

Original Research

Received 27 June 2026

Revised 02 Aug 2026

Accepted 15 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (16s): 768–788

KEYWORDS:

Explainable artificial intelligence; SHAP; Lifespan-aware modelling; Behavioural change prediction; Wearable accelerometry; Explanation stability; Responsible AI; NHANES

Explainable AI for Tracking and Predicting Human Behavioral Changes Across the Lifespan

Dr. S. Mahalakshmi^{1*}, Dr. Gayetri Devi S.V², C. S. Sandhya Sri³, S. A. Althaf Ahamed⁴, R. Gayathri⁵

¹Assistant Professor, Department of Computer Science w/s Artificial Intelligence & Machine Learning, Faculty of Liberal Arts & Business Studies, SRM Institute of Science and Technology, Ramapuram Campus, Chennai – 600 089, India. Email: drmahalakshmiphdc@gmail.com

²Professor of Practice, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, India. Email: gayetri.venkhatraman@gmail.com

³Assistant Professor, Department of Computer Science and Engineering, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, #42, Vel Tech Road, Vel Nagar, Avadi, Chennai – 600 062, India. Email: sandhyacs248@gmail.com

⁴Assistant Professor, Department of Computer Science and Engineering, Hindusthan College of Technology, Sikkanampatti, Omalur, Opposite Salem Airport, Salem – 636 309, India. Email: althafahamed.cse@dgct.ac.in

⁵Assistant Professor, Department of Computer Science and Engineering, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, India. Email: rgayathri@veltechmultitech.org

*Corresponding author: Dr. S. Mahalakshmi (drmahalakshmiphdc@gmail.com)

ABSTRACT: Background: Everyday behaviours such as physical activity, sedentary time and sleep change systematically across the human lifespan, yet most behavioural prediction models are trained on pooled populations and explained, if at all, with a single global attribution profile.

Problem: Pooled models can obscure age-dependent behavioural mechanisms, and post-hoc explanations are rarely tested for stability.

Objective: To develop and evaluate a Lifespan-Aware Explainable Behavioral Prediction Framework (LA-EBPF) that predicts short-term behavioural change and explains its predictions separately for lifespan groups.

Framework: LA-EBPF integrates behavioural-change feature engineering (day-to-day differences, trends, variability, transitions), five literature-informed lifespan strata, partially pooled shared-plus-group gradient-boosted models, lifespan-conditional Shapley attribution, a Lifespan Attribution Divergence (LAD) index and an Explanation Stability Index (ESI).

Data and simulation: The variable schema, wear protocol and age structure follow the public NHANES 2011–2014 wrist-accelerometer Physical Activity Monitor files; the numerical evaluation uses a controlled, calibrated synthetic cohort (N = 11,000) with known, age-varying behavioural drivers. The task is to predict a person-relative downturn in day-7 activity from six preceding days.

Simulation findings: Across ten repeated participant-level splits, LA-EBPF reached ROC-AUC 0.770 ± 0.013 , F1 0.525 ± 0.017 and MCC 0.337 ± 0.023 , versus 0.732 ± 0.013 , 0.495 ± 0.020 and 0.298 ± 0.027 for the strongest baseline (gradient-boosted trees on raw day-level features). Removing behavioural-change features produced the largest loss (AUC 0.725). Group attribution profiles diverged well beyond a permutation null (mean LAD 0.171 vs 0.0018; $p = 0.005$), and 13 of 15 simulated group-specific drivers were recovered among the top-eight attributions. Stability-aware screening reduced the feature set by about 28% without loss of accuracy but did not raise rank-level ESI. Novelty and significance: LA-EBPF contributes an integrated, auditable procedure for lifespan-stratified behavioural prediction with explanation-stability reporting. All numerical results are simulation-based and require empirical replication on real longitudinal data.

1. INTRODUCTION

Human behaviour is not static. Physical activity, sedentary time, sleep duration and sleep timing fluctuate from day to day and shift systematically from childhood to old age [1, 2]. Population accelerometry has shown that device-measured activity falls steeply from childhood through adolescence and declines again in later life [3–5], while sleep duration, architecture and timing follow their own age-dependent trajectories [6]. A lifespan perspective therefore implies that the same observed behavioural signal—an irregular sleep schedule, a weekend, a declining weekly trend—may carry different predictive meaning at different ages.

Ubiquitous sensing has made such behaviour measurable at scale. Digital-phenotyping and personal-sensing research uses passively collected sensor streams to infer behavioural and health-related states [7–9], and machine learning (ML) is now routinely applied to physical-activity, sedentary and sleep data [10]. These models are, however, usually trained on pooled samples or on a single age band, and their outputs are difficult to interpret. Explainable AI (XAI), particularly Shapley-value attribution [11, 12], offers a principled way to quantify each feature’s predictive contribution, but three problems remain. First, a single global attribution profile averages over heterogeneous subpopulations and can hide age-specific mechanisms. Second, attributions can be unstable under resampling or small input perturbations, and different explainers can disagree [13–15]; explanations whose stability is not reported should not be relied upon in sensitive settings [16, 17]. Third, most pipelines represent behaviour by levels (means or totals), not by the change dynamics—differences, trends and variability—that define behavioural change in the first place.

This study addresses these gaps by proposing the **Lifespan-Aware Explainable Behavioral Prediction Framework (LA-EBPF)**. LA-EBPF is not a new learning algorithm; its novelty lies in the integration of lifespan-stratified, partially pooled prediction with change-oriented behavioural representation, lifespan-conditional attribution, quantitative comparison of attribution profiles between lifespan groups and explicit assessment of explanation stability. Because the selected public dataset (NHANES 2011–2014 wrist accelerometry) is cross-sectional with a one-week monitoring window, we frame the work as *lifespan-informed, age-stratified behavioural prediction* rather than lifelong longitudinal tracking. We evaluate LA-EBPF in a controlled simulation calibrated to the NHANES variable schema, which allows the framework’s methodological behaviour—including whether it recovers known age-specific drivers—to be assessed against a known ground truth. Figure 1 gives an overview of the framework.

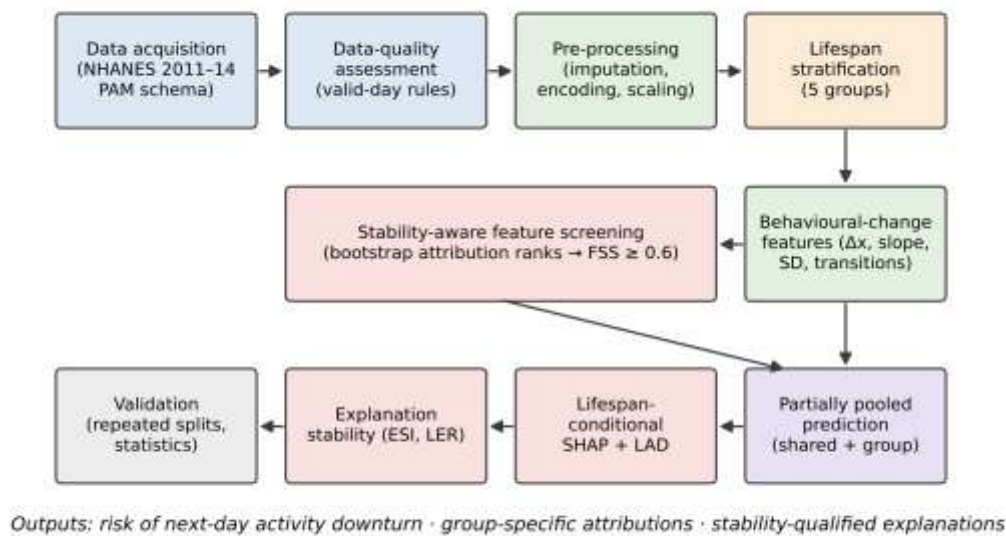


Fig. 1. Overall architecture of the proposed Lifespan-Aware Explainable Behavioral Prediction Framework (LA-EBPF).

The contributions of this article are as follows:

1. A formal, reproducible framework (LA-EBPF) that unifies data-quality screening, lifespan stratification, behavioural-change feature engineering, partially pooled prediction and explanation in a single leakage-controlled pipeline.
2. A lifespan-conditional Shapley attribution procedure that uses group-specific reference distributions, together with a Lifespan Attribution Divergence (LAD) index and a permutation test for comparing attribution profiles across lifespan groups.
3. An explanation-stability layer comprising stability-aware feature screening, a bootstrap Explanation Stability Index (ESI) and a Local Explanation Robustness (LER) score.
4. A transparent simulation-based evaluation against four baselines, with ablation, repeated-split statistical testing and ground-truth driver-recovery analysis, reported with explicit limits on what the simulated evidence supports.

2. RELATED WORK

2.1 Computational behavioural modelling and lifespan analysis

Smartphone and wearable studies have linked passively sensed behaviour to mental health and functioning. StudentLife combined continuous smartphone sensing with surveys in a college cohort [18], and Saeb et al. reported associations between GPS- and phone-usage features and depressive-symptom severity [19]. These studies established feasibility but focused on single life stages and correlational analysis. Population accelerometry provides broader age coverage: NHANES hip-worn data quantified activity across ages [3] and were later used to re-examine age patterns in activity over the lifespan [4], and wrist-worn NHANES 2011–2014 data yielded US age- and sex-specific percentiles in Monitor-Independent Movement Summary (MIMS) units [5, 20]. Life-course and life-span frameworks emphasise that development involves both gains and losses at every age [1, 2], but this insight has rarely been operationalised inside predictive models.

