Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Subject-aware deep learning for forecasting a Mental Health Index from passive smartphone sensing in university students: prediction model development and temporal validation

Desy Iba Ricoida
Institut Teknologi Sepuluh Nopember
Retno Aulia Vinarti
ra_vinarti@its.ac.id

Institut Teknologi Sepuluh Nopember
Ahmad Mukhlason
Institut Teknologi Sepuluh Nopember

Article

Keywords: digital phenotyping, passive sensing, mental health forecasting, personalization, deep learning, data leakage, equivalence testing, ecological momentary assessment, PHQ-9, StudentLife

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11382695/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Subject-aware deep learning for forecasting a Mental Health Index from passive smartphone sensing in university students: prediction model development and temporal validation

Desy Iba Ricoida[1,2], Retno Aulia Vinarti[1,*], Ahmad Muklason[1]

[1]Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

[2]Department of Information Systems, Universitas Multi Data Palembang, Palembang, Indonesia

[*]Corresponding author: ra_vinarti@its.ac.id

Abstract

Background: Deep learning models conditioned on subject identity are increasingly proposed for personalized forecasting of mental health from passive smartphone sensing. Their reported gains are rarely tested under strictly causal evaluation, and data leakage can inflate them. We asked whether identity conditioning improves next-day forecasting of a daily well-being index, and how much common leakage sources inflate apparent performance.

Methods: We analysed 19 students from the public StudentLife dataset (846 participant-days; 682 for development and 164 in a calendar hold-out). A causal, bounded Mental Health Index (MHI, -1 to 1) was derived from ecological momentary assessments relative to each person's own past. Four sequence encoders (gated recurrent unit, temporal convolutional network, Transformer and a TimesFM-inspired patch model) were combined with five identity strategies, trained with 10 random seeds and compared with nine baselines. Additional analyses comprised equivalence testing, validity against the Patient Health Questionnaire-9 (PHQ-9) and a leakage audit that reintroduced four leakage sources, singly and combined.

Results: No model forecast next-day MHI better than the hold-out mean (all hold-out $R^2 < 0$). The lowest error came from an exponentially weighted mean of past MHI (mean absolute error [MAE] 0.481, 95% CI 0.407–0.547); the best sensor-only deep model reached 0.498, versus 0.499 for a zero forecast. Identity conditioning did not improve accuracy (all eight primary comparisons Holm-adjusted $p = 1.00$), and seven of eight comparisons were equivalent within ± 0.027 MAE. The MHI was negatively correlated with PHQ-9 scores (Spearman ρ -0.29 to -0.36 across windows; 95% CI excluding zero in one of four). In the leakage audit, augmenting data before splitting raised the cross-

validated R^2 of a tree ensemble from -0.20 to 0.57, and combining all sources raised that of deep models to 0.08–0.45 while their hold-out R^2 deteriorated to between -0.81 and -0.43.

Conclusions: Under causal evaluation, no sensing model forecast next-day MHI better than the hold-out mean, and subject identity added no accuracy. A single leakage source can make models without predictive value appear strong in cross-validation. We recommend leakage audits and equivalence-based reporting for personalized digital phenotyping models.

Keywords: digital phenotyping; passive sensing; mental health forecasting; personalization; deep learning; data leakage; equivalence testing; ecological momentary assessment; PHQ-9; StudentLife

Background

Depression and anxiety are common among university students [1,2], and many go unrecognised between scheduled contacts with services. Smartphones record behaviour continuously and unobtrusively, and *digital phenotyping*, the quantification of behaviour from personal devices, has been proposed as a way to monitor mental health between those contacts [3,4]. Passive sensing is also being piloted to support mental health services, including in low-resource settings [5]. Studies using mobility, phone use, sleep proxies and social context have reported associations with mood, stress and depressive symptoms [6–9].

Because people differ in both their behavioural baselines and in how behaviour relates to mood, a growing line of work argues for *idiographic* or personalized models rather than a single population model [10,11]. In machine learning this is often implemented by giving the model the identity of the person, as a one-hot vector or a learned embedding [12,13], or by normalising each person's data against their own history [14]. Personalized models of this kind have reported substantial accuracy gains over population models for forecasting mood and stress [14].

Such gains are hard to interpret, first because of data leakage: information from the evaluation data reaches the model during training or model selection [15]. Leakage has been documented across hundreds of machine-learning studies and typically inflates performance [16], and a systematic review of machine-learning prediction of depression from electronic health records found that few studies referred to a formal reporting or assessment standard [17]. Longitudinal sensing data are especially vulnerable, because standard practices such as normalising with full-history statistics, interpolating missing values in both directions, augmenting data before splitting it, or selecting stopping points on ran-

domly sampled rows can each let future or duplicated information into model development [18,19]. Near-duplicate samples of the same person in training and validation can then be matched directly, whether or not the model is told who each person is. The target is a further concern: well-being indices built from ecological momentary assessment (EMA) [20] are often standardised using each person's whole study period, which itself leaks future information into every label, and they are rarely validated against an established clinical instrument.

Small idiographic studies also tend to report a null comparison as "no significant difference", which does not separate the absence of an effect from a lack of precision. Equivalence testing asks instead whether an effect is smaller than a pre-specified practically relevant margin [21,22].

In this study we evaluated subject-aware deep learning for next-day forecasting of mental health from passive smartphone sensing under an evaluation protocol in which every step (label construction, feature imputation, normalisation, augmentation and splitting) uses only past information. We asked three questions. (1) Does conditioning deep sequence models on subject identity improve next-day forecasting of a causally defined well-being index, beyond personal normalisation of the input? (2) Is that index valid against a clinical depression measure? (3) How much do common leakage sources inflate the apparent performance of such models?

To answer these questions we defined a causal, bounded Mental Health Index (MHI) relative to each person's own past and checked it against the Patient Health Questionnaire-9 (PHQ-9) [23]. We compared four compact sequence encoders, one of them modelled on the patch design of the TimesFM foundation model [24], combined with five ways of handling identity, with nine baselines, using an analysis plan fixed in code and ten random seeds. Equivalence tests bound the possible benefit of identity, and a leakage audit reintroduces common leakage sources on the same data and models so that the cross-validated performance each pipeline reports can be set against its performance on the hold-out.

Methods

Study design and reporting

This is a prediction model development and temporal validation study, conducted as a secondary analysis of a public longitudinal dataset of university students. Reporting follows the TRIPOD+AI statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence) [25]; the completed checklist is provided as Additional file 1.

No separate study protocol was published and the study was not registered; the analysis plan, including the primary comparisons, was fixed in the analysis code before hold-out results were computed, and all settings were selected on development data only. Because the dataset is public, the late-term data were not hidden from the authors before the analysis. Patients and members of the public were not involved in the design, conduct or reporting of this study. Figure 1 summarises the design.

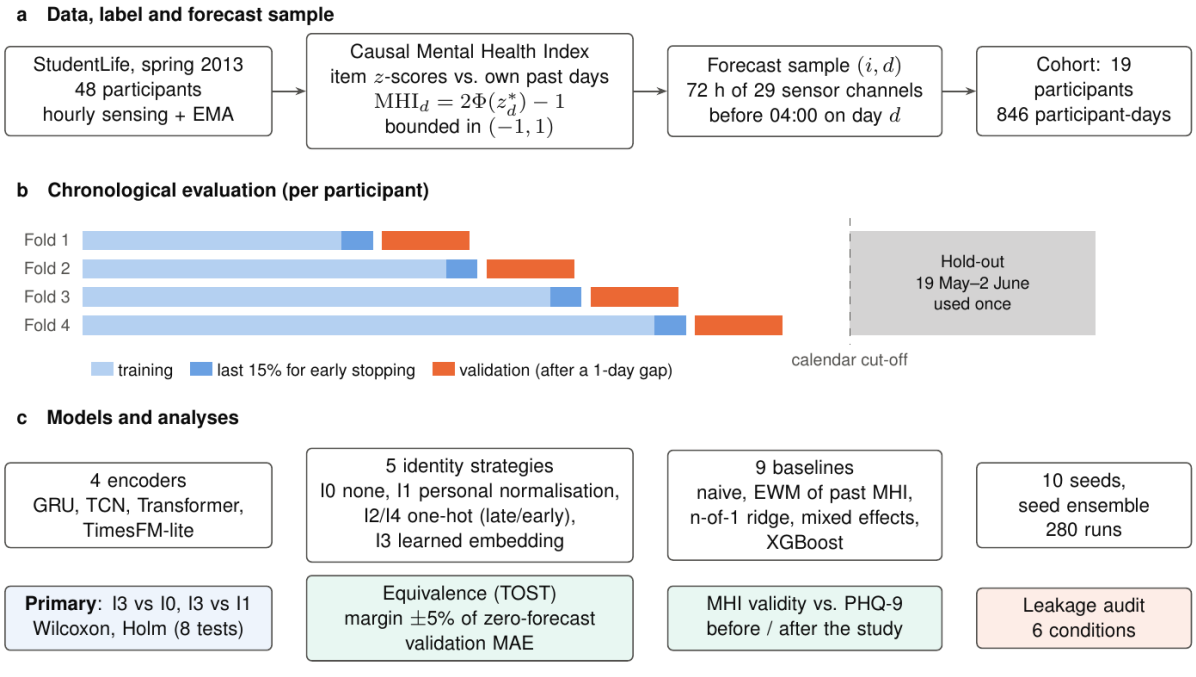


Figure 1: Study design. a Construction of the causal Mental Health Index (MHI) and of forecast samples. **b** Chronological evaluation: a calendar hold-out (19 May 2013 onwards) is used once; model development uses four expanding-window folds per participant with a one-day gap and chronological early stopping. **c** Models and analyses.

