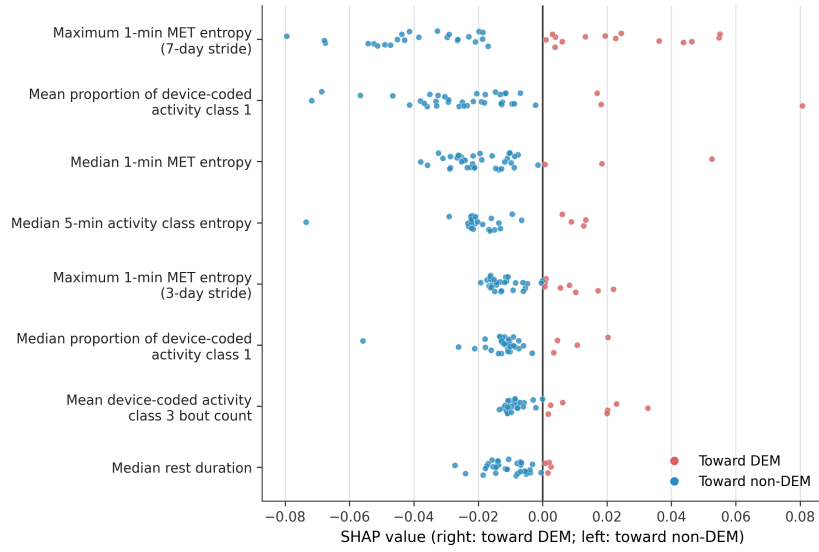
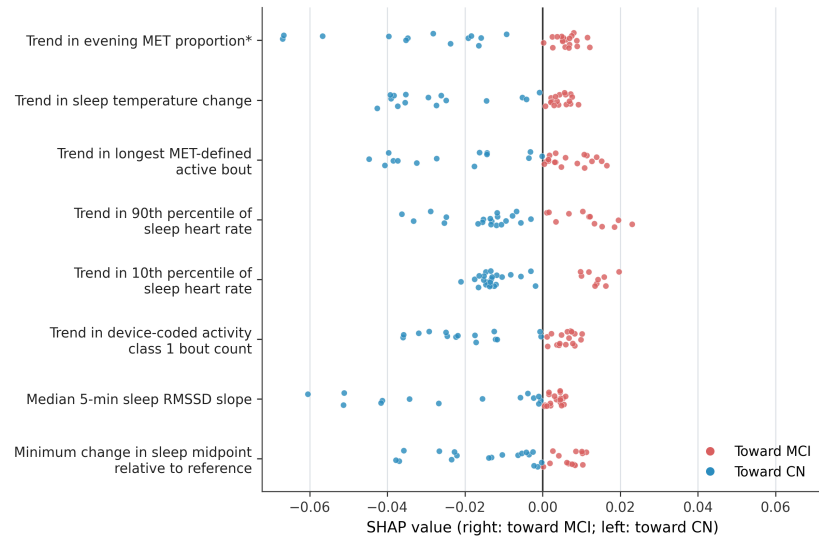


Fig. 4 Participant-level signed SHAP distributions for the eight leading Stage 1 representatives. Red indicates contribution toward DEM; blue indicates contribution toward non-DEM. The order is a model-explanation rank, not a clinical biomarker rank.



* Retained for model transparency; excluded from the three interpreted MCI candidate areas.

Fig. 5 Participant-level signed SHAP distributions for the eight leading Stage 2 representatives. Red indicates contribution toward MCI; blue indicates contribution toward CN. The starred evening MET feature is retained for transparency but excluded from the interpreted candidate domains.

Table 3 Preliminary DEM-Related Candidate Patterns

Pattern	Related features	model-explanation	Direction toward DEM	Interpretation
Activity complexity and diversity	Maximum and median MET entropy; median activity-class entropy	1-min 5-min	Lower	May reflect a narrower activity repertoire; Cliff’s delta = 0.208–0.333
Device-coded state organization	Mean and median proportion of activity class 1; mean class-3 bout count		Higher class-1; fewer class-3 bouts	Raw device states; clinical intensity definition unavailable
Rest and inactivity	Median and mean rest time; restless-sleep summaries		Higher	Greater inactivity and restless rest; Cliff’s delta = 0.302–0.641
Light-sleep structure	Mean, median, and maximum device-recorded light-sleep measures		Higher	DEM participants show more device-recorded light sleep; Cliff’s delta = 0.627–0.688; device-stage reliability limits clinical interpretation

Table 4 Exploratory MCI-Related Candidate Domains

Domain	Related features	model-explanation	Pattern toward MCI	Interpretation
Activity persistence and fragmentation	Slopes of longest active bout and device-coded class 1 bout count	MET-defined	Shorter sustained activity; more fragmented class-1 bouts	May reflect weaker sustained activity and altered temporal structure
Sleep-wake timing instability	Minimum change in sleep midpoint; midpoint trend, standard deviation, and interquartile range		Extreme shifts and greater variability	Timing instability rather than shorter total sleep time
Nocturnal autonomic and temperature dynamics	Slopes of sleep temperature change, heart-rate percentiles, and RMSSD		Higher temperature and RMSSD slopes; lower heart-rate percentile slopes	Joint longitudinal physiologic change, not a single direction of autonomic decline

the internal holdout, whereas the direct reference failed to identify MCI; sensitivity-oriented operating behavior was broadly consistent across the repeated cross-validation schemes. Second, Stage 1 explanations suggested four preliminary DEM patterns involving activity complexity, device-coded state organization, rest, and light-sleep

structure. Third, Stage 2 explanations supported exploratory MCI domains involving activity persistence and fragmentation, sleep-wake timing, and nocturnal autonomic and temperature dynamics. Fourth, the useful engineering unit was not random forest alone; performance depended on longitudinal feature construction, participant-level validation, stage-specific aggregation, class weighting, sensitivity-oriented thresholds, and explicit error routing.