2.2 Machine learning and temporal modelling for behavioural prediction

ML on accelerometry ranges from activity recognition to outcome prediction. Deep convolutional–LSTM architectures improved wearable activity recognition [21], building on recurrent units such as LSTM and GRU [22, 23]. Random forests and hidden Markov models were used to derive sleep and activity phenotypes in UK Biobank [24, 25], and accelerometry-derived features predicted mortality in older NHANES participants [26, 27]. For tabular, moderately sized data, tree ensembles

frequently match or outperform deep networks [28, 29], which motivates gradient-boosted trees [30, 31] as a strong reference. A recent review highlights over-fitting, dimensionality and validation as recurring weaknesses in behavioural ML [10].

2.3 Explainable and responsible behavioural AI

SHAP unifies several additive attribution methods under Shapley-value axioms [11]; TreeSHAP and model-agnostic permutation estimators make it practical [12, 32], while LIME offers local surrogate explanations [33]. XAI has been applied to wearable-derived health prediction, for example SHAP analyses of activity–sleep relations in older NHANES adults [34] and SHAP, partial-dependence and LIME explanations of mortality risk in UK Biobank [35], and to clinical early warning [36]. Surveys stress that explanations must be evaluated quantitatively, including for robustness [37, 38], because attributions can be fragile or manipulable [13–15]. Responsible-AI work further warns of demographic and age-related bias [39–41].

Table 1 summarises representative studies. None of them simultaneously (i) stratifies prediction by lifespan stage, (ii) represents behaviour through explicit change features, (iii) compares attribution profiles between lifespan groups quantitatively and (iv) reports explanation stability. This combination is the research gap that LA-EBPF targets.

Table 1. Comparative analysis of related studies.

Study	Data / population	Age coverage	Task and method	XAI	Lifespan-specific / stability
Wang et al. 2014 [18]	Smartphone sensing, college cohort	Young adults	Associations of sensed behaviour with mental health and performance	No	No / No
Saeb et al. 2015 [19]	Smartphone GPS and usage, small adult sample	Adults	Correlates of depressive-symptom severity	No	No / No
Ordóñez & Roggen 2016 [21]	Benchmark wearable HAR datasets	Not age-focused	Activity recognition, DeepConvLSTM	No	No / No
Varma et al. 2017 [4]	NHANES 2003–2006 hip accelerometry	Childhood to older age	Descriptive age patterns of activity	n/a	Descriptive / n/a
Willetts et al. 2018 [24]	UK Biobank wrist accelerometry	Middle to older age	Sleep/activity phenotypes, RF + HMM	Limited	No / No
Smirnova et al. 2020 [26]	NHANES 2003–2006 accelerometry	Older adults	5-year mortality prediction, ML comparison	Variable importance	No / No
Belcher et al. 2021 [5]	NHANES 2011–2014 wrist MIMS	Childhood to older age	Population-referenced percentiles	n/a	Descriptive / n/a
Cai et al. 2025 [34]	NHANES 2011–2014 wrist accelerometry	Older adults	Sleep duration, RF/GBDT/XGBoost/LightGBM	SHAP	No / No
Graham & Farrell 2025 [35]	UK Biobank wrist accelerometry	Middle to older age	3-year mortality, seven ML models	SHAP, PDP, LIME	No / No

Study	Data / population	Age coverage	Task and method	XAI	Lifespan-specific / stability
This study (LA-EBPF)	NHANES 2011–2014 PAM schema; calibrated simulation	Ages 6–80, five strata	Next-day behavioural-change prediction, partially pooled GBDT	Lifespan-conditional SHAP, LAD	Yes / Yes (ESI, LER)

3. PROBLEM FORMULATION AND RESEARCH HYPOTHESES

3.1 Research problem and questions

Given a short multi-domain behavioural record and static characteristics of a person, the problem is to estimate the probability that the person's next-day behaviour departs downward from his or her own recent baseline, and to explain that estimate in a way that is specific to the person's lifespan stage and stable under resampling. We pose four research questions: **RQ1** Does lifespan-aware, change-oriented modelling improve prediction over conventional pooled models? **RQ2** Which framework components contribute to performance? **RQ3** Do influential behavioural factors differ across lifespan groups? **RQ4** How stable are the resulting explanations?

3.2 Mathematical formulation

Let participant $i \in \{1, \dots, N\}$ be observed on consecutive days $t = 1, \dots, T$ (here $T = 6$). The daily behavioural vector is $\mathbf{x}_i^t \in \mathbb{R}^D$ with $D = 5$ domains (activity volume, sedentary wake time, sleep duration, sleep midpoint and light exposure), so that $X_i = \{\mathbf{x}_i^1, \dots, \mathbf{x}_i^T\}$. Static characteristics are $\mathbf{z}_i$ (age a_i , sex, body-mass index and income-to-poverty ratio) and target-day context is $\mathbf{c}_i^{T+1}$ (weekend indicator and contextual light exposure).

Lifespan representation. A fixed mapping Γ assigns each participant to one of $G = 5$ lifespan groups,

$$g_i = \Gamma(a_i) = k \quad \text{if} \quad b_{k-1} \leq a_i < b_k, \quad k = 1, \dots, G, \quad (1)$$

with boundaries b_k given in Table 3, and $\mathbf{e}_{g_i} \in \{0,1\}^G$ denotes the one-hot group indicator.

Behavioural-change representation. For domain d the change operators are

$$\Delta x_{id}^t = x_{id}^t - x_{id}^{t-1}, \quad \beta_{id} = \frac{\sum_{t=1}^T (t - \bar{t}) x_{id}^t}{\sum_{t=1}^T (t - \bar{t})^2}, \quad s_{id} = [1/T - 1 \sum_t (x_{id}^t - \bar{x}_{id})^2]^{1/2}, \quad (2)$$

namely the last-day difference, the least-squares trend and the within-person variability, where $\bar{x}_{id}$ is the within-person mean. For activity we add a weekend-weekday contrast ω_i and a transition count

$$\rho_i = \sum_{t=2}^T \mathbb{1}[\text{sgn}(x_{i1}^t - m_i) \neq \text{sgn}(x_{i1}^{t-1} - m_i)], \quad (3)$$

where m_i is the person's median activity. The feature map is $\phi(X_i) = [\bar{x}_{id}, s_{id}, x_{id}^T, \Delta x_{id}^T, \beta_{id}]_{d=1}^D \oplus [\omega_i, \rho_i]$, and the candidate input is $\mathbf{u}_i = \phi(X_i) \oplus \mathbf{z}_i \oplus \mathbf{c}_i^{T+1}$ with $|\mathbf{u}_i| = 33$.

Target. With activity as domain $d = 1$, behavioural downturn is defined relative to each person's own baseline,

$$y_i = \mathbb{1}[x_{i1}^{T+1} < \bar{x}_{i1} - \kappa s_{i1}], \quad \kappa = 1.1, \quad (4)$$

which removes between-person differences in absolute activity level that are strongly age-dependent [5].

Prediction function. LA-EBPF combines a shared scorer s_0 , trained on all participants with $\mathbf{e}_g$ as an input, and group-specific scorers s_g in log-odds space:

$$\eta_i = w_{g_i} s_{g_i}(\tilde{\mathbf{u}}_i) + (1 - w_{g_i}) s_0(\tilde{\mathbf{u}}_i, \mathbf{e}_{g_i}), \quad \hat{p}_i = \sigma(\eta_i), \quad \hat{y}_i = \mathbb{1}[\hat{p}_i \geq \tau_{g_i}], \quad (5)$$

so that $\hat{y}_{i,T+1} = f_\theta(X_i, \mathbf{z}_i)$ in the notation of the generic problem. Here $\tilde{\mathbf{u}}_i$ is the stability-screened input (Eq. 9), σ the logistic function, $w_g \in [0,1]$ a pooling weight and τ_g a group-specific decision threshold. Setting $w_g = 0$ recovers a pooled model and $w_g = 1$ fully separate group models, mirroring partial pooling in multilevel modelling [42].

Objective. Each scorer minimises a class-weighted binary cross-entropy with model-specific regularisation Ω :

$$\mathcal{L}(\theta) = -\frac{1}{n} \sum_{i=1}^n \alpha_{y_i} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] + \Omega(\theta), \quad \alpha_c = \frac{n}{2n_c}, \quad (6)$$

where n_c is the number of training cases in class c . The weights $w_g \in \{0, 0.25, 0.5, 0.75, 1\}$ minimise validation log-loss within group g , and τ_g maximises validation F1 within group g .

XAI attribution. For participant i in group g , the Shapley value of feature j with respect to η is

$$\varphi_{ij} = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (p - |S| - 1)!}{p!} [v_i(S \cup \{j\}) - v_i(S)], \quad v_i(S) = \mathbb{E}_{\mathbf{u}' \sim \mathcal{B}_g} [\eta(\mathbf{u}_{i,S}, \mathbf{u}'_S)], \quad (7)$$

where F is the set of p retained features and $\mathcal{B}_g$ is a background sample drawn from the *same* lifespan group in the training data. Using a group-specific reference distribution makes each explanation answer the question “why does this person’s risk differ from that of a typical peer of the same lifespan stage?”. The group attribution profile is $\bar{\mathbf{I}}_{gj} = |\mathcal{J}_g|^{-1} \sum_{i \in \mathcal{J}_g} |\varphi_{ij}|$ and its normalised form is $\boldsymbol{\pi}_g = \bar{\mathbf{I}}_g / \|\bar{\mathbf{I}}_g\|_1$.