Data source and participants

We used the StudentLife dataset [6,26], downloaded from its public copy on Kaggle [27], collected from undergraduate and graduate students at Dartmouth College during the spring term of 2013 (24 March to 5 June). StudentLife was chosen because it is one of the few public datasets that combine continuous multimodal smartphone sensing with frequent, multi-item EMA of mood and stress over a whole academic term, together with PHQ-9 assessments at the start and end of that term. Frequent EMA is required to build a daily index relative to each person's own past, and the PHQ-9 allows that index to be validated; the age of the data is considered in the Limitations. Participants carried Android smartphones that passively recorded physical activity (accelerometer-based activity inference), ambient audio (silence, voice, noise), location (Global Positioning System, GPS), phone lock and unlock events, screen inactivity, running applications and calls, and they answered EMA prompts. We aggre-

gated these streams into hourly summaries and used them together with the EMA responses of the 48 participants with sensing data. Timestamps were stored in Coordinated Universal Time (UTC) and were converted to local time (Eastern Daylight Time, UTC-4).

A forecast day was eligible if at least 72 hours of sensing history were available before it and accelerometer data were present in at least 50% of those hours. Each eligible day of a participant forms one forecast sample, so sample sizes are counted in participant-days rather than calendar days. Demographic characteristics were not part of the data used, so model performance could not be compared across sociodemographic groups and no fairness adjustments were applied. Participants were included if they had at least 20 eligible days before the hold-out cut-off and at least 3 eligible days after it. Of the 48 participants, 46 had at least one eligible day, 28 had at least 20 development days and 19 also had at least 3 hold-out days; these 19 formed the cohort (Table 1).

Outcome: a causal Mental Health Index

Four EMA items were used: net affect (-4 to 4), predicted valence (-1 or 1), valence (-1 to 1) and stress (-3 to 2). Items were rescaled to a common range in which higher values indicate better well-being (net affect divided by 4; stress divided by 2 on its positive side and by 3 on its negative side). Days were defined as behavioural days running from 04:00 to 04:00 local time, so that late-night responses belong to the preceding day. When a person answered several prompts on the same behavioural day, the responses to each item were averaged into one daily value per item. The MHI was derived from EMA responses only; passive sensing data were used exclusively as model inputs and never entered the construction of the outcome.

For each person, item and day d, the daily item mean was converted to a z-score relative to all of that person's previous labelled days (an expanding history from the start of the study, not a fixed window). To stabilise early days, the past mean and variance were shrunk towards an instrument-based prior (mean 0, standard deviation equal to that of the rescaled response options) with a weight of $\kappa = 10$ pseudo-days: with n previous labelled days, the reference mean was $(n\bar{x}_{<d} + \kappa\mu_0)/(n + \kappa)$, and the variance was combined in the same way, so the prior and the person's history carry equal weight after ten labelled days and the history dominates thereafter. The daily composite was the mean of the available item z-scores, and it was re-standardised in the same causal way against the person's previous composites (prior mean 0, variance 1, $\kappa = 10$), giving z_d^*. The index was then

$$\mathrm{MHI}_d = 2\Phi(z_d^*) - 1, \quad \mathrm{MHI}_d \in (-1,1),$$

where Φ is the standard normal distribution function. Positive values indicate a better day than usual for that person, and $\mathrm{MHI}_d = 0$ corresponds to an ordinary day. Because only data up to day d are used and earlier labels are never revised, the label of a day is fixed when it is observed and is not recursive. Days with $z_d^* \leq -1$ ($\mathrm{MHI}_d \leq -0.68$) were labelled below personal baseline (BPB).

Predictors

Hourly sensor data were converted into 29 channels: 19 continuous features (proportions of stationary, walking, running and unknown activity; proportions of silence, voice and noise; log distance travelled, location entropy and log number of GPS points; minutes locked and unlock count; minutes inactive and number of inactive periods; number of running applications; and incoming, outgoing and missed calls and call minutes), 7 binary availability flags (one for each modality except calls, and one for any device data) and 3 time encodings (sine and cosine of hour, weekend); Additional file 2, Table S1 lists every channel with its processing and imputation rule. Activity and audio counts were converted to proportions within each hour. Lock and inactivity durations recorded at the start of a session were carried forward across the following hours, at most 60 minutes per hour. Application counts during hours with no device data were set to missing rather than zero.

Missing values were imputed causally. Activity, audio and GPS features were forward-filled for up to three hours and otherwise replaced by the mean of the same hour of day on previous days. Lock, inactivity and application features were replaced by the median of the same hour on previous days, and calls by zero. The availability flags allowed models to distinguish imputed from observed hours.

The forecast sample for participant i and day d consisted of the 72 hourly vectors preceding 04:00 on day d (a 72×29 matrix); the target was MHI_d. Forecasts therefore use only information available before the behavioural day begins. Each sample thus pairs three days of hourly sensing with a single daily index value, and the windows of consecutive days overlap by 48 hours. A 72-hour window was chosen because it covers three complete day–night cycles, so that sleep-related phone use, mobility and social activity on the preceding nights and days are represented, because effects of behaviour on mood may build up over a few days rather than hours, and because a longer window would have excluded more days through the coverage requirement and would have increased the overlap between consecutive samples. The window length was fixed in advance and was not tuned; other lengths were not examined. The relation between hourly sensing and the daily MHI was not estimated hour by hour: the sequence encoders learned a mapping from the whole 72-hour sequence to the index, whereas the

tabular baselines used daily aggregates of the same channels (see Baselines). Figure 2 illustrates the construction of the samples.

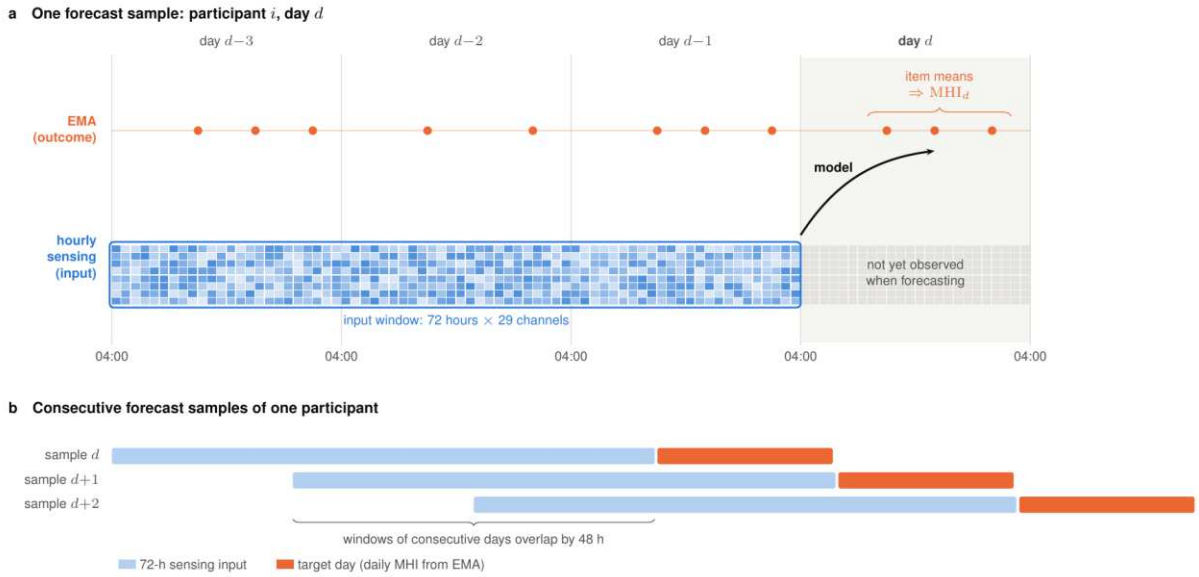


Figure 2: Construction of forecast samples from hourly sensing and daily EMA. a For participant i and day d, the input is the 72×29 matrix of hourly sensing channels ending at 04:00 on day d; the target is MHI_d, computed from the EMA responses given during behavioural day d (04:00 to 04:00). Sensing data never enter the outcome, and hours of day d are not observed when the forecast is made. **b** Each day of a participant forms one sample; windows of consecutive days overlap by 48 hours.