7.2 Hierarchical Screening and Comparison With Prior Work

The hierarchy addresses an asymmetric screening problem. Missing a person with DEM by predicting CN can delay further assessment, whereas routing a CN participant to DEM creates an avoidable referral but does not directly assign a diagnosis. Stage 1 therefore prioritized DEM sensitivity. It identified 5 of 6 DEM participants but produced 6 false-positive DEM routes, a trade-off that is expected at a sensitivity-oriented operating point. Sensitivity-oriented operating behavior was broadly consistent across repeated 3- to 6-fold cross-validation, with a development DEM sensitivity of 0.833 in each scheme. The internal holdout also yielded 5 of 6 detected DEM cases, but the resulting recall of 0.833 is imprecise because one case changes the estimate by 16.7 percentage points. The final system should be understood as a triage aid that prioritizes formal cognitive assessment, not as an autonomous diagnostic tool.

The two-stage design amplified minority-class sensitivity relative to a single three-class model. The direct baseline had MCI recall of 0.000, whereas the hierarchy achieved 0.700; the Oracle Stage 2 recall of 0.800 decreased to 0.778 after Stage 1 routing and to 0.700 in the final three-class flow. This progression makes the error mechanism visible: an early routing error cannot be recovered by the later classifier. Fig. 6 and Appendix Tables D14–D15 trace this end-to-end error path. Prior wearable studies have usually evaluated binary CN–MCI discrimination or cognitive-score prediction [8, 9] rather than a three-class hierarchy, so headline metrics are not directly comparable. We therefore compare task design, feature domains, and error flow rather than claiming superiority from unmatched performance values. The two-stage amplification of DEM sensitivity is the main practical argument for the hierarchy: a single three-class model can reach comparable accuracy while missing every MCI case, which is unacceptable for a screening tool whose purpose is to avoid missed cognitive impairment.

7.3 Interpretation of DEM-Related Patterns

The leading DEM explanations emphasized lower entropy, less diverse device-coded activity states, fewer bouts of one state, more rest, and higher device-recorded light-sleep measures. These patterns are consistent at the construct level with reports that movement complexity and transitions can change before or during cognitive decline [19]. Prospective accelerometer studies have also linked lower activity, weaker rest-activity rhythm, and sleep-wake disruption with later MCI or dementia risk [28–30]. These comparisons support behavioral repertoire and temporal structure as plausible constructs.

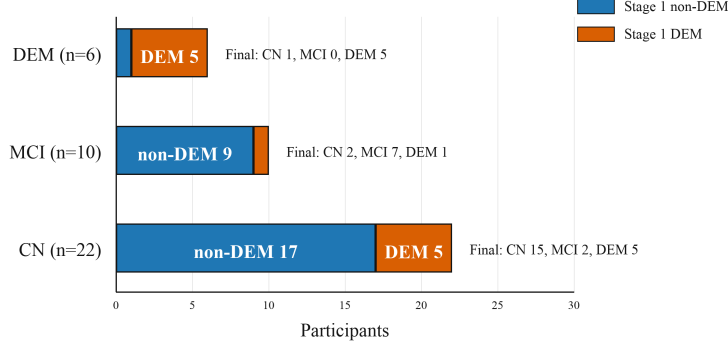


Fig. 6 Participant flow through Stage 1 and the resulting final class counts, showing how routing errors propagate through the hierarchy.

The DEM signal is not confined to daytime activity. Device-recorded light-sleep measures were retained as a fourth candidate pattern because they were selected among the top 100 Stage 1 features and showed a consistent DEM-higher group difference (Cliff’s delta 0.627–0.688). This is consistent with actigraphic reports linking altered sleep structure and rest-activity rhythms with preclinical and early Alzheimer disease [20, 21]. However, several other sleep-related representatives in the broader set reflected sequence length, valid counts, or other measurement-quality properties; we did not promote these to a clinical domain because their physiologic meaning was weaker than that of the light-sleep, activity-complexity, and rest features. Consumer sleep-stage estimates are device dependent and less reliable than total sleep time [41, 42], so the light-sleep finding should be read as model use of a device-recorded sleep construct, not as evidence that greater physiologic light sleep causes or diagnoses dementia. Larger DEM cohorts with validated sleep and autonomic measurements are needed to determine whether nocturnal physiology adds independent information.

The strength of the DEM evidence is limited by sample size. Only six independent DEM participants were available for development, although each produced many windows. Repeated windows improve temporal coverage but do not create new independent patients. The four DEM patterns are therefore explicitly labeled preliminary and carry less evidential weight than the MCI domains.

7.4 Interpretation of MCI-Related Domains

The MCI explanations combined activity, sleep timing, temperature, and autonomic features. A decreasing longest active bout and an increasing device-coded class-1 bout count may represent weaker sustained activity and greater fragmentation. However, class 1 was not defined clinically, so the finding cannot be translated into a specific activity intensity. The evening MET proportion ranked first by mean absolute SHAP, yet its median was 0.000 in both CN and MCI, Cliff’s delta was only -0.190 , and the SHAP direction did not align with the univariate distribution. We retained the feature for transparent model reporting but excluded it from the interpreted domains. Excluding a high-SHAP feature without a consistent group effect is an important

negative result: model importance alone is insufficient for a biomarker claim, and a single time-of-day activity ratio should not be over-read as an independent MCI marker.

Sleep-wake timing was supported by several related features rather than one rank. Extreme changes in sleep midpoint, long-term midpoint trend, and within-person variability formed a coherent temporal domain. This agrees at the construct level with reports linking activity onset and fragmented sleep-wake rhythms with poorer cognition [22] and with work using sleep-midpoint and rhythm features for CN-MCI classification [50]. It does not replicate an individual marker.