Lifespan Attribution Divergence. Differences between groups are quantified with the base-2 Jensen–Shannon divergence [43],

$$\text{LAD}(g, h) = 1/2 \text{KL}(\boldsymbol{\pi}_g \parallel \mathbf{m}) + 1/2 \text{KL}(\boldsymbol{\pi}_h \parallel \mathbf{m}), \quad \mathbf{m} = 1/2 (\boldsymbol{\pi}_g + \boldsymbol{\pi}_h), \quad (8)$$

which lies in $[0,1]$; its null distribution is obtained by permuting group labels.

Explanation stability. Stability-aware screening retains feature j if its feature-stability score across B bootstrap fits satisfies

$$\text{FSS}_j = \frac{1}{B} \sum_{b=1}^B \mathbb{1}[\text{rank}_b(j) \leq k] \geq 0.6, \quad (9)$$

in the spirit of stability selection [44]. Global explanation stability is the mean pairwise Kendall rank correlation [45] between attribution profiles from B independently retrained pipelines, and local robustness is the mean cosine similarity between attributions before and after a small input perturbation:

$$\text{ESI} = \frac{2}{B(B-1)} \sum_{b < b'} \tau_K(\bar{\mathbf{I}}^{(b)}, \bar{\mathbf{I}}^{(b')}), \quad (10a)$$

$$\text{LER} = \frac{1}{n} \sum_i \frac{\boldsymbol{\varphi}_i \cdot \boldsymbol{\varphi}_i^\epsilon}{\|\boldsymbol{\varphi}_i\| \|\boldsymbol{\varphi}_i^\epsilon\|}, \quad \mathbf{u}^\epsilon = \mathbf{u} + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}_u, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.05^2 \mathbf{I}), \quad (10b)$$

where τ_K is Kendall’s rank correlation, $\bar{\mathbf{I}}^{(b)}$ the global attribution profile of pipeline b and $\boldsymbol{\sigma}_u$ the vector of training-set feature standard deviations.

Evaluation metrics. With TP, TN, FP and FN denoting confusion-matrix counts, we report balanced accuracy, precision, recall, F1, ROC-AUC and the Matthews correlation coefficient [46],

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \quad \text{BA} = 1/2 (\text{TP}/\text{TP} + \text{FN} + \text{TN}/\text{TN} + \text{FP}). \quad (11)$$

3.3 Assumptions and hypotheses

We assume that (A1) the five lifespan groups are fixed before modelling and are not tuned to optimise performance; (A2) the six-day observation window precedes the target day, so that no target-day behaviour enters the features; (A3) target-day light exposure acts as an externally available contextual proxy (for example, daylight or weather information); and (A4) Shapley attributions describe *predictive association* within the fitted model, not causal effects. The hypotheses are: **H1** LA-EBPF attains higher ROC-AUC and MCC than each baseline over repeated splits; **H2** removing lifespan stratification or behavioural-change features lowers discrimination; **H3** attribution profiles differ between lifespan groups more than expected under random group assignment; and **H4** stability-aware screening increases explanation stability without degrading prediction.

4. MATERIALS AND METHODS

4.1 Selected public dataset

The National Health and Nutrition Examination Survey (NHANES) is a nationally representative US survey conducted by the National Center for Health Statistics. In the 2011–2012 and 2013–2014 cycles, participants were asked to wear an ActiGraph GT3X+ accelerometer on the non-dominant wrist for seven consecutive days (nine calendar days including partial first and last days), sampling acceleration at 80 Hz [47]. Data were summarised as MIMS units [20], and an open-source machine-learning algorithm classified each minute as wake wear, sleep wear, non-wear or unknown [20, 47]. The public day-level files contain 61,168 day records from 6,917 participants aged ≥ 6 years (2011–2012) and 69,018 day records from 7,776 participants aged ≥ 3 years (2013–2014) [47]. The dataset was selected because it is public, well documented, covers childhood to older adulthood with a common device and protocol, and has been used extensively in accelerometry research [5, 27]. It is, however, cross-sectional: each person contributes a single week, so it supports lifespan-*informed* analysis (between-age comparison of within-week dynamics) rather than lifelong tracking of individuals. Table 2 separates actual dataset properties from the simulated components used in this study.

Table 2. Dataset characteristics and behavioural variables, distinguishing real NHANES properties from simulated components.

Element	Description	Status
Source	NHANES 2011–2012 and 2013–2014, Physical Activity Monitor (PAM) and demographic/examination files [47]	Real dataset property
Device and protocol	ActiGraph GT3X+, non-dominant wrist, 80 Hz, 7 consecutive days (9 calendar days)	Real dataset property
Public day-level files	PAXDAY_G: 61,168 days, 6,917 participants (≥ 6 y); PAXDAY_H: 69,018 days, 7,776 participants (≥ 3 y)	Real dataset property
Activity volume (ACT)	Daily MIMS sum (PAXMTSD), in thousands	Variable defined on real schema
Sleep duration (SLP)	Daily valid sleep-wear minutes (PAXSWMD)	Variable defined on real schema
Sedentary wake time (SED)	Wake-wear minutes below an activity threshold, from minute-level file (PAXMIN)	Derived variable (threshold set by authors)
Sleep midpoint (MID)	Midpoint of the main sleep-wear bout, from minute-level wear predictions	Derived variable
Light exposure (LUX)	$\log_{10}$ of daily lux sum (PAXLXSD)	Variable defined on real schema
Static covariates	Age (RIDAGEYR), sex (RIAGENDR), income-to-poverty ratio (INDFMPIR), BMI (BMXBMI)	Variables defined on real schema

Element	Description	Status
Analytic cohort	N = 11,000 synthetic participants, 7 days each; 33 candidate features; outcome prevalence 24.8%	Simulated
Behavioural mechanisms	Person-level baselines, trends, AR(1) noise, group-specific day-7 drivers (Table 3)	Simulated (author assumptions)

4.2 Simulation-based cohort

To evaluate the framework against a known ground truth, we generated a synthetic cohort of $N = 11,000$ participants on the NHANES PAM day-level schema. Each simulated participant has an age between 6 and 80 years, sex, BMI, income-to-poverty ratio, a weekday of monitoring start and seven consecutive days of five behavioural domains. Daily activity follows a person-specific level, a person-specific linear trend, a group-specific weekend effect and first-order autoregressive noise; sleep, sleep timing, light exposure and sedentary time are generated around person-specific baselines with group-specific variability. Day-7 activity additionally depends on group-specific drivers (Table 3) that encode plausible but *assumed* age-varying mechanisms—for example, weekend and outdoor-light effects in childhood, irregular sleep timing in adolescence and declining weekly trends in older adulthood. These parameters are simulation assumptions informed by the general direction of published lifespan patterns [4–6]; they are not estimates from NHANES and should not be read as empirical findings. The resulting downturn prevalence (Eq. 4) was 24.8% overall, ranging from 15.4% to 37.7% across groups.

4.3 Pre-processing, inclusion and missing data

The protocol specifies, for application to the real files, inclusion of participants aged ≥ 6 years with seven consecutive valid full days (≥ 16 h of valid wear and ≥ 10 h of wake wear per day, using the NHANES wear predictions) and exclusion of days with data-quality flags. Missing covariates (for example, the income-to-poverty ratio) are imputed with training-fold medians plus a missingness indicator, categorical variables are one-hot encoded and continuous inputs are standardised for the linear and recurrent models using training-fold statistics only. Because the simulated cohort was generated complete and already satisfies the inclusion rule, the imputation step is specified but was not exercised in this evaluation.

4.4 Behavioural feature engineering

Five daily domains are derived (Table 2): activity volume (daily MIMS total, in thousands), sedentary wake time (wake-wear minutes below an activity threshold), sleep duration (sleep-wear minutes), sleep midpoint (hours after midnight) and light exposure ($\log_{10}$ daily lux). For each domain, the operators of Eq. (2) produce the mean, variability, last-day value, last-day difference and trend; activity additionally receives the weekend–weekday contrast and transition count (Eq. 3). These features are motivated by the definition of behavioural change itself: a downturn is more likely to follow a declining trend or high day-to-day volatility than a stable routine, irrespective of absolute activity level. Together with four static covariates and two target-day context variables, this yields 33 candidate features (Fig. 2).