Data partition and model development

Days from 19 May 2013 onwards (last forecast day 2 June) formed a calendar hold-out set that was used once, after all modelling decisions had been fixed. Within the development period, each participant's days were split chronologically: the first 40% formed the initial training set and the remainder was divided into four consecutive validation blocks, giving four expanding-window folds that contain all 19 participants (Figure 1b). Training days within one day of a validation block were removed. In each fold, the last 15% of each participant's training days were used for early stopping (patience 15, maximum 200 epochs). Final models were trained on all development days for the median number of epochs selected across folds. Continuous features were standardised with statistics from the training portion only. Automated tests verified that labels and features of day d do not change when later data are altered, that folds are disjoint and chronological, and that the hold-out never enters development.

Identity strategies

Five ways of handling subject identity were compared. I0 used no identity information. I1 normalised each continuous channel per person, using that person's hours before the target day and shrinking to-

wards cohort training statistics with a weight of 72 hours. I2 to I4 added explicit identity to I1: I2 concatenated a one-hot identity vector to the encoder output (late fusion); I3 concatenated a learned two- or four-dimensional subject embedding with weight decay, analogous to a random effect [12,13] (late fusion); and I4 concatenated the one-hot vector to the input at every hour (early fusion) (Figure 3; a schematic including the past-MHI inputs of variant S1, described below, is given in Additional file 2, Figure S1). As a secondary input variant (S1), an exponentially weighted mean of the person's past MHI values (half-life of 3 labelled days) and the number of days since the last label were added as static inputs; the main analysis used sensors only (S0).

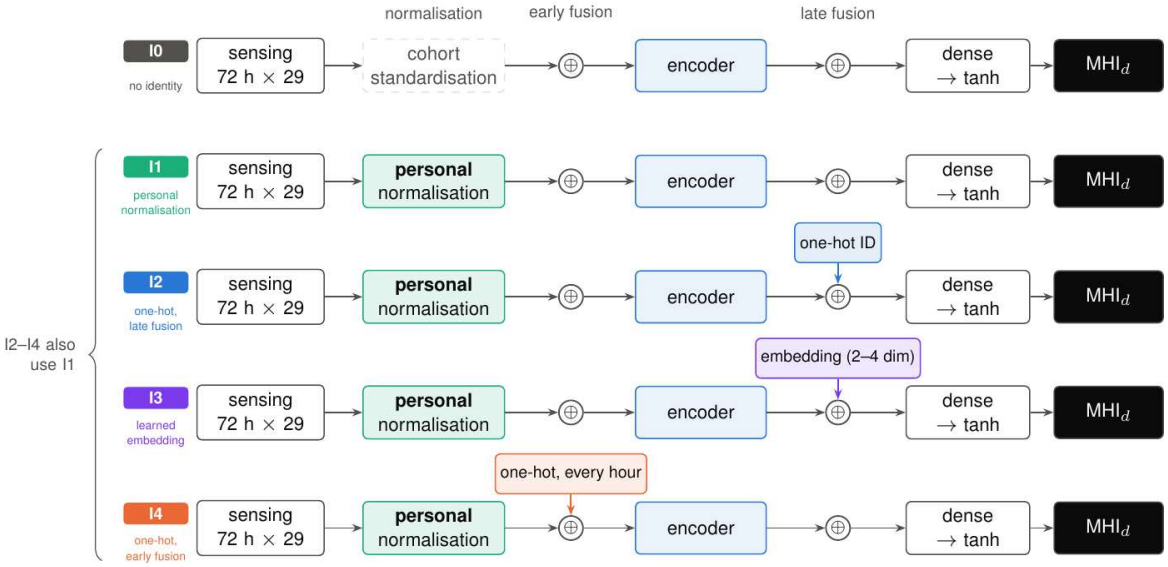


Figure 3: Identity strategies. Each row shows where information about the participant enters the model. I0 standardises inputs with cohort statistics; I1 normalises each channel against the participant's own past hours; I2 and I3 concatenate a fixed one-hot vector or a learned embedding with the encoder output (late fusion); I4 concatenates the one-hot vector with the input at every hour (early fusion). I2 to I4 also use the personal normalisation of I1. Encoder and head are identical across strategies. $\oplus$, concatenation.

Models

Four compact encoders were compared (Figure 4): a single-layer gated recurrent unit (GRU) [28]; a temporal convolutional network (TCN) with three residual causal convolution blocks (dilations 1, 2 and 4) [29]; a Transformer encoder with learned positional embeddings and average pooling [30]; and TimesFM-lite, a small model trained from scratch (without pretraining) that borrows the patch-based design of the TimesFM decoder [24]; it embeds non-overlapping patches of 4 to 12 hours with a residual block and processes them with causal self-attention [31]. Each encoder fed a small dense head with a hyperbolic tangent output, matching the range of the MHI, and was trained with mean squared error and the Adam optimiser. Optional online augmentation (Gaussian jitter of 0.02 or 0.05 standard devia-

tions with ±5% scaling, applied to continuous channels during training only) was treated as a hyperparameter.

Hyperparameters were tuned separately for each encoder with 30 trials of the tree-structured Parzen estimator in Optuna [32], minimising the validation mean absolute error (MAE) under I3 and S0, with models limited to 30 000 trainable parameters. The selected settings were then used for every identity strategy of that encoder, so that identity strategies were compared with the same architecture (Additional file 2, Table S2). The final models had between 1 400 and 19 700 parameters. Each configuration was trained with ten random seeds, and the mean of the ten predictions (seed ensemble) was evaluated; the spread across seeds is reported as a measure of stability. In total, 200 sensor-only runs (4 encoders × 5 identity strategies × 10 seeds) and 80 S1 runs (4 encoders × {I0, I3} × 10 seeds) were trained.

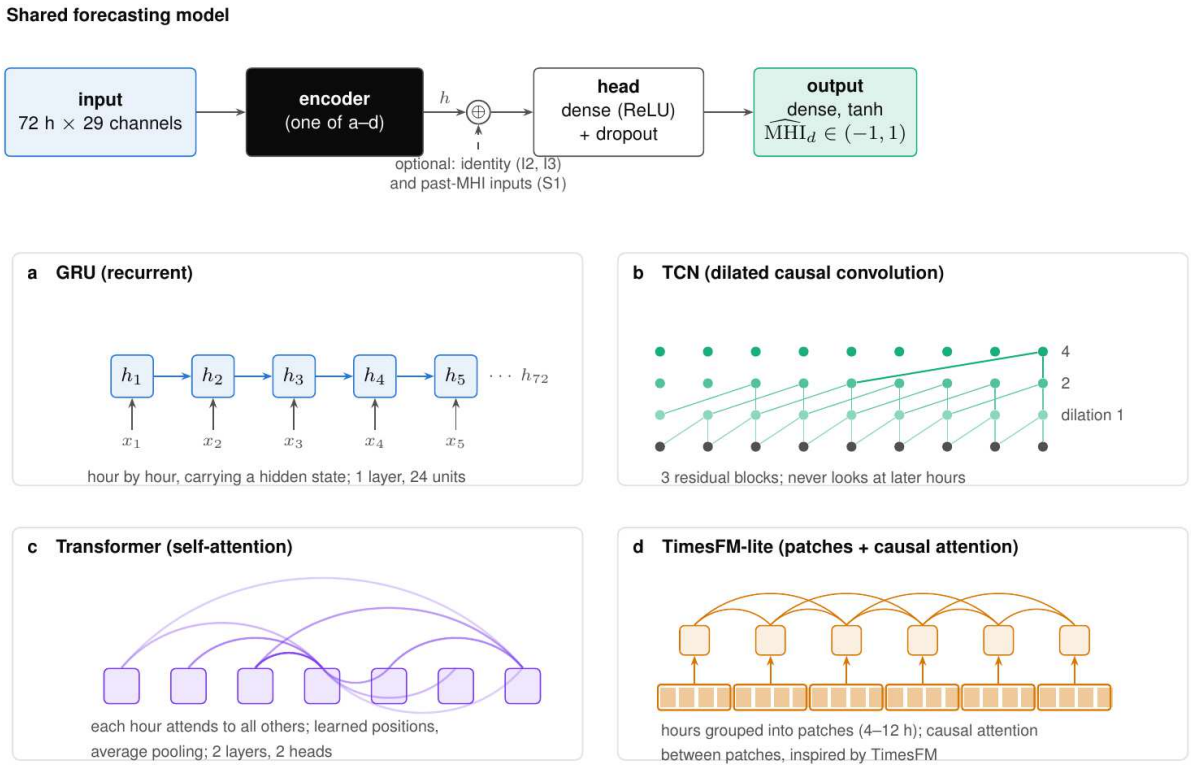


Figure 4: Forecasting model and the four encoders. All models map the 72-hour input to a summary vector h, optionally concatenated with identity (I2, I3) or past-MHI inputs (S1), followed by a dense head with a hyperbolic tangent output. **a** Single-layer gated recurrent unit. **b** Temporal convolutional network with causal convolutions dilated by 1, 2 and 4. **c** Transformer encoder with learned positional embeddings and average pooling. **d** TimesFM-lite: non-overlapping patches of 4 to 12 hours processed with causal self-attention. Panel d shows the range of patch lengths searched; the other sizes are those selected by tuning (selected settings, including a patch length of 4 hours for TimesFM-lite, in Additional file 2, Table S2).