Temperature and nocturnal autonomic changes form a second physiologic domain. Prior work found altered sleep heart-rate variability in MCI [37], and autonomic function has been associated with later cognitive decline [38]. Temperature has also been associated with cognitive performance in older adults with and without MCI [40]. Our features were longitudinal slopes across wearable windows, not the same absolute physiologic measures used in those studies. They therefore provide construct-level support only.

7.5 Machine Learning and Digital Health Implications

Tree ensembles performed more reliably than the evaluated recurrent baselines in this small, high-dimensional tabular setting. Random forest was selected through the development comparisons rather than assumed a priori; the methodological contribution is the full hierarchical and longitudinal design rather than the classifier alone [47]. This does not imply that random forest is universally superior to deep learning. The number of independent participants, especially DEM participants, was too small to exploit a high-capacity sequence model safely, and a tree ensemble with exact TreeSHAP support kept the explanation method consistent with the predictive model. The results instead suggest that structured longitudinal features and explicit participant-level aggregation can be effective when the independent sample size is limited. Appendix Tables B5–B6 report the comparison details.

In community or nursing-led monitoring, the framework could help identify people whose changes in activity, sleep timing, or nocturnal physiology warrant formal assessment. Longitudinal monitoring may also reveal within-person change that a single visit does not capture. After formal assessment, digital cognitive interventions evaluated against active controls and usual care in MCI and dementia offer a downstream care option [51]. Such use would require clear escalation pathways, review by qualified clinicians, and monitoring of false-alert burden. Consumer wearable outputs must remain complementary to validated cognitive assessment and clinical judgment.

8 Limitations and Future Work

Several limitations affect the certainty of these findings. The full dataset included only 12 DEM participants, and development and holdout each contained six. One case changed holdout DEM recall by 16.7 percentage points. The clinically important DEM-to-CN error also shows how mistakes propagate through a sequential system.

The direct model was one reference, not an exhaustive test of all one-stage architectures. Prediction-model sample-size guidance further emphasizes that the number of independent participants, rather than repeated observations, determines the precision and stability of model estimates [52]. Point estimates from 38 holdout participants cannot establish stable model rankings, and small-sample validation studies can show high variance even when resampling is carefully designed [53].

The study used one dataset and one source-labeling process. Diagnostic procedures, assessment timing, depression, medication, pain, sleep disorders, and mobility limitations could not be fully verified or controlled. The closest prior CN-MCI study used the same AI Hub source cohort [50], so agreement is not external replication. Overlapping windows and many derived features increase correlation and overfitting risk. SHAP does not establish causality, and the reported domains are not validated biomarkers.

Consumer wearable measurements are device dependent. Sleep stages are generally less reliable than total sleep time, and heart-rate and heart-rate-variability estimates vary by device and population [41, 42]. People with cognitive impairment may also need support to maintain device use [25, 54]. Future studies should include larger MCI and DEM samples, external multicenter and cross-device validation, prospective follow-up, calibration assessment, adherence, false-alert burden, and outcomes after a positive screen.

Medical foundation models such as MedGemma may help integrate heterogeneous modalities in later work [55]. A useful future system should accept different combinations of wearable, clinical, voice, imaging, or questionnaire data because hospitals and patients rarely provide the same modalities. Foundation-model pretraining does not remove the need for reliable labels, missing-modality evaluation, and prospective clinical validation.

9 Conclusion

We developed a hierarchical and interpretable framework for participant-level cognitive screening from longitudinal wearable data without MMSE or device-generated composite scores. The internal-holdout evaluation yielded an accuracy of 0.711, balanced accuracy of 0.738, macro F1 of 0.692, MCI recall of 0.700, and DEM recall of 0.833, while sensitivity-oriented operating behavior was broadly consistent across repeated cross-validation schemes. The framework also separated model-explanation rank from clinical interpretation, yielding four preliminary DEM patterns—activity complexity, device-state organization, rest, and light-sleep structure—and three exploratory MCI candidate domains. These findings support further investigation of wearable-based risk screening and nursing-assisted longitudinal monitoring. Multicenter external validation is required before clinical use.

Appendix A Extended Data: Source Signals and Feature Derivation

This appendix documents the source wearable signals, longitudinal feature derivation, fixed model settings, development cross-validation summaries, internal-holdout performance, operating-point sensitivity, and representative model-explanation evidence. Model and threshold selection were performed using the development cohort. Candidate-configuration results on the same internal holdout are reported only as post hoc descriptive audits; the frozen two-stage hierarchy is the primary internal-holdout result. The holdout was drawn from the same data source and is not an external validation cohort.

Table A1 Available source signals and longitudinal feature-derivation framework

Source signal	Domain	Resolution	Derived feature families	Examples
1-min MET series	Activity	1 min	Level, percentiles, entropy, unique count, transitions, bouts, and slope/change features	MET entropy, transition rate, percentile trend
5-min activity-class sequence	Activity	5 min	Device-state proportions, bout counts, longest bouts, entropy, unique states, transitions, and slope/change features	Class-1 proportion, class-3 bout count, class entropy
Daily activity summaries	Activity	Daily / participant	Steps, total/high/low/rest duration, average MET, inactivity alerts, and time-of-day MET proportions	Evening MET-proportion trend, step-count trend
5-min sleep-stage sequence	Sleep	5 min	Stage proportions, bout counts, longest bouts, sequence length, transitions, and longitudinal change	Light-sleep summaries, longest stage-specific bout
Sleep heart-rate series	Sleep physiology	5 min	Central tendency, percentiles, variability, entropy, first/last values, nonzero summaries, and temporal slopes	Nocturnal HR p90 trend, HR-entropy trend
Sleep RMSSD series	Sleep physiology	5 min	Nonzero summaries, variability, transitions, first/last values, and temporal slopes	RMSSD trend, transition-ratio trend
Sleep timing, temperature, and breathing	Sleep / physiology	Window derived	Sleep midpoint, temperature change, breathing, and their longitudinal changes	Sleep-midpoint change, temperature-change trend
Longitudinal windowing and participant aggregation	Derived	7-day windows; 3-day overlapping or 7-day nonoverlapping stride	Window summaries; chronologically ordered slope/change; participant-level mean, median, or top-3 mean	Stage 1 top-3 probability mean; Stage 2 median feature aggregation

Malformed sequence tokens and empty, **nan**, **none**, or **null** entries were treated as missing. Device-coded activity classes 0–4 were retained as integer states because an authoritative clinical intensity mapping was unavailable; in particular, classes 1 and 3 were not relabeled as low- or high-intensity activity. Each analysis window covered 7 calendar days and required at least 3 valid days. Calendar gaps were retained when records were reindexed.