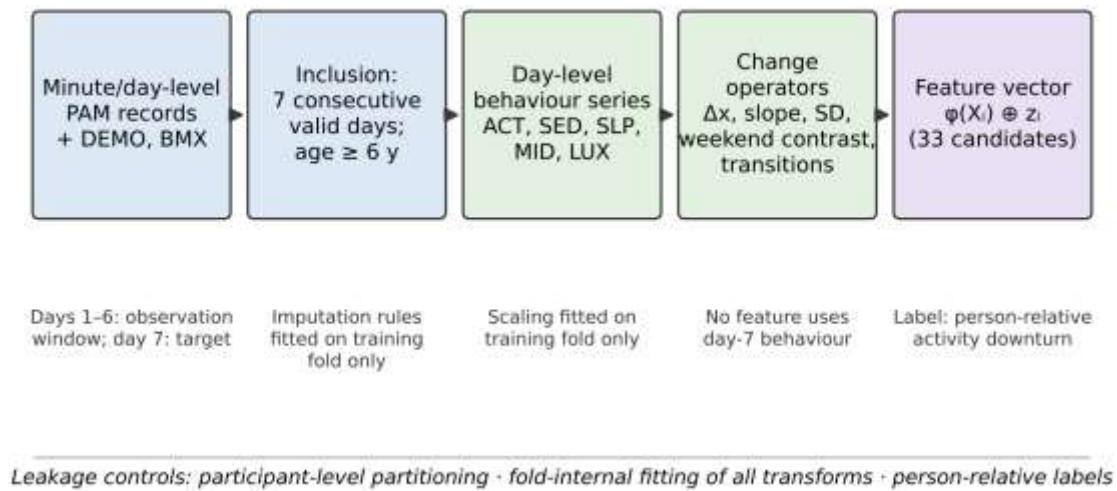


Fig. 2. Data pre-processing and behavioural feature-engineering workflow of LA-EBPF. All transformations are fitted on the training fold only.

4.5 Lifespan stratification

Five groups were defined a priori from developmental stages and the age bands commonly used in population accelerometry reporting [1, 5]: childhood (6–11 years), adolescence (12–19), young adulthood (20–39), middle adulthood (40–59) and older adulthood (60–80; NHANES top-codes age at 80). The boundaries were fixed before any model was fitted (assumption A1).

Table 3. Lifespan group configuration. Boundaries are fixed a priori; sample sizes, prevalences and drivers belong to the simulation.

Group (code)	Age (years)	Developmental rationale	Simulated n	Downturn prevalence (%)	Simulated group-specific drivers
Childhood (CH)	6–11	Primary-school age; externally structured routines	1,482	26.7	Weekend target day; target-day light; last-day activity change
Adolescence (AD)	12–19	Puberty; shift to later sleep timing	1,780	37.7	Last-night sleep midpoint; sleep-timing variability; weekend
Young adulthood (YA)	20–39	Education, work and family formation	2,600	15.4	Last-night sleep duration; mean sedentary time; weekend
Middle adulthood (MA)	40–59	Established occupational routines	2,666	19.9	Mean sedentary time; activity trend; BMI
Older adulthood (OA)	60–80	Retirement transition and ageing	2,472	29.7	Activity trend; activity variability; long last-night sleep

4.6 Data partitioning and leakage prevention

Each of $R = 10$ repetitions used a different random seed to split participants 60/20/20 into training, validation and test sets, stratified jointly by lifespan group and outcome. Splitting is at participant level, so no individual's days appear in more than one partition. Hyper-parameters, pooling weights, thresholds, scalars and screening decisions are all learned on the training and

validation sets, and the test set is used once per repetition, following established guidance on leakage [48] and on reporting prediction models [49].

5. PROPOSED LA-EBPF FRAMEWORK

5.1 Architecture

Figure 1 (Section 1) shows the end-to-end architecture. The data layer acquires the PAM and covariate files and applies valid-day rules. The representation layer performs pre-processing, lifespan stratification and behavioural-change feature engineering. The modelling layer applies stability-aware screening and partially pooled prediction. The explanation layer computes lifespan-conditional Shapley values, group attribution profiles, LAD and stability indices, and the validation layer assesses performance and statistical robustness.

5.2 Lifespan-aware predictive mechanism

The shared scorer s_0 learns behaviour–outcome associations common to all ages while receiving the group indicator, and therefore borrows strength across the whole cohort. Group scorers s_g capture associations specific to each lifespan stage but are trained on fewer participants. The pooling weight w_g of Eq. (5) lets the validation data decide how far each group departs from the shared model; groups with distinctive mechanisms or larger samples receive larger w_g . Group-specific thresholds τ_g compensate for the different outcome prevalences across groups (Fig. 3). Both s_0 and s_g are histogram-based gradient-boosted decision trees [31, 50]; the framework is agnostic to this choice.

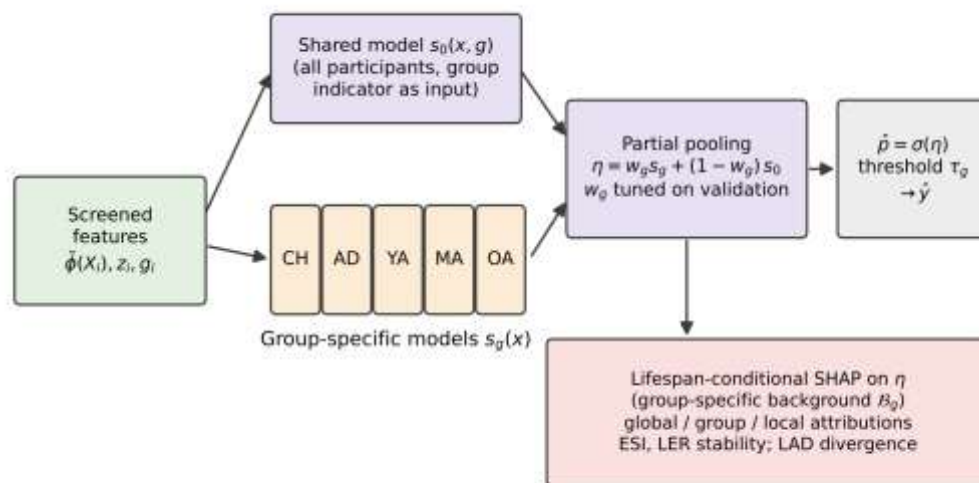


Fig. 3. Lifespan-aware behavioural prediction architecture: shared and group-specific scorers combined by partial pooling, with lifespan-conditional explanation.

5.3 Explainability and explanation-stability mechanisms

Explanations are computed on the combined log-odds η (Eq. 5), so that they describe the deployed predictor rather than a single component. Because LA-EBPF composes several models, we estimate the Shapley values of Eq. (7) with a model-agnostic Monte-Carlo permutation estimator [32], using antithetic permutation pairs and the group-specific background $\mathcal{B}_g$; this targets the same interventional quantity as KernelSHAP [11]. The stability layer acts twice. Before final training, $B = 5$ bootstrap models rank features by mean $|\text{SHAP}|$ and only features with $\text{FSS}_j \geq 0.6$ (Eq. 9, $k = 0.7p$) are retained. After training, ESI and LER (Eqs. 10a–10b) quantify how reproducible the reported explanations are: ESI uses $B = 6$ pipelines retrained on bootstrap resamples of the training set and explained on a fixed reference set of 500 test participants, and LER uses 300 of

these participants. Reported attributions use $M = 12$ permutations (six antithetic pairs) per participant and a background of 200 same-group training participants; the stability analyses use $M = 6$. Algorithm 1 summarises the procedure.

Algorithm 1. LA-EBPF training, prediction and explanation

Input: day-level records $\{X_i\}$, covariates $\{z_i\}$, context $\{c_i\}$, ages $\{a_i\}$; boundaries $\{b_k\}$; B, k, κ .

Output: probabilities $\hat{p}_i$, decisions $\hat{y}_i$, attributions φ_{ij} , profiles π_g , LAD, ESI, LER.

- 1: Apply valid-day rules; assign $g_i \leftarrow \Gamma(a_i)$ (Eq. 1); compute y_i (Eq. 4).
- 2: Build $u_i = \phi(X_i) \oplus z_i \oplus c_i$ using Eqs. (2)–(3) on days $1..T$ only.
- 3: Split participants into train/validation/test, stratified by (g_i, y_i) .
- 4: **for** $b = 1..B$: fit a boosted model on a bootstrap of the training set; rank features by mean $|\varphi|$.
- 5: Retain features with $FSS_j \geq 0.6$ (Eq. 9) to obtain $\tilde{u}$.
- 6: Fit $s_0(\tilde{u}, e_g)$ on all training data, tuning hyper-parameters on validation log-loss.
- 7: **for** $g = 1..G$: fit $s_g(\tilde{u})$ on group g ; choose w_g (validation log-loss) and τ_g (validation F1).
- 8: Predict $\eta_i, \hat{p}_i, \hat{y}_i$ with Eq. (5).
- 9: **for** $g = 1..G$: estimate φ_{ij} (Eq. 7) with background $\mathcal{B}_g$; compute $\bar{I}_g, \pi_g$.
- 10: Compute LAD (Eq. 8) with a label-permutation null; compute ESI and LER (Eqs. 10a–10b).

6. EXPERIMENTAL SETUP

6.1 Simulation protocol and baselines

The simulated cohort was generated once (seed 2026) and evaluated over $R = 10$ repeated participant-level splits (Section 4.6). Four baselines represent common practice. Logistic regression (LR), random forest (RF) [51] and gradient-boosted decision trees (GBDT) receive a conventional representation: the static covariates, target-day context and the raw day-level values of all five domains for days 1–6 (36 inputs), without engineered change features or lifespan stratification (age enters as a continuous covariate). The GRU baseline [23] receives the six-day multivariate sequence (five domains plus a weekend flag), with static and context variables entering a linear output layer; a recurrent model is justified here because the input is an ordered daily sequence. A histogram-based GBDT was used instead of XGBoost [30] to keep the whole experiment within one dependency-controlled environment; XGBoost can be substituted but was not evaluated. All models used class weighting rather than resampling, and every model's decision threshold was chosen on validation data (single global threshold for baselines). Table 4 lists the configurations.