Baselines

Nine baselines were evaluated on the same partition: a zero forecast (an ordinary day for the person); persistence (the last observed MHI); the mean of the person's past MHI values; an exponentially weighted mean of past MHI (EWM, half-life of 3 labelled days); an n-of-1 ridge regression per participant; extreme gradient boosting (XGBoost) [33] with features prepared as in I0, I1 or I2; and a linear mixed-effects model. The ridge and XGBoost models used daily features: means of the 19 continuous channels over the 24 hours before 04:00, data coverage, and night-time (22:00–04:00) phone lock minutes, unlocks, stationary and silence proportions. The mixed-effects model used five personally normalised daily features (night-time lock minutes and stationary proportion, location entropy, voice proportion and coverage) and was specified with a random intercept and a random slope for night-time lock minutes. Because the random-slope model did not converge, a random-intercept model was used, and the coverage feature, which was constant after normalisation, was removed.

Statistical analysis

The analysis plan was fixed before the hold-out was examined. The primary metric was the MAE on the hold-out. Root mean squared error, R^2 and the mean within-person Pearson correlation were secondary. Confidence intervals (CIs) were obtained with a cluster bootstrap over participants (2000 resamples).

Primary analysis. For each encoder, I3 was compared with I0 and with I1 using the per-participant MAE in the hold-out (Wilcoxon signed-rank test [34]), with Holm correction across the eight tests [35]. Using the standard deviation of per-participant I3 minus I0 differences in validation, the smallest difference detectable with 80% power at the most stringent Holm level ($\alpha = 0.05/8$) was 0.017 (averaged over the four encoders in the development folds); this planning value is not used to interpret the hold-out results. Other comparisons (I4 vs I2, I2 vs I3, S1 vs S0, encoders and baselines) were secondary and descriptive.

Equivalence. Because a non-significant result cannot show absence of benefit, we added two one-sided tests (TOST) [21,22] for the eight primary comparisons. The equivalence margin was set to 5% of the zero-forecast MAE in the development folds (± 0.027), a value derived without reference to the hold-out; it corresponds to about 70% of the improvement over the zero forecast that the EWM of past MHI achieved in the development folds (development MAE 0.499 for the EWM versus 0.537 for the zero forecast), so a difference within the margin is smaller than the gain obtained from recent self-reports

alone. We report the 90% confidence interval of the mean per-participant difference, Holm-adjusted TOST p-values and the smallest margin supported by the data. As a rank-based check consistent with the primary Wilcoxon tests, 90% Hodges–Lehmann intervals of the per-participant differences were also compared with the margin. To show how much the forecasts varied, the standard deviation of the seed-ensembled hold-out predictions was computed overall and within participants and compared with that of the observed MHI. This analysis was specified after the primary analysis had been completed and is reported as additional.

Below-personal-baseline days. The ability to flag BPB days was summarised by the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve, the calibration slope, alert rates (alerts on single days and on two consecutive days) and decision curve analysis [36], with alert thresholds set on validation predictions. Whether the benefit of identity (I3 minus I0) differed between participants with baseline PHQ-9 below 5 and at or above 5 (the conventional threshold for mild symptoms) was examined with the Mann–Whitney test.

Validity of the MHI. PHQ-9 questionnaires [23] were completed at the start and the end of the study. For every participant with at least three labelled days, mean MHI over the first two study weeks (24 March to 6 April) and the last two weeks before the end-of-study survey (18 to 31 May) were correlated with the pre- and post-study PHQ-9 totals, respectively, using Spearman's ρ with bootstrap 95% confidence intervals. Windows covering the first and last 14 days of each participant's labelled period served as a sensitivity analysis. Change in MHI between the two windows was related to change in PHQ-9 (for per-participant windows, only participants whose labelled period spanned at least 28 days). A mean of the rescaled EMA items without personal standardisation was analysed for comparison. To examine how much the instrument prior contributes to these associations, the analyses were repeated with prior weights $\kappa = 1$, 5 and 20 and, for $\kappa = 10$, after excluding each person's first 10 labelled days, during which the prior carries at least half the weight. Responses to PHQ-9 item 9 are reported only in aggregate.

Feature contribution. For the deep configuration with the lowest validation MAE, permutation importance was computed on the hold-out for seven feature groups (10 permutations, 10 seeds).

Leakage audit. To quantify how much common pipeline choices inflate apparent performance, four leakage sources were reintroduced into the clean pipeline one at a time and then together, keeping data, cohort, encoders and hyperparameters fixed: (L1) labels standardised with each person's full-period

mean and standard deviation; (L2) linear interpolation of missing sensor values in both directions; (L3) early stopping on a random 20% of training rows; (L4) data augmentation in which each sample was supplemented with two jittered copies (Gaussian jitter of 0.05 standard deviations and $\pm 5\%$ scaling) applied before splitting, followed by a per-participant split by row position in which augmented copies are appended after the original rows, as in a group time-series split; and (L5) all four together. Identity was supplied as a one-hot vector at every hour (early fusion, as in I4 but without personal normalisation). Besides the four deep encoders, an ExtraTrees regressor [37] on window summaries was included as a fast non-deep reference. For every condition we computed the *apparent* performance (the pipeline's own cross-validated R^2, the figure such a pipeline would report) and the *honest* performance (hold-out R^2 with clean features and labels), using five seeds. In additional conditions, run with all five models, identity input was removed (clean, L4 and L5) and augmentation before splitting was followed by a conventional random row-level K-fold split (L6), with and without identity input. The audit was designed after the primary analysis and is reported as additional.

Software and computing

Analyses used Python 3.12 with TensorFlow 2.21 and Keras 3.14, scikit-learn 1.8, XGBoost 3.2, statsmodels 0.15, Optuna 5.0, NumPy 2.4, pandas 3.0 and SciPy 1.17; exact package versions are provided with the code. All models were trained on a laptop with an Intel Core Ultra 9 275HX processor running Windows 11, using the central processing unit only (no graphics processing unit); hyperparameter tuning took about 1.2 hours, the 280 runs of the main experiment about 4.7 hours and the 150 runs of the leakage audit about 1.8 hours and the 125 runs of the additional audit conditions about 3.6 hours.

Use of large language models

The authors used Claude (Anthropic) and Gemini (Google) as assistants for code development and debugging, discussion of analysis options, drafting and language editing, and consistency checks between the manuscript and the analysis outputs. The study design, all analytical decisions and the interpretation of results are the authors' own; the authors reviewed and verified all code, results and text and take full responsibility for the article. No data or results were generated by the language models: all reported numbers were produced by the analysis code provided in Additional file 3, and no generative AI image tools were used.

Results

Participants and the Mental Health Index

The 19 participants contributed 846 participant-days (forecast samples): 682 in development (median 35 per participant, interquartile range [IQR] 31.5–42.5) and 164 in the hold-out (median 10, IQR 5–11; Table 1). Accelerometer coverage of the 72-hour windows was high (median 100%). By construction the MHI was centred on each person's own baseline: only 3% of its variance lay between participants (intraclass correlation 0.03), and participant means ranged from -0.36 to 0.22. BPB days were less frequent in the hold-out (9 of 164, 5.5%, in six participants) than in development (118 of 682, 17.3%).

Table 1: Characteristics of the cohort. BPB, below personal baseline; IQR, interquartile range; MHI, Mental Health Index; PHQ-9, Patient Health Questionnaire-9; SD, standard deviation.

	Overall	Development	Hold-out
Participants, n	19		
Participant-days (forecast samples), n	846	682	164
Participant-days per participant, median (IQR)		35 (31.5–42.5)	10 (5–11)
MHI, mean (SD)	-0.06 (0.59)	-0.10 (0.58)	0.09 (0.57)
Below-personal-baseline days ($z^* \leq -1$), n (%)	127 (15.0)	118 (17.3)	9 (5.5)
Accelerometer coverage of 72-h window, median (IQR)	100 (97–100)%		
PHQ-9 before study, n; mean (SD)	19; 4.3 (3.7)		
PHQ-9 after study, n; mean (SD)	15; 6.8 (5.3)		
PHQ-9 ≥ 10 before / after, n	2 / 4		

Validity against PHQ-9

Across all participants with sufficient EMA in each window, higher MHI was associated with lower PHQ-9 scores in all four level analyses (Figure 5; Additional file 2, Table S3). Spearman correlations were -0.31 before the study ($n = 39$; 95% CI -0.57 to 0.01) and -0.36 after it ($n = 22$; -0.70 to 0.05) with calendar windows, and -0.32 ($n = 41$; -0.57 to -0.03) and -0.29 ($n = 33$; -0.54 to 0.02) with per-participant windows. The non-standardised EMA mean showed correlations of similar size (-0.27 to -0.48; after the study with calendar windows -0.48, 95% CI -0.81 to -0.02). Changes in MHI were not associated with changes in PHQ-9 (ρ between -0.12 and 0.06, all intervals including zero). The associations were weaker with a small prior weight ($\kappa = 1$: ρ -0.12 to -0.29 across the four level analyses; with $\kappa = 20$: -0.29 to -0.40) and, with per-participant windows, after excluding each person's first 10 labelled days (-0.10 before and -0.24 after the study), so part of the

association reflects the instrument prior that dominates early labels (Additional file 2, Table S4). PHQ-9 item 9 was endorsed (score above zero) by 4 of 46 respondents before and 5 of 38 after the study.