For time-of-day MET features, the clock blocks were night (00:00–06:00), morning (06:00–12:00), afternoon (12:00–18:00), and evening (18:00–24:00). Morning and

evening MET proportions were calculated as the number of active minutes in the corresponding block divided by all active minutes in that day; active minutes were defined as $\text{MET} \geq 1.5$. Daily numeric variables were summarized by mean, standard deviation, minimum, maximum, median, linear slope, valid and missing counts, range, interquartile range, and coefficient of variation.

Table A2 Preselection dimensions of the feature representations used in model development

Representation	Candidate features before within-fold selection
All eligible Stage 1 window features	3737
Stage 2 integrated activity-and-sleep features	748
Stage 2 peak-valley and distributional features	1104
Stage 2 variability features	1076
Stage 2 slope-and-change features	520

Note. The representation counts are not additive because the sets were defined by overlapping naming and derivation rules. The final Stage 2 model used the 520-dimensional slope-and-change representation before within-fold univariate selection of the top 100 features.

Appendix B Model Development and Performance

All windows from a participant remained in the same development fold or holdout partition. The fixed hierarchy used a Stage 1 random forest for DEM versus non-DEM screening and a Stage 2 random forest for CN versus MCI classification. The primary internal holdout contained 38 participants (CN, $n = 22$; MCI, $n = 10$; DEM, $n = 6$). Because only six DEM participants were available in this partition, candidate-model point estimates should not be interpreted as stable rankings.

Table B3 Fixed configurations used in the final two-stage hierarchy

Stage	Task	Model and confirmed settings	Feature representation / selection	Participant-level aggregation and imbalance handling	Threshold	Positive class
1	DEM vs non-DEM	Random forest; 120 trees	Selected window-derived candidates; SelectKBest(F statistic), $k = 100$	Mean of the three highest window-level DEM probabilities; balanced class weights	0.210	DEM
2	CN vs MCI	Random forest; 180 trees; minimum leaf size 2; square-root feature sampling	Slope-and-change representation; SelectKBest(F statistic), $k = 100$	Median of each window-derived feature; class weight CN:1, MCI:3	0.350	MCI

Table B4 Shared internal-holdout comparison of direct and hierarchical three-class prediction

Architecture	Model	n	Acc.	Bal. acc.	Macro F1	CN rec.	MCI rec.	DEM rec.
Direct one-stage reference	DEM-augmented three-class reference	38	0.289	0.369	0.223	0.273	0.000	0.833
Two-stage hierarchy	Stage 1 RF + Stage 2 RF	38	0.711	0.738	0.692	0.682	0.700	0.833

Note. The direct model is one reference configuration rather than an exhaustive one-stage benchmark. Both rows use the same 38-participant internal holdout; the comparison remains exploratory because the reference set is limited and the class counts are small.

Table B5 Stage 1 DEM-versus-non-DEM model comparison. Panel A reports development cross-validation mean $\pm$ standard deviation under fixed comparison settings. Panel B is a post hoc descriptive audit on the shared internal holdout.

Model	Threshold	AUROC	AUPRC	DEM sens.	Non-DEM spec.	Bal. acc.	Precision	DEM F1
<i>Panel A. Development cross-validation under fixed comparison settings</i>								
Logistic regression	0.790	0.867 $\pm$ 0.101	0.422 $\pm$ 0.186	0.750 $\pm$ 0.378	0.777 $\pm$ 0.069	0.764 $\pm$ 0.181	0.145 $\pm$ 0.117	0.233 $\pm$ 0.165
Random forest	0.240	0.839 $\pm$ 0.149	0.438 $\pm$ 0.274	0.688 $\pm$ 0.372	0.723 $\pm$ 0.214	0.705 $\pm$ 0.206	0.109 $\pm$ 0.074	0.182 $\pm$ 0.118
Extra Trees	0.395	0.825 $\pm$ 0.163	0.381 $\pm$ 0.147	0.812 $\pm$ 0.259	0.689 $\pm$ 0.168	0.751 $\pm$ 0.158	0.122 $\pm$ 0.062	0.204 $\pm$ 0.092
MLP	0.205	0.627 $\pm$ 0.230	0.304 $\pm$ 0.324	0.188 $\pm$ 0.372	0.954 $\pm$ 0.052	0.571 $\pm$ 0.195	0.188 $\pm$ 0.372	0.167 $\pm$ 0.309
LSTM	0.070	0.859 $\pm$ 0.137	0.457 $\pm$ 0.307	0.250 $\pm$ 0.378	0.977 $\pm$ 0.027	0.613 $\pm$ 0.198	0.375 $\pm$ 0.518	0.292 $\pm$ 0.415
GRU	0.040	0.839 $\pm$ 0.114	0.391 $\pm$ 0.293	0.438 $\pm$ 0.417	0.931 $\pm$ 0.036	0.684 $\pm$ 0.221	0.250 $\pm$ 0.231	0.312 $\pm$ 0.288
<i>Panel B. Post hoc descriptive audit on the internal holdout</i>								
Logistic regression	0.790	0.677	0.387	0.500	0.781	0.641	0.300	0.375
Random forest	0.210	0.807	0.512	0.833	0.812	0.823	0.455	0.588
Extra Trees	0.410	0.719	0.442	0.500	0.938	0.719	0.600	0.545
MLP	0.205	0.698	0.436	0.167	0.969	0.568	0.500	0.250
LSTM	0.070	0.682	0.471	0.500	0.969	0.734	0.750	0.600
GRU	0.040	0.693	0.472	0.500	0.969	0.734	0.750	0.600

Note. The random-forest threshold of 0.240 in Panel A belongs to the fixed-comparison run. The deployed threshold of 0.210 was selected in the separate repeated 3- to 6-fold operating-point analysis in Table B10. XGBoost and LightGBM are omitted from Stage 1 because final comparable artifacts were unavailable. Panel B is not interpreted as a stable candidate-model ranking.