Table 4. Model and experimental configuration.

Model	Input representation	Configuration (search space $\rightarrow$ selection)
LR	Raw day-level values + static + context (36)	L2 penalty, $C \in \{0.1, 1\}$; class-balanced; standardised inputs
RF [51]	Same as LR (36)	300 trees, $\text{max_features} = \sqrt{p}$, $\text{min_samples_leaf} \in \{5, 20\}$; balanced subsample
GBDT [31]	Same as LR (36)	Histogram GBDT; (learning rate, max leaf nodes) $\in \{(0.1, 15), (0.1, 31), (0.05, 15)\}$; ≤ 400 iterations, early stopping (20); $L2 = 1$; class-balanced
GRU [23]	6×6 sequence + 6 static/context inputs	1 layer, 32 hidden units; Adam (lr 0.003), batch 128, $L2 = 1e-4$, ≤ 40 epochs, patience 6; class-weighted loss
LA-EBPF	33 engineered candidates $\rightarrow$ stability-screened subset + group indicator	Shared and group GBDT (same grid; group models inherit shared setting); $w_g \in \{0, 0.25, 0.5, 0.75, 1\}$; τ_g on validation F1; $B = 5$, $k = 0.7p$, $FSS \geq 0.6$

Model	Input representation	Configuration (search space → selection)
All models	Participant-level 60/20/20 split, stratified by group × outcome; R = 10 repetitions	Selection by validation log-loss; threshold by validation F1; seeds fixed

6.2 Validation, metrics and statistical procedure

Balanced accuracy, precision, recall, F1, ROC-AUC and MCC (Eq. 11) were computed on each test set and summarised as mean $\pm$ standard deviation over the ten repetitions. MCC and balanced accuracy were emphasised because the outcome is imbalanced [46]. LA-EBPF was compared with each baseline using (i) the Nadeau–Bengio corrected resampled t -test, which accounts for the overlap of training sets across repetitions [52], with Holm adjustment over the four comparisons per metric; (ii) the exact two-sided Wilcoxon signed-rank test as a non-parametric check [53] (with ten pairs, its smallest attainable p -value is 0.002); and (iii) the paired effect size d_z (mean difference divided by the standard deviation of differences). A participant-level bootstrap with 2,000 resamples [54] provided a 95% confidence interval for the AUC difference between LA-EBPF and GBDT on the first repetition's test set. These analyses quantify variability *across simulation repetitions*; they do not constitute population-level or clinical inference.

6.3 Reproducibility configuration

All experiments ran in Python 3.13.16 with NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1, scikit-learn 1.9.1 and Matplotlib 3.11.2. The GRU was implemented directly in NumPy with back-propagation through time (verified by numerical gradient checking) and trained with Adam, early stopping on validation loss (patience 6, at most 40 epochs). The proposed reproducible simulation configuration was a 2-core x86-64 Linux CPU environment without a GPU; the complete pipeline (ten repetitions, explanation and stability analyses) required approximately 21 minutes. Seeds were fixed for data generation (2026), splitting (1000 + r for repetition r) and every model.

7. Simulation-Based Experimental Results

All numerical values in this section are produced by the controlled simulation described in Sections 4.2–6.3. They demonstrate the expected methodological behaviour of LA-EBPF under the defined simulation configuration and require independent empirical reproduction on real data before any real-world performance claim can be made.

7.1 Overall model comparison

Table 5 and Fig. 4 summarise test performance. LA-EBPF achieved the highest ROC-AUC (0.770 ± 0.013), F1 (0.525 ± 0.017), MCC (0.337 ± 0.023) and balanced accuracy (0.689 ± 0.015). The strongest baselines were GBDT (AUC 0.732 ± 0.013) and RF (AUC 0.729 ± 0.012), which performed almost identically. The GRU (AUC 0.711 ± 0.013) was below the tree ensembles, consistent with prior observations that recurrent networks offer little advantage on short, low-dimensional sequences [28, 29], and LR was weakest (AUC 0.675 ± 0.011). The AUC gain of LA-EBPF over GBDT was 0.038 (MCC gain 0.039). Absolute performance is moderate, as expected for forecasting a single day of behaviour; the person-relative outcome is intrinsically noisy because each baseline is estimated from only six days.

Table 5. Simulation-based predictive performance comparison on held-out test sets (mean $\pm$ SD over ten repeated participant-level splits).

Model	Balanced acc.	Precision	Recall	F1	ROC-AUC	MCC
LR	0.621 ± 0.007	0.342 ± 0.018	0.683 ± 0.084	0.453 ± 0.009	0.675 ± 0.011	0.213 ± 0.011
RF	0.664 ± 0.012	0.421 ± 0.037	0.615 ± 0.062	0.496 ± 0.015	0.729 ± 0.012	0.298 ± 0.027
GBDT	0.664 ± 0.016	0.423 ± 0.031	0.609 ± 0.080	0.495 ± 0.020	0.732 ± 0.013	0.298 ± 0.027

Model	Balanced acc.	Precision	Recall	F1	ROC-AUC	MCC
GRU	0.646 ± 0.013	0.384 ± 0.016	0.629 ± 0.056	0.475 ± 0.015	0.711 ± 0.013	0.258 ± 0.021
LA-EBPF (proposed)	0.689 ± 0.015	0.432 ± 0.021	0.674 ± 0.066	0.525 ± 0.017	0.770 ± 0.013	0.337 ± 0.023

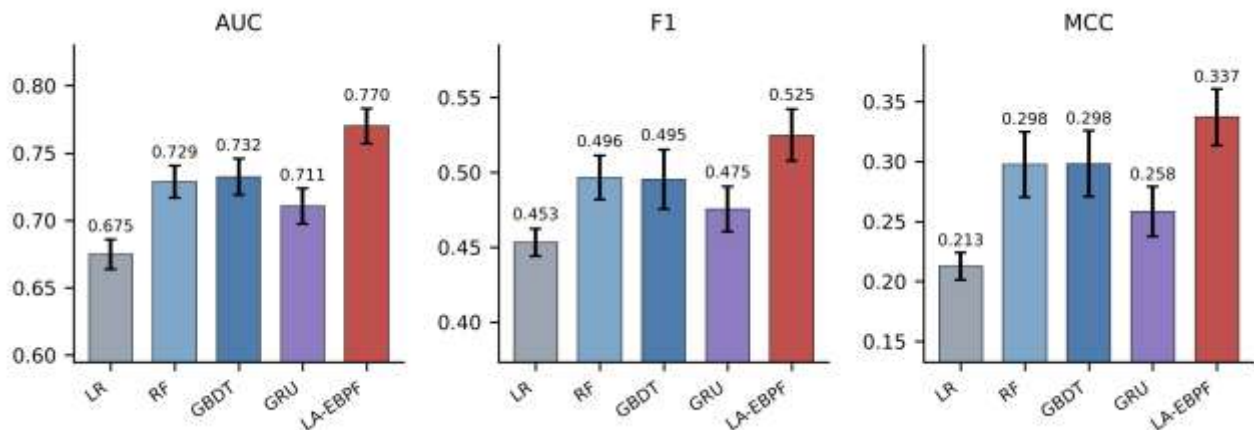


Fig. 4. Simulation-based model performance comparison (mean $\pm$ SD over ten repeated participant-level splits).

7.2 Lifespan-specific performance

Table 6 reports performance within each lifespan group. GBDT discriminated well in childhood and adolescence, where the target-day weekend effect is strong and directly observable, but poorly in young and middle adulthood (AUC 0.631 and 0.623). LA-EBPF improved AUC in every group, and the largest gains occurred in young adulthood (+0.086) and middle adulthood (+0.079), the groups whose simulated drivers (recent sleep, sedentary level, trend and BMI) are least visible in raw day-level values. Gains were smaller in older adulthood (+0.027). Pooling weights selected on validation data averaged 0.57, 0.57, 0.45, 0.42 and 0.72 for the five groups, indicating that every group benefited from a blend of shared and group-specific scorers rather than from either extreme.

Table 6. Lifespan-specific performance (mean over ten repetitions) and lifespan-conditional SHAP analysis (first repetition) under the simulation.

Group	AUC (GBDT / LA-EBPF)	F1 (GBDT / LA-EBPF)	MCC (GBDT / LA-EBPF)	Top-3 LA-EBPF attributions (mean)	SHAP
Childhood (CH)	0.797 / 0.822	0.591 / 0.607	0.418 / 0.461	Target day is weekend; Light exposure, target day; Activity Δ (last day)	3/3
Adolescence (AD)	0.816 / 0.846	0.685 / 0.708	0.461 / 0.524	Target day is weekend; Sleep-timing variability; Sleep midpoint, last night	3/3
Young adulthood (YA)	0.631 / 0.717	0.252 / 0.375	0.117 / 0.245	Target day is weekend; Activity variability (SD); Activity trend (slope)	2/3
Middle adulthood (MA)	0.623 / 0.703	0.324 / 0.423	0.136 / 0.241	Activity variability (SD); BMI; Activity trend (slope)	2/3

Group	AUC (GBDT / LA-EBPF)	F1 (GBDT / LA-EBPF)	MCC (GBDT / LA-EBPF)	Top-3 LA-EBPF attributions (mean)	SHAP
Older adulthood (OA)	0.681 / 0.708	0.503 / 0.519	0.222 / 0.261	Activity trend (slope); Sleep, last night; Activity variability (SD)	3/3

7.3 Ablation analysis

Panel A of Table 7 shows the ablation. Removing behavioural-change features caused the largest deterioration (AUC 0.725 ± 0.011 , MCC 0.256 ± 0.023), confirming that change-oriented representation is the main source of the predictive gain. Removing lifespan stratification reduced AUC from 0.770 to 0.759 but left F1 and MCC essentially unchanged (0.524 and 0.343), so its contribution to threshold-based classification was not demonstrated. Removing stability-aware screening produced indistinguishable accuracy (AUC 0.771) while using all 33 features, compared with a mean of 23.6 retained features in the full framework. Its effect on explanation stability is discussed in Section 8.4.