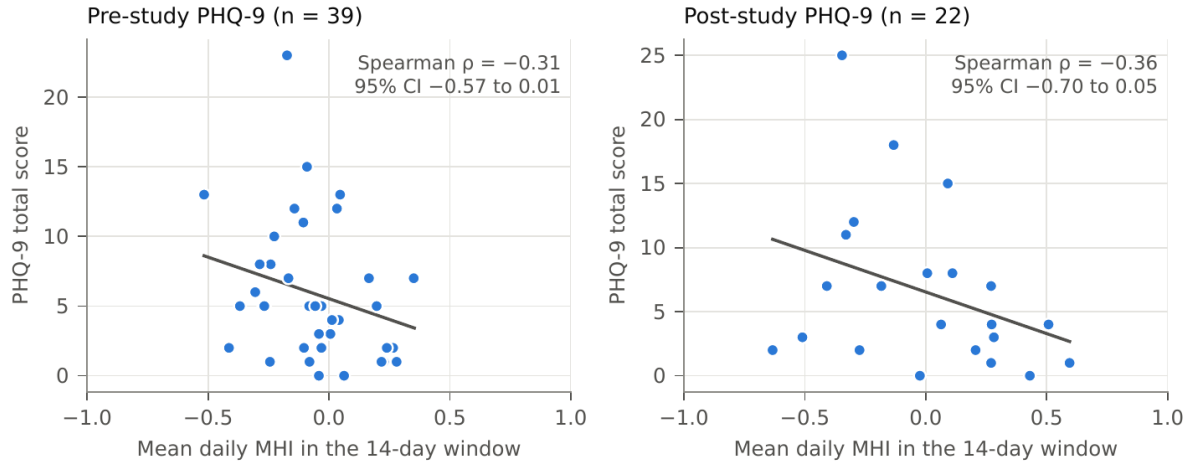


Figure 5: Convergent validity of the Mental Health Index (MHI). Mean MHI in the first two study weeks (left) and the last two weeks before the end-of-study survey (right) against PHQ-9 totals, with least-squares lines for orientation. Correlations are Spearman's ρ with bootstrap 95% confidence intervals.

Forecasting performance

Hyperparameter tuning reached validation MAEs of 0.508 to 0.513 for the four encoders. In the main experiment, the attention-based encoders reached their best validation loss after very few epochs (median early-stopping epoch across folds and seeds 3 for both), compared with 13 for the GRU and 46 for the TCN.

In the hold-out, no model forecast the MHI better than the hold-out mean: every R^2 was negative (Table 2, Figure 6; all configurations in Additional file 2, Tables S5 and S6). The lowest MAE was obtained by the EWM of past MHI (MAE 0.481, 95% CI 0.407–0.547), followed by XGBoost with I2 features (0.488); confidence intervals of all models overlapped widely. The best sensor-only deep configuration (Transformer, I4) reached 0.498 (0.411–0.569), practically the same as the zero forecast (0.499; 0.423–0.562), and the 20 sensor-only configurations ranged from 0.498 to 0.516. For deep models, mean within-person correlations between forecasts and observed MHI did not exceed 0.17, and those of the EWM and personal-mean baselines were negative (-0.23 and -0.20), so their lower MAE reflected the level rather than the day-to-day course of the MHI. The forecasts of the deep models varied little: the standard deviation of their hold-out predictions was 0.03 to 0.15 (0.02 to 0.11 within participants), compared with 0.57 (0.51) for the observed MHI (Additional file 2, Table S7). Mean MHI was higher in the hold-out than in development (0.09 versus -0.10; Table 1), so R^2, which is computed against the hold-out mean, partly reflects this shift. Adding past MHI (S1) improved most

deep models slightly (best: Transformer, I3, MAE 0.491; TimesFM-lite with I3 did not improve); in secondary comparisons the improvement reached nominal significance only for the Transformer with I3 (median per-participant difference -0.008, $p = 0.03$) and approached it for the GRU with I0 (-0.021, $p = 0.06$). The spread across seeds was small (standard deviation of MAE 0.003–0.021).

Table 2: Hold-out mean absolute error (MAE) of the seed-ensembled deep models and of the baselines. Lower is better. I0, no identity; I1, personal normalisation; I2, one-hot late fusion; I3, learned embedding; I4, one-hot early fusion; S0, sensors only; S1, sensors plus past MHI. The lowest deep-model MAE is in bold. R^2 for deep models is the range across the five S0 identity strategies. EWM, exponentially weighted mean.

	Sensors only (S0)					Sensors + past MHI (S1)		
Encoder	I0	I1	I2	I3	I4	I0	I3	R^2 (S0)
GRU	0.516	0.508	0.503	0.505	0.507	0.497	0.493	−0.17 to −0.11
TCN	0.511	0.509	0.504	0.505	0.509	0.505	0.502	−0.16 to −0.12
Transformer	0.502	0.501	0.499	0.501	0.498	0.493	**0.491**	−0.10 to −0.08
TimesFM-lite	0.502	0.503	0.499	0.508	0.500	0.500	0.509	−0.12 to −0.09

Baseline	MAE (95% CI)	R^2
EWM of past MHI	0.481 (0.407 to 0.547)	−0.02
XGBoost, I2	0.488 (0.417 to 0.553)	−0.05
XGBoost, I1	0.491 (0.425 to 0.555)	−0.05
Personal mean	0.496 (0.424 to 0.563)	−0.06
Zero forecast	0.499 (0.423 to 0.562)	−0.03
Mixed effects	0.503 (0.426 to 0.569)	−0.10
XGBoost, I0	0.523 (0.435 to 0.599)	−0.23
Persistence	0.535 (0.433 to 0.634)	−0.66
Ridge (n-of-1)	0.576 (0.486 to 0.671)	−0.71