Table B6 Post hoc Stage 2 CN-versus-MCI model comparison under conditional-pipeline evaluation. The subset contains true CN/MCI participants who were not assigned to DEM by Stage 1 ($n = 26$).

Model	Feature set / aggregation	k	Imbalance / threshold	AUROC	Bal. acc.	Macro F1	MCI rec.	CN spec.
Logistic regression	Slope/change; median	20	MCI weight 1.5; 0.550	0.686	0.716	0.710	0.667	0.765
Random forest	Slope/change; median	100	MCI weight 3; 0.350	0.797	0.830	0.830	0.778	0.882
Extra Trees	Slope/change; median	100	MCI weight 3; 0.350	0.719	0.797	0.764	0.889	0.706
XGBoost	Slope/change; median	100	SMOTE $k = 3$; 0.400	0.804	0.745	0.745	0.667	0.824
LightGBM	Slope/change; median	50	SMOTE $k = 3$; 0.400	0.745	0.716	0.710	0.667	0.765
MLP	Slope/change; median	200	Balanced; 0.450	0.673	0.712	0.685	0.778	0.647
LSTM	Integrated activity/sleep; median	50	Balanced; 0.500	0.379	0.556	0.505	0.111	1.000
GRU	Variability; top-3 mean	200	Balanced; 0.500	0.712	0.712	0.685	0.778	0.647
Linear SVM	Variability; top-3 mean	100	SMOTE $k = 3$; 0.550	0.797	0.686	0.675	0.667	0.706

Note. Metrics are participant-level internal-holdout point estimates. They are presented descriptively and should not be interpreted as a confirmatory algorithm ranking.

Table B7 Post hoc internal-holdout sensitivity analysis across Stage 2 feature representations. Oracle evaluation includes all true CN/MCI participants ($n = 32$); conditional-pipeline evaluation includes the true CN/MCI participants passed through Stage 1 ($n = 26$).

Feature set	Best configuration in the audit	Oracle AUROC	Oracle AUPRC	Oracle bal. acc.	Oracle macro F1	Oracle MCI rec.	Pipeline bal. acc.	Pipeline macro F1
Slope and change	RF; median; $k = 100$; MCI weight 3; threshold 0.350	0.805	0.645	0.832	0.823	0.800	0.830	0.830
Variability	GRU; top-3 mean; $k = 200$; balanced; threshold 0.500	0.686	0.422	0.695	0.648	0.800	0.712	0.685
Integrated activity and sleep	LR; top-3 mean; $k = 200$; MCI weight 1.5; threshold 0.250	0.641	0.402	0.682	0.561	1.000	0.676	0.571
Peak-valley	Linear SVM; top-3 mean; $k = 100$; SMOTE $k = 3$; threshold 0.100	0.650	0.487	0.682	0.561	1.000	0.676	0.571

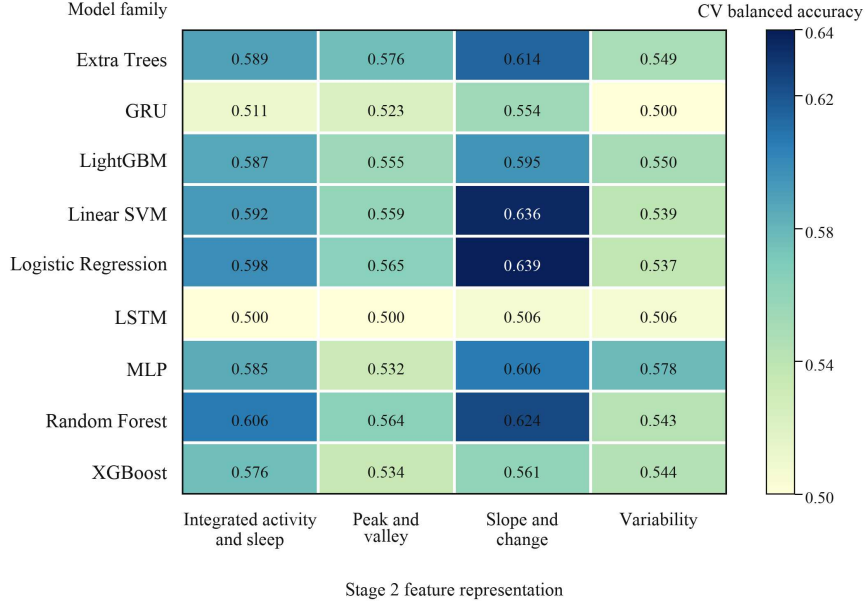


Fig. B1 Development cross-validation balanced accuracy across model families and Stage 2 feature representations under fixed comparison settings. This screening heatmap was one component of development and was not the sole criterion for selecting the final random-forest configuration.