7.4 Statistical validation

Panel B of Table 7 reports the paired comparisons. For ROC-AUC, LA-EBPF exceeded every baseline in all ten repetitions (Wilcoxon $p = 0.002$ for each comparison; all Holm-adjusted corrected t -test $p < 0.001$), with large paired effects (d_z from 3.3 to 7.6). The bootstrap 95% CI for the AUC difference against GBDT on the first test set was 0.041 (95% CI 0.020 to 0.063). For MCC, the advantage over LR and GRU remained significant after correction, whereas the differences from RF (Holm-adjusted $p = 0.069$) and GBDT (Holm-adjusted $p = 0.141$) were not significant under the conservative corrected t -test, even though LA-EBPF was better in 9 and 8 of ten repetitions, respectively. H1 is therefore supported for discrimination but only partly for threshold-dependent classification.

Table 7. Ablation and statistical validation. Panel A — Ablation of LA-EBPF components (performance: mean $\pm$ SD over ten repetitions; ESI and Jaccard: six bootstrap-retrained pipelines on the first repetition).

Variant	F1	ROC-AUC	MCC	ESI (Kendall τ)	Top-10 Jaccard
w/o lifespan stratification	0.524 ± 0.012	0.759 ± 0.015	0.343 ± 0.022	0.489 ± 0.087	0.543
w/o behavioural-change features	0.479 ± 0.013	0.725 ± 0.011	0.256 ± 0.023	0.677 ± 0.116	0.772
w/o stability-aware screening	0.526 ± 0.011	0.771 ± 0.010	0.336 ± 0.016	0.647 ± 0.072	0.656
Full LA-EBPF	0.525 ± 0.017	0.770 ± 0.013	0.337 ± 0.023	0.562 ± 0.104	0.675

Table 7 (continued). Panel B — Statistical validation of paired differences across the ten simulation repetitions.

Comparison	Metric	Mean Δ	d_z	Corrected t-test p (Holm)	Wilcoxon p (exact)	Wins
LA-EBPF vs LR	AUC	+0.095	7.58	< 0.001	0.002	10/10
LA-EBPF vs LR	MCC	+0.124	4.81	< 0.001	0.002	10/10
LA-EBPF vs RF	AUC	+0.041	4.68	< 0.001	0.002	10/10
LA-EBPF vs RF	MCC	+0.039	1.64	0.069	0.004	9/10
LA-EBPF vs GBDT	AUC	+0.038	3.26	< 0.001	0.002	10/10
LA-EBPF vs GBDT	MCC	+0.039	1.06	0.141	0.014	8/10
LA-EBPF vs GRU	AUC	+0.059	4.07	< 0.001	0.002	10/10
LA-EBPF vs GRU	MCC	+0.079	2.69	0.008	0.002	10/10

8. EXPLAINABILITY AND LIFESPAN-SPECIFIC INTERPRETATION

8.1 Global attribution

Figure 5a ranks features by mean $|\text{SHAP}|$ on the test set of the first repetition. The three most influential features were the activity trend, whether the target day was a weekend and within-person activity variability, followed by the last-day activity change, last-night sleep and BMI. Change-oriented features (trend, variability, last-day difference) (13 of the 25 inputs retained in this repetition) accounted for 56% of total attribution. Global importance is, however, an average over lifespan groups that, as shown next, rely on markedly different information.

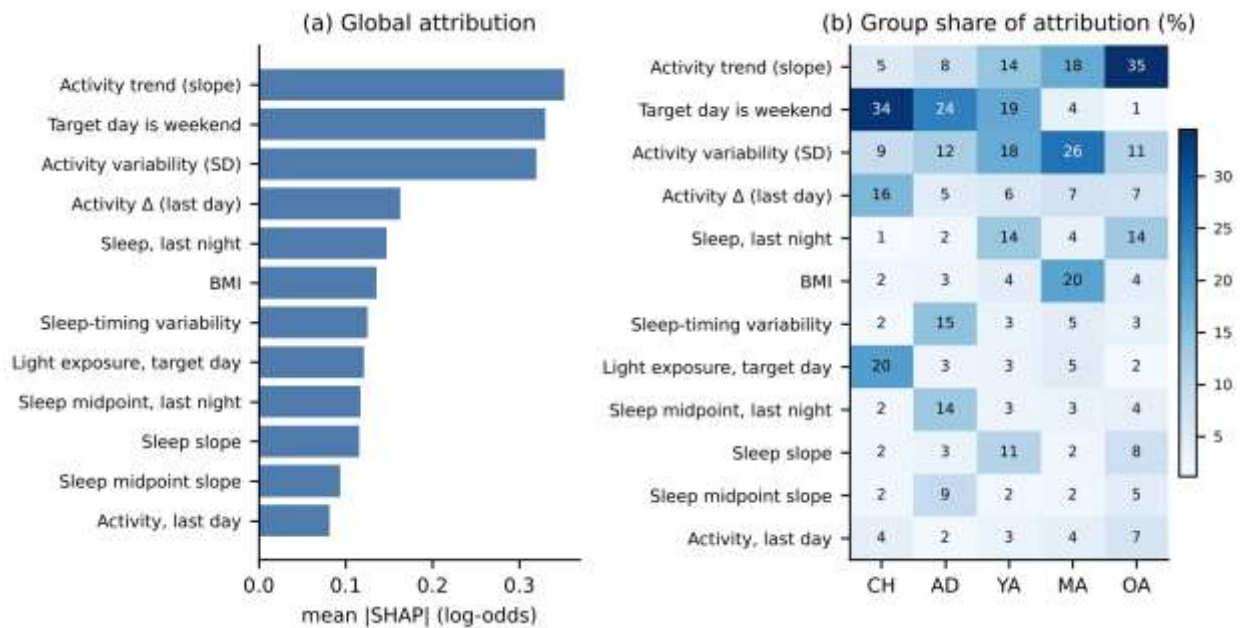


Fig. 5. SHAP-based global and lifespan-specific feature attribution: (a) mean $|\text{SHAP}|$ over all test participants; (b) percentage share of each group's attribution across the 12 globally most influential features.

8.2 Lifespan-specific attribution and behavioural differences

Figure 5b and Table 6 show distinct attribution profiles. In childhood, the weekend indicator, target-day light exposure and last-day activity change dominated; in adolescence, the weekend indicator, sleep-timing variability and last-night sleep midpoint; in young adulthood, the weekend indicator, activity variability and trend together with last-night sleep; in middle adulthood, activity variability, BMI and trend; and in older adulthood, the activity trend, last-night sleep and activity variability. Pairwise LAD values ranged from 0.087 (young vs older adulthood) to 0.295 (childhood vs older adulthood), and the mean LAD (0.171) was two orders of magnitude above the 95th percentile of the label-permutation null (0.0018; $p = 0.005$, 200 permutations), supporting H3. Because the simulation encodes known drivers, recovery can be checked directly: 13 of the 15 group-specific drivers appeared among the top-eight attributions of their group. The two misses—mean sedentary time in young and middle adulthood—are instructive: in the simulated data, sedentary time is strongly collinear with activity volume, and the attribution was absorbed by correlated activity features. This illustrates that Shapley values distribute credit among correlated predictors and must not be read as identifying a unique mechanism.

8.3 Representative local interpretation

Figure 6b–c shows two correctly flagged test participants. For a 13-year-old (predicted probability 0.80), the dominant contribution came from the target day falling on a weekend (+3.20 log-odds), reinforced by high recent activity variability (+0.97); a last-night sleep midpoint earlier than the adolescent reference partially offset the risk (−0.48). For an 80-year-old (predicted probability 0.62), a declining six-day activity trend (slope −0.60 thousand MIMS per day) contributed +1.47 log-odds, with smaller contributions from the last-day change and sedentary trend. These explanations are relative to a typical peer of the same lifespan stage, which is the comparison a lifespan-aware practitioner would ordinarily make.