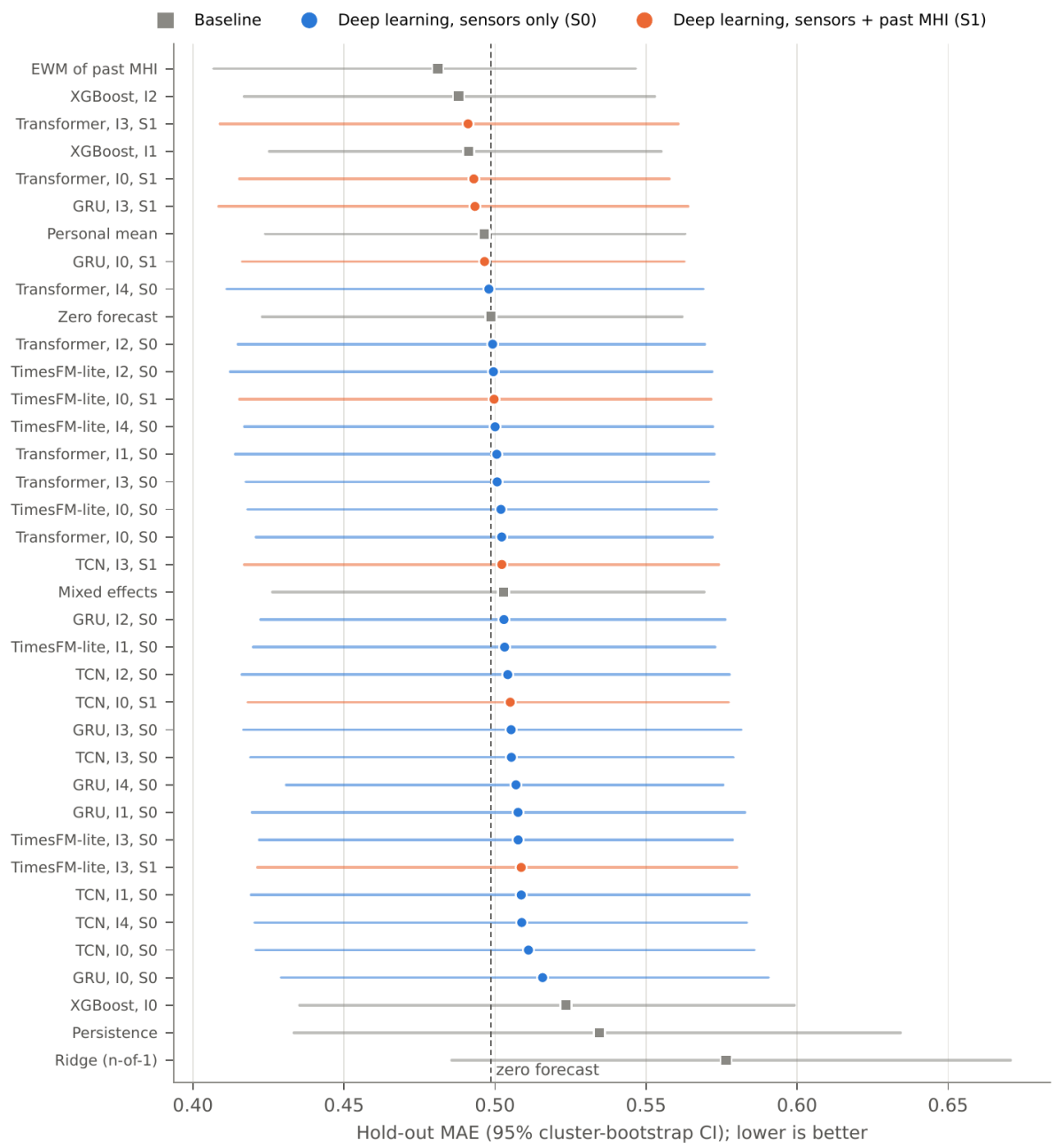


Figure 6: Hold-out forecasting error of all configurations. Points show the MAE of the seed ensemble (deep models) or of the baseline; lines show 95% cluster-bootstrap confidence intervals over participants. The dashed line marks the zero forecast (an ordinary day for each person).

Effect of subject identity

None of the eight primary comparisons was significant (Table 3; Additional file 2, Table S8). Median per-participant differences between I3 and its comparators ranged from -0.009 to $+0.006$ MAE, and all Holm-adjusted p-values were 1.00. In the equivalence analysis, seven of the eight comparisons were statistically equivalent within ± 0.027 MAE (Holm-adjusted TOST $p < 0.01$); the data supported margins as small as 0.005 to 0.016 (Figure 7; Additional file 2, Table S9). The exception was GRU

I3 versus I0, where the 90% interval (-0.031 to 0.006) was compatible with a modest benefit of the embedding as well as with none. With rank-based Hodges–Lehmann intervals, however, all eight comparisons fell within the margin (smallest bounds 0.005 to 0.024; Additional file 2, Table S10). Among secondary comparisons, early and late fusion of one-hot identity did not differ for any encoder, and for TimesFM-lite the one-hot late fusion was more accurate than the learned embedding (median difference -0.006, $p = 0.005$, unadjusted). Among the baselines, XGBoost with personal normalisation (I1) or one-hot identity (I2) had a lower hold-out MAE than with cohort standardisation (I0; 0.491 and 0.488 versus 0.523), whereas in validation the order was reversed (0.546, 0.550 and 0.538). The benefit of identity did not differ between participants with low (< 5, $n = 10$) and higher (≥ 5, $n = 9$) baseline PHQ-9 scores (all $p \geq 0.65$).

Table 3: Primary comparisons of identity strategies. ΔMAE is the per-participant MAE of I3 (learned subject embedding) minus that of the comparator; negative values favour I3. Superiority: Wilcoxon signed-rank test with Holm correction across eight tests. Equivalence: two one-sided tests with margin ± 0.027; the smallest bound is the narrowest margin at which equivalence would be concluded.

		Superiority (Wilcoxon)			Equivalence (TOST)		
Encoder	Comparison	Median ΔMAE	p	p_{Holm}	90% CI of mean Δ	p_{Holm}	Smallest bound
GRU	I3 vs I0	−0.009	0.29	1.00	−0.031 to 0.006	0.105	0.031
GRU	I3 vs I1	−0.001	0.59	1.00	−0.015 to 0.004	0.002	0.015
TCN	I3 vs I0	−0.001	0.98	1.00	−0.016 to 0.011	0.005	0.016
TCN	I3 vs I1	−0.004	0.52	1.00	−0.012 to 0.012	0.002	0.012
Transformer	I3 vs I0	−0.004	0.54	1.00	−0.010 to 0.004	<0.001	0.010
Transformer	I3 vs I1	−0.001	0.98	1.00	−0.005 to 0.005	<0.001	0.005
TimesFM-lite	I3 vs I0	0.006	0.18	1.00	−0.003 to 0.016	0.002	0.016
TimesFM-lite	I3 vs I1	0.005	0.23	1.00	−0.001 to 0.010	<0.001	0.010

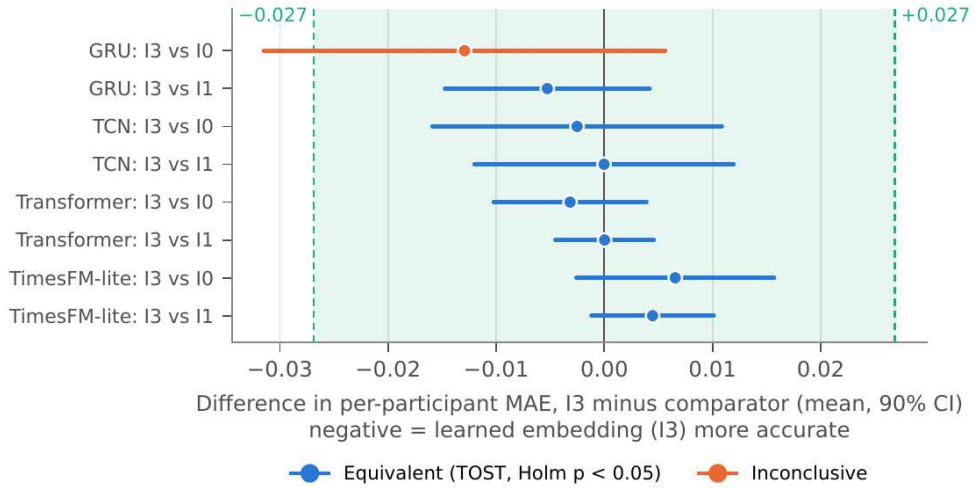


Figure 7: Equivalence of identity strategies. Mean per-participant difference in hold-out MAE between I3 and I0 or I1 with 90% confidence intervals. The shaded band is the equivalence margin (± 0.027 MAE).

Below-personal-baseline days and feature contribution

With only nine BPB days in the hold-out, discrimination was uninformative: AUROC ranged from 0.35 to 0.60 for deep models and from 0.24 to 0.50 for baselines, and no model provided net benefit over alerting everyone or no one at any threshold between 0.05 and 0.50 (Additional file 2, Table S11). In the configuration with the best validation MAE (Transformer, I1), permuting any feature group increased hold-out MAE by at most 0.004 (applications and calls; Additional file 2, Table S12), consistent with the weak signal.

Leakage audit

In the clean pipeline, apparent and honest performance were both poor (R^2 below zero for all five models; Table 4, Figure 8; Additional file 2, Table S13). Standardising labels with full-history statistics (L1), interpolating in both directions (L2) or stopping on random rows (L3) did not raise apparent performance appreciably. Augmenting before splitting (L4) caused 84% of validation rows to be jittered copies of training rows; this alone raised the apparent R^2 of ExtraTrees from -0.20 to 0.57, while its hold-out R^2 remained -0.08. For deep encoders, L4 alone had a smaller effect (-0.02 to 0.10), but combining all sources (L5) raised apparent R^2 to 0.08–0.45 (0.52 for the best seed) and simultaneously made hold-out performance worse than in the clean pipeline (R^2 -0.81 to -0.43 versus -0.25 to -0.12). The largest deterioration occurred with GRU and TimesFM-lite. In the additional conditions, the inflation did not depend on identity. Without identity input, the apparent R^2 of ExtraTrees was 0.57 under L4 and 0.61 under L5, as with identity, and that of the deep encoders was -0.03 to 0.09 under L4 and 0.03 to 0.47 under L5 (-0.02 to 0.10 and 0.08 to 0.45 with identity). Augmentation followed by a random row-level K-fold split (L6), in which 94% of validation rows had a copy in training, gave apparent R^2 of 0.74 for ExtraTrees and 0.26 to 0.74 for the deep encoders (0.72 and 0.15 to 0.75 without identity), whereas hold-out R^2 was -0.09 to -0.08 for ExtraTrees and -0.99 to -0.42 for the deep encoders (Additional file 2, Table S14).

Table 4: Leakage audit. Apparent R^2 is the cross-validated performance reported by the pipeline itself; honest R^2 is performance in the clean hold-out. Values are means over five seeds; for deep encoders the range across the four encoders is shown. Duplicated validation rows: share of validation rows whose participant-day also occurs in the training data.