Table B8 Final two-stage class-wise metrics on the 38-participant internal holdout

Class	Support	Predicted	TP	FN	FP	Precision	Recall	Specificity	F1
CN	22	18	15	7	3	0.833	0.682	0.812	0.750
MCI	10	9	7	3	2	0.778	0.700	0.929	0.737
DEM	6	11	5	1	6	0.455	0.833	0.812	0.588
Macro average	—	—	—	—	—	0.689	0.738	0.851	0.692

Table B9 Final two-stage confusion matrix and class-specific error rates

True class	Predicted CN	Predicted MCI	Predicted DEM	Class error rate
CN ($n = 22$)	15	2	5	31.8%
MCI ($n = 10$)	2	7	1	30.0%
DEM ($n = 6$)	1	0	5	16.7%

Table B10 Operating-point sensitivity analysis for Stage 1 across alternative development cross-validation schemes

<i>Panel A. Development operating-point selection</i>						
CV folds	Repeats	Selected threshold	DEM sensitivity	Non-DEM specificity	Precision	
3	5	0.205	0.833		0.723	0.122
4	5	0.180	0.833		0.654	0.100
5	5	0.195	0.833		0.754	0.135
6	5	0.210	0.833		0.769	0.143

<i>Panel B. Application of each selected threshold to the shared internal holdout</i>								
CV folds	Applied threshold	DEM sensitivity	Non-DEM specificity	Precision	DEM F1	CN→DEM	MCI→DEM	
3	0.205	0.833	0.781	0.417	0.556	0.227	0.200	
4	0.180	0.833	0.688	0.333	0.476	0.364	0.200	
5	0.195	0.833	0.750	0.385	0.526	0.273	0.200	
6	0.210	0.833	0.812	0.455	0.588	0.227	0.100	

Note. Each fold-count scheme independently selected an operating threshold using development predictions and then applied it to the same internal holdout. The 6-fold, five-repeat scheme selected 0.210 and provided the deployed Stage 1 threshold. This analysis assesses sensitivity to alternative development allocations; it is not external validation. With six DEM participants in the holdout, one classification changes DEM recall by 16.7 percentage points.

Appendix C Model-Explanation Evidence

The main text groups model-explanation features into clinically interpretable candidate domains. Tables C11 and C12 provide representative technical evidence supporting those domains. SHAP rank describes how the fitted model used a feature and is not interpreted as a causal, clinical, or validated biomarker rank. Complete listings of the 34 Stage 1 and 66 Stage 2 nonredundant representatives are maintained in the separate candidate-feature multimedia appendix.

Table C11 Representative DEM-related model-explanation features

Candidate domain	Representative feature	Original SHAP rank	Cliff's δ	Evidence-consistent interpretation
Activity complexity	1-min MET entropy maximum	1	-0.333	Lower feature values in DEM
Activity complexity	Activity-class entropy median	4	-0.208	Lower feature values in DEM
Device-state organization	Activity class-1 ratio mean	2	0.302	Higher device-coded class-1 occupancy in DEM
Device-state organization	Activity class-3 bout-count mean	12	-0.290	Fewer device-coded class-3 bouts in DEM
Rest and inactivity	Activity rest mean	5	0.302	Higher rest values in DEM
Light-sleep structure	Light-sleep median	28	0.627	Higher device-recorded light-sleep value in DEM
Light-sleep structure	Light-sleep maximum	50	0.688	Higher device-recorded light-sleep value in DEM

Note. For Stage 1, Cliff's δ was oriented as DEM minus non-DEM; positive values indicate higher participant-level values in DEM. The interpretation column summarizes the observed distribution after checking agreement with the corresponding SHAP pattern. Device-state codes and consumer-derived light-sleep values remain device constructs rather than validated physiologic states.

Table C12 Representative MCI-related model-explanation features

Candidate domain / status	Representative feature	Original SHAP rank(s)	Cliff's δ	Interpretation status
Excluded high-SHAP feature	Evening MET-proportion slope	1	-0.190	Reported transparently but excluded from domains because SHAP and univariate direction were not aligned
Activity persistence and fragmentation	Longest MET-defined active-bout slope	3	-0.184	More negative longitudinal change, consistent with reduced sustained activity
Activity persistence and fragmentation	Activity class-1 bout-count slope	7	0.259	Increasing device-coded bout segmentation
Sleep-wake timing	Minimum sleep-midpoint change	2	-0.255	More negative minimum timing change
Sleep-wake timing	Sleep-midpoint trend and variability	4-8	Mixed	Joint timing-instability pattern; not interpreted from one feature alone
Nocturnal autonomic and temperature dynamics	Sleep temperature-change slope	5	0.141	Small positive longitudinal change
Nocturnal autonomic and temperature dynamics	Sleep heart-rate percentile slopes	6-10	Negative across representatives	Lower longitudinal HR-percentile slopes

Note. For Stage 2, Cliff's δ was oriented as MCI minus CN; positive values indicate higher participant-level values in MCI. Rank ranges summarize related representatives and do not imply identical effects. Model direction, group distribution, missingness, and construct coherence were reviewed jointly before promotion to an exploratory domain.

Table C13 Correlation-based redundancy reduction of SHAP-ranked features

Stage	Initial SHAP-ranked features	Retained representatives	Correlation measure	Threshold	Representative-selection rule
Stage 1: DEM vs non-DEM	100	34	Spearman feature–feature correlation	$ \rho \geq 0.85$	Retain the feature with the largest mean absolute SHAP value in each correlated group
Stage 2: CN vs MCI	100	66	Spearman feature–feature correlation	$ \rho \geq 0.85$	Retain the feature with the largest mean absolute SHAP value in each correlated group

Note. This table describes redundancy reduction; it does not quantify agreement between native random-forest importance and SHAP rank.

Appendix D Robustness and Error Propagation

The final hierarchy preserves the full routing path. Participants assigned to DEM in Stage 1 do not enter Stage 2; therefore, an early false-positive DEM route cannot be corrected by the CN-versus-MCI classifier, and a missed DEM case can receive an erroneous CN or MCI label downstream.