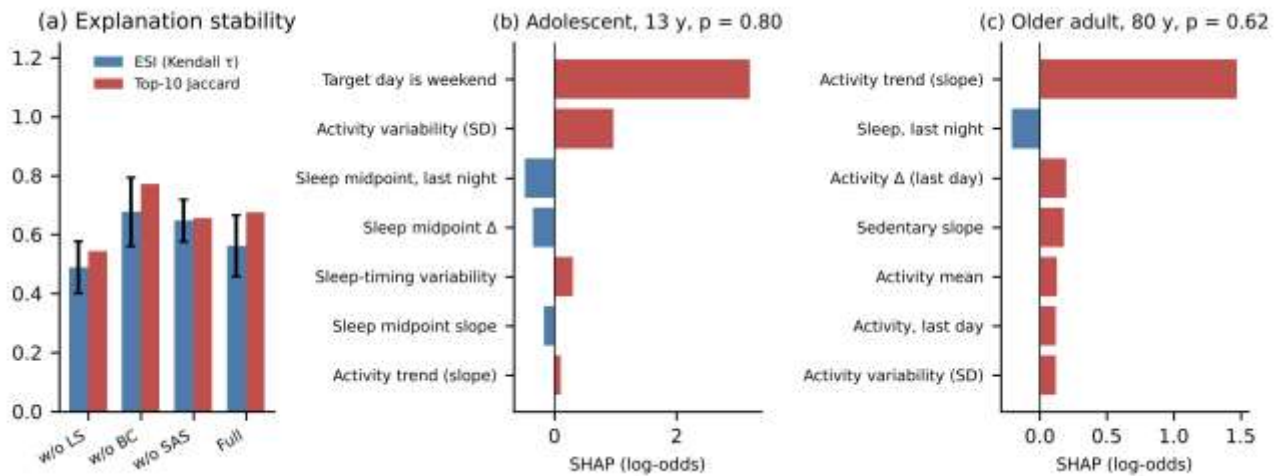


Fig. 6. Explanation stability and behavioural-change interpretation: (a) ESI (mean pairwise Kendall τ , $\pm$ SD) and top-10 Jaccard overlap across six bootstrap-retrained pipelines for each ablation variant (LS, lifespan stratification; BC, behavioural-change features; SAS, stability-aware screening); (b, c) local SHAP explanations for two representative correctly flagged participants.

8.4 Explanation stability

Local explanations were robust: under a 5% feature-SD perturbation, the mean cosine similarity between original and perturbed attribution vectors was 0.979 for LA-EBPF, 0.979 without screening and 0.980 without stratification. Global stability under retraining was lower and more informative (Fig. 6a, Table 7). The full framework reached ESI = 0.562 and a top-10 Jaccard overlap of 0.675; removing stratification reduced both (ESI 0.489, Jaccard 0.543). Contrary to H4, removing stability-aware screening *increased* rank-level ESI to 0.647, while slightly decreasing top-10 overlap (0.656). The explanation is methodological: screening is itself re-estimated in each bootstrap pipeline, so features near the FSS threshold are retained in some pipelines and dropped (zero attribution) in others, which perturbs the full ranking even though the leading features agree more often. The variant without change features had the highest stability (ESI 0.677, Jaccard 0.772), but with markedly lower accuracy, illustrating a trade-off between parsimony of explanation and predictive adequacy. H4 is therefore not supported in its rank-level form, and we report this as a substantive finding rather than a limitation to be hidden.

9. DISCUSSION

Predictive performance. Under the simulation, LA-EBPF improved discrimination by about four AUC points over the strongest conventional baseline, a moderate and credible gain for next-day behavioural forecasting. The ablation localises most of this gain in the behavioural-change representation: trend, variability and last-day change make person-relative downturns easier to anticipate than raw daily values, which tree models struggle to recombine into such summaries. That the GRU, which in principle could learn these summaries from the raw sequence, did not do so with roughly 6,600 training participants is consistent with evidence that engineered representations and tree ensembles remain competitive for small tabular problems [28, 29].

Contribution of lifespan modelling. Stratification contributed modestly to AUC and not measurably to MCC. Its main value in this study lies in explanation rather than accuracy: group-conditional attribution revealed clearly different drivers across groups (LAD far above its null), recovered most simulated drivers and was more stable than explanations from the pooled variant. Pooled models with age as a covariate can represent age interactions in principle, but their single global explanation conflates mechanisms, exactly the problem identified in Section 1. The selected pooling weights (mostly between 0.4 and 0.75) suggest that full separation would waste data, while full pooling would ignore heterogeneity, echoing the rationale for partial pooling in multilevel models [42].

XAI findings in relation to prior literature. The simulated lifespan patterns—weekend sensitivity in youth, sleep-timing irregularity in adolescence and declining trends in older age—were designed to be consistent in direction with published descriptive work [4–6]. Their recovery therefore validates the *procedure*, not the patterns themselves. Prior XAI studies of

wearable data typically report one global SHAP ranking for a single age band [34, 35]; LA-EBPF adds group-conditional reference distributions, a divergence statistic with a null distribution and stability reporting, which addresses recurring concerns about the reliability of post-hoc explanations [13, 14, 38]. The collinearity-induced misses and the screening–stability interaction show why such checks are needed.

Practical implications. In a real deployment, LA-EBPF could support population health programmes or research studies that monitor activity and sleep with consumer or research wearables, for example to flag weeks in which an older adult’s activity trend is declining. The group-specific thresholds and attributions make it possible to communicate *why* a flag was raised in terms appropriate to the person’s life stage. Any such use requires empirical validation, calibration in the target population and human oversight.

Responsible interpretation. All results here derive from a synthetic cohort whose mechanisms were specified by the authors. They show that the framework behaves as intended when age-varying mechanisms exist; they do not show that such mechanisms exist in NHANES or any other population, nor do they estimate the accuracy that LA-EBPF would achieve on real data. Attributions describe associations inside a fitted model and do not establish that changing a feature would change behaviour [16, 17].

10. ETHICAL CONSIDERATIONS AND RESPONSIBLE AI

Privacy and secondary data. NHANES public-use files are de-identified, collected under protocols approved by the NCHS Research Ethics Review Board and released for secondary research; analyses of the real files must respect the NCHS data-use restrictions, including the prohibition on attempts to re-identify participants. The present evaluation used only synthetic data.

Data minimisation. LA-EBPF uses day-level summaries rather than raw signals or location, and stability-aware screening reduced the input set by more than a quarter, which also limits the information that would need to be collected in practice.

Bias and fairness. Age stratification is intended to prevent the needs of smaller or atypical age groups from being averaged away, a form of the age-related bias described as digital ageism [41]. It does not, by itself, guarantee fairness across sex, socioeconomic or racial/ethnic groups [39, 40]; group-wise error rates (for example, equal-opportunity gaps [55]) should be audited in every real application, and Table 6 already shows that error rates differ substantially between lifespan groups. **Explainability and surveillance.** Behavioural prediction from wearables can be misused for surveillance, insurance pricing or employment decisions [56]. LA-EBPF outputs are probabilistic, group-relative indicators of a short-term activity downturn; they cannot determine personality, intent, mental state or future behaviour with certainty and should not be used for consequential decisions about individuals without consent, transparency and human review. **Limits of behavioural prediction.** Even in this favourable simulated setting, recall and precision were moderate; false positives and negatives would be common, and users should be informed accordingly.

11. LIMITATIONS AND FUTURE RESEARCH

The principal limitation is that the evaluation is simulation-based. The synthetic cohort reproduces the structure of the NHANES PAM files but not their real distributions, missingness patterns, survey weights or measurement errors, and the generating mechanisms were chosen by the authors; better-than-baseline performance is partly a consequence of having encoded age-varying effects. Second, the selected dataset is cross-sectional with a single monitoring week, so neither the real data nor the simulation provides genuine lifelong longitudinal tracking; age differences in the framework reflect between-person comparisons, not within-person development. Third, the target-day light-exposure variable was treated as an externally available contextual proxy (assumption A3); in a strictly prospective deployment it would need to be replaced by forecast information. Fourth, the fixed five-group stratification, although literature-based, cannot capture gradual change within groups or cohort effects, and the oldest group includes top-coded ages. Fifth, permutation-based Shapley estimates carry Monte-Carlo error and, as Section 8.2 showed, distribute credit among correlated features. Sixth, the stability-aware screening rule did not improve rank-level ESI, and a consensus-based screening (fixing the retained set across bootstraps) should be investigated. Finally, demographic fairness beyond age was not evaluated.

Future work should (i) apply LA-EBPF to the real NHANES 2011–2014 PAM files with survey-design-aware evaluation; (ii) validate it externally on independent longitudinal cohorts with repeated monitoring waves—such as accelerometry sub-studies of long-running ageing cohorts—so that within-person behavioural change can be modelled across years rather than days; (iii)

replace discrete strata with continuous age-varying coefficients; (iv) assess calibration and subgroup fairness; and (v) compare permutation SHAP with TreeSHAP and other explainers to quantify explainer disagreement.

12. CONCLUSION

This article proposed LA-EBPF, a lifespan-aware explainable framework for predicting short-term behavioural change. The framework integrates behavioural-change feature engineering, literature-based lifespan stratification, partially pooled prediction, lifespan-conditional Shapley attribution, a Lifespan Attribution Divergence index and explicit explanation-stability assessment within a leakage-controlled pipeline built on the public NHANES wrist-accelerometry schema. In a controlled simulation with known age-varying drivers, LA-EBPF improved discrimination over logistic regression, random forest, gradient boosting and GRU baselines (AUC 0.770 vs at most 0.732), with change-oriented features providing most of the gain. Its explanation layer revealed clearly different attribution profiles across lifespan groups, recovered 13 of 15 simulated drivers and exposed two important caveats: credit-sharing among collinear features and a trade-off between feature screening and rank-level explanation stability. These results establish the methodological behaviour of LA-EBPF, not its real-world accuracy. The next step is empirical validation on the real NHANES files and on independent longitudinal datasets, accompanied by calibration, fairness auditing and human-centred evaluation.