Condition	Duplicated validation rows	ExtraTrees R^2 Apparent	ExtraTrees R^2 Honest	Deep encoders R^2 (range of 4) Apparent	Deep encoders R^2 (range of 4) Honest
Clean pipeline	0%	−0.20	−0.06	−0.09 to −0.05	−0.25 to −0.12
L1 Full-history label	0%	−0.29	−0.12	−0.20 to −0.06	−0.24 to −0.08
L2 Two-sided interpolation	0%	−0.19	−0.05	−0.15 to −0.06	−0.29 to −0.13
L3 Random early-stopping set	0%	−0.18	−0.06	−0.12 to −0.07	−0.24 to −0.12
L4 Augment-then-split	84%	0.57	−0.08	−0.02 to 0.10	−0.28 to −0.15
L5 All sources	84%	0.61	−0.13	0.08 to 0.45	−0.81 to −0.43

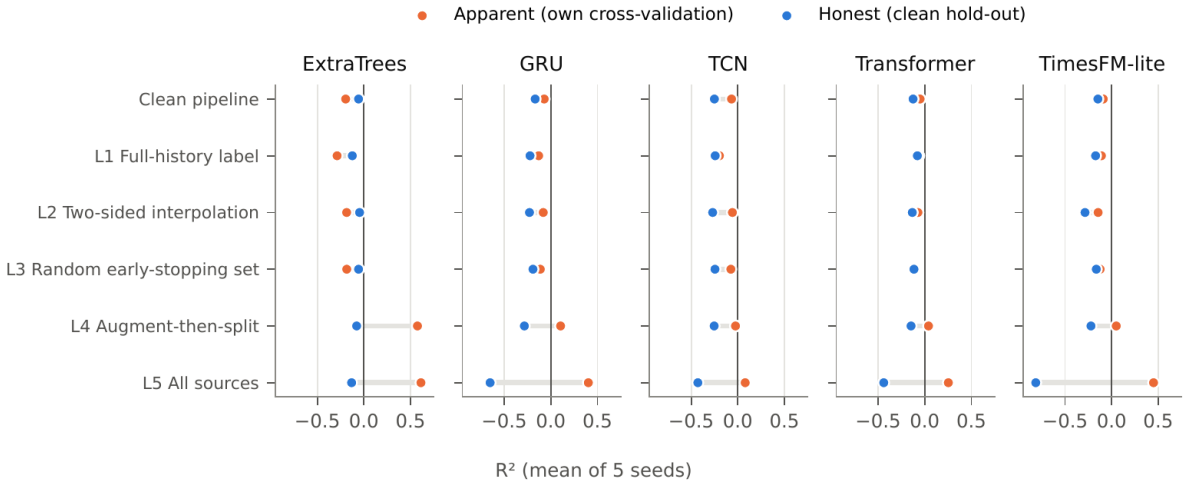


Figure 8: Apparent versus honest performance under each leakage condition. Orange: cross-validated R^2 reported by the pipeline; blue: R^2 in the clean hold-out. L1–L4 reintroduce one leakage source each; L5 combines them.

Discussion

Principal findings

Under an evaluation in which labels, features, normalisation, augmentation and splits used only past information, deep models forecasting a causal, person-relative Mental Health Index from 72 hours of passive smartphone sensing did not forecast better than the hold-out mean or a zero forecast, and conditioning them on subject identity did not improve them. For seven of eight comparisons, equivalence tests excluded differences larger than 5% of the baseline error, although this margin was set after the primary analysis and, given how little signal there was, should be read as bounding the benefit rather than ruling it out. The lowest error, in a comparison that was not tested formally, came from recent self-reports alone. The index was centred on each person and was negatively, though imprecisely, as-

sociated with depressive symptoms measured by the PHQ-9 at the level of between-person differences, but changes in the index were not associated with changes in PHQ-9. Finally, a leakage audit on the same data showed that a single common practice, augmenting before splitting, can produce cross-validated R^2 values near 0.6 from models with no predictive value, and that the combination of leakage sources can make models both look better and perform worse.

Why identity did not help

Identity conditioning is expected to help when people differ systematically in their outcome level or in how predictors relate to it [11,13]. The MHI removes the first source by design: it is defined relative to each person's past, and only 3% of its variance lay between participants. Identity could therefore help only by learning person-specific relations between behaviour and day-to-day deviations in mood, which requires more days per person than the 21 to 49 available here, and a behavioural signal that is present in the first place. Personal normalisation of the inputs (I1), which adjusts for differences in behavioural baselines without extra parameters, performed as well as explicit identity. These results do not show that personalization can never help, but they bound its benefit for this task and data, and because the deep models' forecasts varied little, the equivalence of identity strategies mainly shows that identity did not create predictable signal where there was little to learn.

Limited predictive value of passive sensing

The finding that next-day MHI could not be forecast from sensing better than the hold-out mean is consistent with reports that passive sensors predict daily mood only modestly and inconsistently across individuals [9]. Day-to-day fluctuations in a person-relative index are probably largely transient, the behavioural proxies recorded by 2013-era smartphones are coarse, and labels were sparse (a median of 35 development days per person). The EWM of past MHI and the S1 models had slightly lower errors, which suggests that recent self-reports carry at least as much information about tomorrow as the preceding 72 hours of sensing; these differences were small, were not tested formally, and reflected the level rather than the day-to-day course of the MHI. Low performance was not specific to deep learning: no sensor-based model, including gradient boosting (best MAE 0.488), mixed-effects and n-of-1 models, forecast better than the hold-out mean, and permutation importance showed no feature group of consequence. That the attention-based models stopped after about three epochs also points to little stable signal.

What the leakage audit shows

The audit does not show that any particular published result is inflated. It does show that cross-validated R^2 values usually regarded as good can arise from evaluation choices alone. Augmentation is safe only inside the training portion of each fold: once copies exist before the split, a model can match a validation sample to its near-duplicate in training, and the additional conditions showed that it did so just as well without identity input. The sources also appeared to compound. With all of them combined, deep models looked better in cross-validation and generalised worse, possibly because a randomly sampled early-stopping set also contains copies of training samples and so fails to signal overfitting; the audit did not isolate this interaction. Full-history standardisation of the labels changed apparent performance little here, but it alters what the target means and should still be avoided. For studies of personalized forecasting from longitudinal sensing, we suggest building labels and features causally, splitting chronologically within each person while keeping augmentation, normalisation and early stopping inside the training folds, reporting recent-history baselines such as persistence and exponentially weighted self-reports, using equivalence tests or interval estimates before concluding that a method does not help and, where feasible, reporting a leakage audit.

Implications for decision support

A monitoring service would need forecasts good enough to trigger timely outreach. In this cohort, no model discriminated BPB days, and decision curves showed no net benefit. A system that simply tracked a person's own recent EMA responses would have forecast the next day about as well as any sensing model. Passive sensing may still be useful for longer horizons, for detecting sustained change, or for reducing the burden of self-report, but these uses require separate evaluation.

Strengths and limitations

The pipeline was causal, with its main properties checked by automated tests; the hold-out was used once, the analysis plan was fixed in advance, and forecasts were averaged over ten seeds and compared with a wide range of baselines. The target was assessed against a clinical measure, and the audit linked evaluation choices to inflated performance. The main limitation is the small cohort (19 students from one university). The data were also collected in 2013 with sensing capabilities that have since improved, and EMA was sparse, which limits per-person learning. Only nine BPB days occurred in the hold-out, so alert performance could not be estimated reliably. PHQ-9 was available at two time points, and correlations were moderate with wide intervals; with calendar windows the 95% intervals

for the MHI included zero, so convergent validity is supported but not precisely estimated. A simple mean of the EMA items without personal standardisation correlated with the PHQ-9 at least as strongly as the MHI; this is expected, because the PHQ-9 measures between-person differences in symptom level, which a person-relative index removes by design, but it means that the added value of the MHI lies in tracking within-person change rather than ranking people. The equivalence margin and the leakage audit were specified after the primary analysis, although the margin was derived from development data only. Our encoders were deliberately compact; larger pretrained models were not evaluated. Demographic data were not available, so we could not examine whether performance differed across sociodemographic groups. Because the dataset is public, the hold-out period was not hidden from the authors, although no setting was selected on it. Finally, the audit reproduced generic leakage mechanisms in one pipeline configuration rather than any specific published pipeline, so the size of the inflation it shows is specific to that configuration.

Conclusions

Under causal evaluation, subject-aware deep learning did not improve next-day forecasting of a causal, person-relative Mental Health Index from passive smartphone sensing, and no sensing model forecast better than the hold-out mean. The index was negatively associated with PHQ-9 scores, supporting but not precisely establishing its convergent validity. Because a single leakage source can make models without predictive value appear strong, we suggest that causal pipelines, chronological evaluation, recent-history baselines, equivalence-based reporting and leakage audits be used more widely in personalized digital phenotyping research.