Table D14 Stage-wise error propagation in the two-stage pipeline

Evaluation stage	True class	Support	Misclassified	Correct	Error rate	Error-propagation consequence
Stage 1	CN	22	5	17	22.7%	CN incorrectly routed to DEM and excluded from Stage 2
Stage 1	MCI	10	1	9	10.0%	MCI incorrectly routed to DEM and excluded from Stage 2
Stage 1	DEM	6	1	5	16.7%	DEM missed as non-DEM, passed to Stage 2, and ultimately classified as CN
Stage 2 conditional	CN	17	2	15	11.8%	CN classified as MCI among Stage 1 pass-through cases
Stage 2 conditional	MCI	9	2	7	22.2%	MCI classified as CN among Stage 1 pass-through cases

Table D15 ROC and precision-recall summary metrics for the frozen hierarchy

Stage	Evaluation condition	n	Positive n	Prevalence	Threshold	AUROC	AUPRC
Stage 1	Internal holdout	38	6	0.158	0.210	0.807	0.512
Stage 2	Conditional internal holdout	26	9	0.346	0.350	0.797	0.652

Note. Stage 2 metrics are based only on true CN/MCI participants who passed Stage 1. AUPRC values should be interpreted relative to the positive-class prevalence and should not be compared across stages without accounting for the different evaluation subsets.

Funding. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25489560) and the Gachon University research fund (GCU-202500670001).

Author contributions. H.K. and T.C. conceived the study. T.C. developed the computational pipeline and performed the analyses. H.K. led the clinical interpretation and manuscript preparation. A.J.C. and S.-H.L. supervised the study. All authors reviewed and approved the manuscript.

Competing interests. The authors declare no competing interests.

Data availability. The source data are available through the AI Hub High-Risk Dementia Wearable Lifelog Dataset [46], subject to the AI Hub application process and terms of use.

Code availability. The custom analysis code is not publicly available. It can be made available to editors and reviewers for evaluation and may be provided for non-commercial research on reasonable request to the corresponding authors, subject to institutional and data-use restrictions.

References

- [1] Albert, M. S. *et al.* The diagnosis of mild cognitive impairment due to Alzheimer’s disease: Recommendations from the national institute on aging–alzheimer’s association workgroups on diagnostic guidelines for alzheimer’s disease. *Alzheimer’s & Dementia* **7**, 270–279 (2011).
- [2] Petersen, R. C. *et al.* Practice guideline update summary: Mild cognitive impairment [retired]: Report of the guideline development, dissemination, and implementation subcommittee of the american academy of neurology. *Neurology* **90**, 126–135 (2018).
- [3] Livingston, G. *et al.* Dementia prevention, intervention, and care: 2024 report of the Lancet standing commission. *The Lancet* **404**, 572–628 (2024).
- [4] Kourtis, L. C., Regele, O. B., Wright, J. M. & Jones, G. B. Digital biomarkers for Alzheimer’s disease: The mobile/wearable devices opportunity. *npj Digital Medicine* **2**, 9 (2019).
- [5] Iulita, M. F., Streel, E. & Harrison, J. Digital biomarkers: Redefining clinical outcomes and the concept of meaningful change. *Alzheimer’s & Dementia: Translational Research & Clinical Interventions* **11**, e70114 (2025).
- [6] Qi, W. *et al.* Alzheimer’s disease digital biomarkers multidimensional landscape and AI model scoping review. *npj Digital Medicine* **8**, 366 (2025).
- [7] Hong, S., Kim, C. & Lee, S.-H. Digital tools and biomarkers for cognitive screening: A systematic review. *Digital Health* **12**, 20552076261461194 (2026).
- [8] Xu, Q., Kim, Y., Chung, K., Schulz, P. & Gottlieb, A. Prediction of mild cognitive impairment status: Pilot study of machine learning models based on longitudinal data from fitness trackers. *JMIR Formative Research* **8**, e55575 (2024).
- [9] Rykov, Y. G. *et al.* Predicting cognitive scores from wearable-based digital physiological features using machine learning: Data from a clinical trial in mild cognitive impairment. *BMC Medicine* **22**, 36 (2024).

- [10] Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91 (2006).
- [11] Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079–2107 (2010). URL <https://www.jmlr.org/papers/v11/cawley10a.html>.
- [12] Yagis, E. *et al.* Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports* **11**, 22544 (2021).
- [13] Collins, G. S. *et al.* TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
- [14] Moons, K. G. M. *et al.* PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).
- [15] Ghosh, A., Puthusserypady, V., Chan, D., Mascolo, C. & Hornberger, M. Machine learning detects altered spatial navigation features in outdoor behaviour of Alzheimer’s disease patients. *Scientific Reports* **12**, 3160 (2022).
- [16] Park, I. *et al.* Machine learning-based estimation of the mild cognitive impairment stage using multimodal physical and behavioral measures. *Scientific Reports* **15**, 35369 (2025).
- [17] Delmastro, F., Di Martino, F. & Dolciotti, C. Cognitive training and stress detection in MCI frail older people through wearable sensors and machine learning. *IEEE Access* **8**, 65573–65590 (2020).
- [18] Shahzad, A., Dadlani, A., Lee, H. & Kim, K. Automated prescreening of mild cognitive impairment using shank-mounted inertial sensors based gait biomarkers. *IEEE Access* **10**, 15835–15844 (2022).
- [19] Khan, T. & Jacobs, P. G. Prediction of mild cognitive impairment using movement complexity. *IEEE Journal of Biomedical and Health Informatics* **25**, 227–236 (2021).
- [20] Musiek, E. S. *et al.* Circadian rest-activity pattern changes in aging and preclinical Alzheimer disease. *JAMA Neurology* **75**, 582–590 (2018).
- [21] Alfini, A. *et al.* Associations of actigraphic sleep and circadian rest/activity rhythms with cognition in the early phase of Alzheimer’s disease. *Sleep Advances* **2**, zpab007 (2021).