Declarations

Ethics approval: Not applicable. This study analysed only synthetic, simulated data and did not involve human participants or identifiable personal data. The NHANES protocols referred to for the variable schema were approved by the NCHS Research Ethics Review Board; any future analysis of the real files will follow the NCHS data-use conditions. [Authors to confirm institutional requirements.]

Consent to participate: Not applicable (no human participants were recruited).

Consent for publication: Not applicable.

Data availability: The NHANES 2011–2012 and 2013–2014 Physical Activity Monitor and demographic files are publicly available from the National Center for Health Statistics (<https://wwwn.cdc.gov/nchs/nhanes/>). The synthetic cohort is fully reproducible from the simulation code and seeds described in Section 6.3.

Code availability: The simulation, modelling and figure-generation code will be made available [authors to insert repository link or “from the corresponding author on reasonable request”].

Competing interests: The authors declare no competing interests. [Authors to confirm.]

Funding: [Authors to state funding, or “This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.”]

Author contributions: [To be completed by the authors using the CRediT taxonomy, e.g., conceptualisation, methodology, software, formal analysis, writing—original draft, writing—review and editing, supervision.]

Acknowledgements: The authors acknowledge the National Center for Health Statistics and the NHANES participants for making the survey data and documentation publicly available. [Authors to add further acknowledgements if any.]

REFERENCES

- [1] Baltes PB (1987) Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. *Dev Psychol* 23(5):611–626. <https://doi.org/10.1037/0012-1649.23.5.611>
- [2] Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C (2003) Life course epidemiology. *J Epidemiol Community Health* 57(10):778–783. <https://doi.org/10.1136/jech.57.10.778>
- [3] Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M (2008) Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc* 40(1):181–188. <https://doi.org/10.1249/mss.0b013e31815a51b3>
- [4] Varma VR, Dey D, Leroux A, Di J, Urbanek J, Xiao L, Zipunnikov V (2017) Re-evaluating the effect of age on physical activity over the lifespan. *Prev Med* 101:102–108. <https://doi.org/10.1016/j.ypmed.2017.05.030>

- [5] Belcher BR, Wolff-Hughes DL, Dooley EE, Staudenmayer J, Berrigan D, Eberhardt MS, Troiano RP (2021) US population-referenced percentiles for wrist-worn accelerometer-derived activity. *Med Sci Sports Exerc* 53(11):2455–2464. <https://doi.org/10.1249/MSS.0000000000002726>
- [6] Ohayon MM, Carskadon MA, Guilleminault C, Vitiello MV (2004) Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep* 27(7):1255–1273. <https://doi.org/10.1093/sleep/27.7.1255>
- [7] Insel TR (2017) Digital phenotyping: technology for a new science of behavior. *JAMA* 318(13):1215–1216. <https://doi.org/10.1001/jama.2017.11295>
- [8] Onnela JP, Rauch SL (2016) Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41(7):1691–1696. <https://doi.org/10.1038/npp.2016.7>
- [9] Mohr DC, Zhang M, Schueller SM (2017) Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol* 13:23–47. <https://doi.org/10.1146/annurev-clinpsy-032816-044949>
- [10] Farrahi V, Rostami M (2024) Machine learning in physical activity, sedentary, and sleep behavior research. *J Act Sedentary Sleep Behav* 3:5. <https://doi.org/10.1186/s44167-024-00045-9>
- [11] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* 30, pp 4765–4774
- [12] Lundberg SM, Erion G, Chen H, et al. (2020) From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2(1):56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- [13] Alvarez-Melis D, Jaakkola TS (2018) On the robustness of interpretability methods. *arXiv:1806.08049*. <https://doi.org/10.48550/arXiv.1806.08049>
- [14] Krishna S, Han T, Gu A, Pombra J, Jabbari S, Wu S, Lakkaraju H (2022) The disagreement problem in explainable machine learning: a practitioner’s perspective. *arXiv:2202.01602*. <https://doi.org/10.48550/arXiv.2202.01602>
- [15] Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp 180–186. <https://doi.org/10.1145/3375627.3375830>
- [16] Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 3(11):e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
- [17] Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [18] Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, Zhou X, Ben-Zeev D, Campbell AT (2014) StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp ’14)*, pp 3–14. <https://doi.org/10.1145/2632048.2632054>
- [19] Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, Mohr DC (2015) Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res* 17(7):e175. <https://doi.org/10.2196/jmir.4273>
- [20] John D, Tang Q, Albinali F, Intille S (2019) An open-source monitor-independent movement summary for accelerometer data processing. *J Meas Phys Behav* 2(4):268–281. <https://doi.org/10.1123/jmpb.2018-0068>
- [21] Ordóñez FJ, Roggen D (2016) Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16(1):115. <https://doi.org/10.3390/s16010115>
- [22] Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [23] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of EMNLP 2014*, pp 1724–1734. <https://doi.org/10.3115/v1/D14-1179>
- [24] Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A (2018) Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Sci Rep* 8:7961. <https://doi.org/10.1038/s41598-018-26174-1>
- [25] Doherty A, Jackson D, Hammerla N, et al. (2017) Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank study. *PLoS One* 12(2):e0169649. <https://doi.org/10.1371/journal.pone.0169649>

- [26] Smirnova E, Leroux A, Cao Q, et al. (2020) The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National Health and Nutritional Examination Survey 2003–2006. *J Gerontol A Biol Sci Med Sci* 75(9):1779–1785. <https://doi.org/10.1093/gerona/glz193>
- [27] Leroux A, Di J, Smirnova E, et al. (2019) Organizing and analyzing the activity data in NHANES. *Stat Biosci* 11(2):262–287. <https://doi.org/10.1007/s12561-018-09229-9>
- [28] Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on typical tabular data? In: *Advances in Neural Information Processing Systems 35 (Datasets and Benchmarks Track)*
- [29] Shwartz-Ziv R, Armon A (2022) Tabular data: deep learning is not all you need. *Inf Fusion* 81:84–90. <https://doi.org/10.1016/j.inffus.2021.11.011>
- [30] Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 785–794. <https://doi.org/10.1145/2939672.2939785>
- [31] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems 30*, pp 3146–3154
- [32] Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst* 41(3):647–665. <https://doi.org/10.1007/s10115-013-0679-x>
- [33] Ribeiro MT, Singh S, Guestrin C (2016) “Why should I trust you?”: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- [34] Cai X, Xian Y, Zhou YX, Liu TY, Zhang XY, Chen Q (2025) Association between accelerometer-measured physical activity volume and sleep duration in older adults: a cross-sectional interpretable machine learning analysis. *Front Public Health* 13:1635020. <https://doi.org/10.3389/fpubh.2025.1635020>
- [35] Graham B, Farrell M (2025) Mortality prediction using data from wearable activity trackers and individual characteristics: an explainable artificial intelligence approach. *Expert Syst Appl* 267:126195. <https://doi.org/10.1016/j.eswa.2024.126195>
- [36] Lundberg SM, Nair B, Vavilala MS, et al. (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2(10):749–760. <https://doi.org/10.1038/s41551-018-0304-0>
- [37] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [38] Nauta M, Trienes J, Pathak S, et al. (2023) From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput Surv* 55(13s):295. <https://doi.org/10.1145/3583558>
- [39] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM Comput Surv* 54(6):115. <https://doi.org/10.1145/3457607>
- [40] Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453. <https://doi.org/10.1126/science.aax2342>
- [41] Chu CH, Nyrupe R, Leslie K, et al. (2022) Digital ageism: challenges and opportunities in artificial intelligence for older adults. *Gerontologist* 62(7):947–955. <https://doi.org/10.1093/geront/gnab167>
- [42] Gelman A, Hill J (2007) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge
- [43] Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 37(1):145–151. <https://doi.org/10.1109/18.61115>
- [44] Meinshausen N, Bühlmann P (2010) Stability selection. *J R Stat Soc Series B Stat Methodol* 72(4):417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- [45] Kendall MG (1938) A new measure of rank correlation. *Biometrika* 30(1–2):81–93. <https://doi.org/10.1093/biomet/30.1-2.81>
- [46] Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6. <https://doi.org/10.1186/s12864-019-6413-7>
- [47] Centers for Disease Control and Prevention, National Center for Health Statistics. *National Health and Nutrition Examination Survey 2011–2012 and 2013–2014: Physical Activity Monitor – Day data documentation, codebook, and*

- frequencies (PAXDAY_G, PAXDAY_H). Hyattsville, MD. <https://wwwn.cdc.gov/nchs/nhanes/> (accessed 5 October 2026)
- [48] Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data* 6(4):15. <https://doi.org/10.1145/2382577.2382579>
- [49] Collins GS, Moons KGM, Dhiman P, et al. (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385:e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [50] Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
- [51] Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- [52] Nadeau C, Bengio Y (2003) Inference for the generalization error. *Mach Learn* 52(3):239–281. <https://doi.org/10.1023/A:1024068626366>
- [53] Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
- [54] Efron B, Tibshirani RJ (1993) *An introduction to the bootstrap*. Chapman & Hall, New York
- [55] Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems* 29, pp 3315–3323
- [56] Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. *Nat Mach Intell* 1(9):389–399. <https://doi.org/10.1038/s42256-019-0088-2>