List of abbreviations

AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; BPB, below personal baseline; CI, confidence interval; EMA, ecological momentary assessment; EWM, exponentially weighted mean; GPS, Global Positioning System; GRU, gated recurrent unit; I0–I4, identity strategies (see Methods); IQR, interquartile range; L1–L4, leakage sources, L5, all four combined, and L6, augmentation followed by a random K-fold split (see Methods); MAE, mean absolute error; MHI, Mental Health Index; PHQ-9, Patient Health Questionnaire-9; S0, sensors-only input; S1, sensors plus past-MHI input; TCN, temporal convolutional network; TOST, two one-sided tests; TRIPOD+AI,

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence; UTC, Coordinated Universal Time.

Declarations

Ethics approval and consent to participate

This study is a secondary analysis of the publicly available, de-identified StudentLife dataset [6,27]. The original study was approved by the Institutional Review Board at Dartmouth College, and all participants signed an informed consent form before taking part [6]; the original publication does not report an approval reference number. The present authors used only the anonymised public release and had no access to identifiable information. The authors' institution did not require additional ethics approval for this secondary analysis of publicly available, de-identified data.

Consent for publication

Not applicable.

Availability of data and materials

The analysis in this study used a processed version of the public StudentLife dataset. The original data are available from the dataset authors at https://studentlife.cs.dartmouth.edu/dataset.html and as a public copy on Kaggle at https://www.kaggle.com/datasets/dartweichen/student-life. The processed, analysis-ready dataset derived from it for this study (hourly sensor channels after causal imputation, daily causal Mental Health Index labels, forecast samples with their partition, daily EMA means and PHQ-9 totals, with a data dictionary), the hourly input summaries read by the analysis code, and all analysis code, configuration files and locked package versions are deposited in Zenodo at https://doi.org/10.5281/zenodo.23443540 (data under CC BY 4.0; code under the MIT licence) and are also provided as Additional file 3. During peer review the Zenodo files are access-restricted and available to the editors and reviewers through the private link https://zenodo.org/records/23443540?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjQxOTAxZTNiLTJhYjAtNDgzYy05OTI3LWE2NzEyOWU3YjU1YiIsImRhdGEiOnt9LCJyYW5kb20iOiJjN2E3MTAzNDc0MmFlN2I0ZjE4NjhiMWIzYzEzNjA0ZSJ9.9ULyBGX-0ifQAv7VMabz09m7g539nxd3-SjWxGS0eALoqvWrbgJ7oIojpBFLbnT6LrOrZFYOn7RzSqDPj7esYQ; they will be made publicly

available upon acceptance. Item-level PHQ-9 responses are not included; they are part of the original StudentLife release.

Competing interests

The authors declare no competing interests.

Funding

This work was supported in part by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, as the research organiser and funder, under Grant No. 2202/PKS/ITS/2026.

Authors' contributions

DIR: conceptualization, methodology, software, investigation, data curation, writing (original draft), visualization. RAV: supervision, validation, writing (review and editing). AM: supervision, formal analysis, writing (review and editing). All authors read and approved the final manuscript.

Acknowledgements

This work was partly supported by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia under Grant No. 2202/PKS/ITS/2026. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funder.

References

[1] Auerbach R, Mortier P, Bruffaerts R, Alonso C, Benjet C, Cuijpers P, et al. WHO World Mental Health Surveys International College Student project: prevalence and distribution of mental disorders. Journal of Abnormal Psychology. 2018;127(7):623–638. doi:10.1037/abn0000362.

[2] Beiter R, Nash R, McCrady M, Rhoades D, Linscomb M, Clarahan M, et al. The prevalence and correlates of depression, anxiety, and stress in a sample of college students. Journal of Affective Disorders. 2015;173:90–96. doi:10.1016/j.jad.2014.10.054.

[3] Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017;318(13):1215–1216. doi:10.1001/jama.2017.11295.

[4] Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annual Review of Clinical Psychology. 2017;13:23–47. doi:10.1146/annurev-clinpsy-032816-044949.

[5] Maharjan SM, Poudyal A, van Heerden A, Byanjankar P, Thapa A, Islam C, et al. Passive sensing on mobile devices to improve mental health services with adolescent and young mothers in low-resource settings: the role of families in feasibility and acceptability. BMC Medical Informatics and Decision Making. 2021;21:117. doi:10.1186/s12911-021-01473-2.

[6] Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014. p. 3–14. doi:10.1145/2632048.2632054.

[7] Wang R, Wang W, daSilva A, Huckins J, Kelley W, Heatherton T, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018;2(1):1–26. doi:10.1145/3191775.

[8] Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2015. p. 1293–1304. doi:10.1145/2750858.2805845.

[9] Pratap A, Atkins DC, Renn BN, Tanana MJ, Mooney SJ, Anguera JA, et al. The accuracy of passive phone sensors in predicting daily mood. Depression and Anxiety. 2019;36(1):72–81. doi:10.1002/da.22822.

[10] Molenaar PCM. A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. Measurement: Interdisciplinary Research and Perspectives. 2004;2(4):201–218. doi:10.1207/s15366359mea0204_1.

[11] Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proceedings of the National Academy of Sciences. 2018;115(27):E6106–E6115. doi:10.1073/pnas.1711978115.

[12] Guo C, Berkhahn F. Entity embeddings of categorical variables. arXiv:1604.06737 [Preprint]. 2016. doi:10.48550/arXiv.1604.06737.

[13] Simchoni G, Rosset S. Using random effects to account for high-cardinality categorical features and repeated measures in deep neural networks. In: Advances in Neural Information Processing Systems. 2021. p. 25111–25122.

[14] Taylor S, Jaques N, Nosakhare E, Sano A, Picard R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. IEEE Transactions on Affective Computing. 2020;11(2):200–213. doi:10.1109/TAFFC.2017.2784832.

[15] Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data. 2012;6(4):15. doi:10.1145/2382577.2382579.

[16] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.

[17] Nickson D, Meyer C, Walasek L, Toro C. Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. BMC Medical Informatics and Decision Making. 2023;23:271. doi:10.1186/s12911-023-02341-x.

[18] Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. GigaScience. 2017;6(5):gix019. doi:10.1093/gigascience/gix019.

[19] Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–929. doi:10.1111/ecog.02881.

[20] Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annual Review of Clinical Psychology. 2008;4:1–32. doi:10.1146/annurev.clinpsy.3.022806.091415.

[21] Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics. 1987;15(6):657–680. doi:10.1007/BF01068419.

[22] Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science. 2017;8(4):355–362. doi:10.1177/1948550617697177.

[23] Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine. 2001;16(9):606–613. doi:10.1046/j.1525-1497.2001.016009606.x.

[24] Das A, Kong W, Sen R, Zhou Y. A decoder-only foundation model for time-series forecasting. In: Proceedings of the 41st International Conference on Machine Learning. 2024. p. 10148–10167.

[25] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

[26] StudentLife dataset [Internet]. Hanover (NH): Dartmouth College; 2014 [cited 2026 Oct 7]. Available from: https://studentlife.cs.dartmouth.edu/dataset.html

[27] StudentLife [Internet]. Kaggle; [cited 2026 Oct 7]. Available from: https://www.kaggle.com/datasets/dartweichen/student-life

[28] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014. p. 1724–1734. doi:10.3115/v1/D14-1179.

[29] Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271 [Preprint]. 2018. doi:10.48550/arXiv.1803.01271.

[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017. p. 5998–6008.

[31] Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: long-term forecasting with transformers. In: International Conference on Learning Representations (ICLR). 2023. doi:10.48550/arXiv.2211.14730.

[32] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019. p. 2623–2631. doi:10.1145/3292500.3330701.

[33] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–794. doi:10.1145/2939672.2939785.

[34] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945;1(6):80–83. doi:10.2307/3001968.

[35] Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6(2):65–70.

[36] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565–574. doi:10.1177/0272989X06295361.

[37] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63(1):3–42. doi:10.1007/s10994-006-6226-1.

Additional files

Additional file 1. TRIPOD+AI checklist. Format: PDF. Completed TRIPOD+AI checklist with the location of each item in the manuscript.
Additional file 2. Supplementary tables and figure. Format: PDF. Definitions of the 29 input channels (Table S1), selected hyperparameters (Table S2), validity against PHQ-9 and its sensitivity to the prior weight (Tables S3–S4), hold-out and validation results for every configuration (Tables S5–S6), spread of the forecasts (Table S7), primary and secondary comparisons (Table S8), equivalence tests and a rank-based check (Tables S9–S10), below-personal-baseline alerts (Table S11), feature importance (Table S12), the leakage audit per model and additional audit conditions (Tables S13–S14) and a schematic of the identity fusion points (Figure S1).
Additional file 3. Data and code. Format: ZIP. Analysis-ready dataset derived from StudentLife with data dictionary, hourly input summaries, complete analysis code with locked package versions, and the result tables reported in the article; see the README file in the archive.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile2.pdf
- Additionalfile3.zip