- [22] Jones, C. D. *et al.* Detecting sleep/wake rhythm disruption related to cognition in older adults with and without mild cognitive impairment using the myRhythmWatch platform: Feasibility and correlation study. *JMIR Aging* **8**, e67294 (2025).
- [23] Okuda, M. *et al.* Effects of long sleep time and irregular sleep–wake rhythm on cognitive function in older people. *Scientific Reports* **11**, 7039 (2021).
- [24] Leahy, S., Xiao, Q., Yeung, C. H. C. & Figueiro, M. G. Associations between circadian alignment and cognitive functioning in a nationally representative sample of older adults. *Scientific Reports* **14**, 13509 (2024).
- [25] Gabb, V. G. *et al.* Longitudinal remote sleep and cognitive research in older adults with mild cognitive impairment and dementia: Prospective feasibility cohort study. *JMIR Aging* **8**, e72824 (2025).
- [26] Sakal, C., Li, T., Li, J., Yang, C. & Li, X. Association between sleep efficiency variability and cognition among older adults: Cross-sectional accelerometer study. *JMIR Aging* **7**, e54353 (2024).
- [27] Liu, G. H. *et al.* Associations of wearable-measured sleep and physical activity with memory performance in older adults: Cross-sectional study with actigraphy and MRI. *JMIR Aging* **8**, e80584 (2025).
- [28] Tranah, G. J. *et al.* Circadian activity rhythms and risk of incident dementia and mild cognitive impairment in older women. *Annals of Neurology* **70**, 722–732 (2011).
- [29] Haghayegh, S. *et al.* Association of rest-activity rhythm and risk of developing dementia or mild cognitive impairment in the middle-aged and older population: Prospective cohort study. *JMIR Public Health and Surveillance* **10**, e55211 (2024).
- [30] Cavallès, C. *et al.* Digital sleep-wake cycle metrics and dementia prediction in older adults. *JAMA Neurology* **83**, 694–704 (2026).
- [31] Wales, J. L. *et al.* Use of wearable technology for measuring and characterizing sedentary behavior in people with mild cognitive impairment and dementia: Systematic review. *JMIR Aging* **9**, e85361 (2026).
- [32] Shubar, A. G., Ramakrishnan, K. & Ho, C.-K. Optimizing machine learning models for accessible early cognitive impairment prediction: A novel cost-effective model selection algorithm. *IEEE Access* **12**, 180792–180814 (2024).
- [33] Alsuhaibani, M. *et al.* A review of machine learning approaches for non-invasive cognitive impairment detection. *IEEE Access* **13**, 56355–56384 (2025).

- [34] Ormeño-Arriagada, P., Robles, D., Márquez, G. & Taramasco, C. Reproducible framework for mild cognitive impairment classification from nocturia-derived behavioral patterns under extreme class imbalance. *IEEE Access* **14**, 103261–103275 (2026).
- [35] Kim, D. *et al.* Early detection of patients with mild cognitive impairment through EEG-SSVEP-based machine learning model. *IEEE Access* **12**, 172101–172114 (2024).
- [36] Pacheco-Lorenzo, M. R., Christensen, H., Anido-Rifón, L. E., Fernández-Iglesias, M. J. & Valladares-Rodríguez, S. M. Analysis of voice biomarkers for the detection of cognitive impairment. *IEEE Access* **12**, 122840–122851 (2024).
- [37] Kong, S. D. X. *et al.* Altered heart rate variability during sleep in mild cognitive impairment. *Sleep* **44**, zsaa232 (2021).
- [38] Nicolini, P. *et al.* Autonomic function predicts cognitive decline in mild cognitive impairment: Evidence from power spectral analysis of heart rate variability in a longitudinal study. *Frontiers in Aging Neuroscience* **14**, 886023 (2022).
- [39] Ye, E. M. *et al.* Dementia detection from brain activity during sleep. *Sleep* **46**, zsac286 (2023).
- [40] Eggenberger, P., Bürgisser, M., Rossi, R. M. & Annaheim, S. Body temperature is associated with cognitive performance in older adults with and without mild cognitive impairment: A cross-sectional analysis. *Frontiers in Aging Neuroscience* **13**, 585904 (2021).
- [41] Green, S. F. *et al.* A systematic review of the validity of non-invasive sleep-measuring devices in mid-to-late life adults: Future utility for Alzheimer’s disease research. *Sleep Medicine Reviews* **65**, 101665 (2022).
- [42] Miller, D. J., Sargent, C. & Roach, G. D. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. *Sensors* **22**, 6317 (2022).
- [43] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30**, 4765–4774 (2017).
- [44] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).
- [45] Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* **3**, e745–e750 (2021).

- [46] AI Hub. High-risk dementia wearable lifelog dataset. National Information Society Agency (2026). URL <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=226>. Accessed: Aug. 3, 2026.
- [47] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
- [48] Nogueira, S., Sechidis, K. & Brown, G. On the stability of feature selection algorithms. *Journal of Machine Learning Research* **18**, 1–54 (2018). URL <https://jmlr.org/papers/v18/17-514.html>.
- [49] Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002).
- [50] Kim, H. & Park, J.-Y. Longitudinal variability of wearable-derived sleep and heart rate as a digital biomarker for early detection of mild cognitive impairment. *Yonsei Medical Journal* **67**, 524–533 (2026).
- [51] Lee, H., Lim, Y. & Lee, S.-H. Comparing digital cognitive interventions to active controls and usual care for mild cognitive impairment and dementia: A systematic review and meta-analysis. *Medicina* **62**, 1162 (2026).
- [52] Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**, m441 (2020).
- [53] Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2018).
- [54] Chien, S.-Y., Zaslavsky, O. & Berridge, C. Technology usability for people living with dementia: Concept analysis. *JMIR Aging* **7**, e51987 (2024).
- [55] Sellergren, A. *et al.* MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081* (2026).