MIXED-EFFECTS MACHINE LEARNING: INTEGRATING RANDOM-EFFECTS INTO REPRESENTATION LEARNING FOR LONGITUDINAL MULTIMODAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal multimodal time-series data present two distinct modeling challenges: (1) learning predictive representations from high-dimensional time series within individual assessments and (2) accounting for participant-specific dependence across repeated assessments. We introduce Mixed-Effects Representation Learning (MERL), which builds on the architecture-agnostic mixedML framework by combining a multimodal CNN–Transformer with a linear mixed-effects model. This formulation separates population-level representation learning from structured participant-specific effects while allowing both components to contribute to prediction. We evaluate MERL using a longitudinal simulation of multimodal physiological signals with known population-level and participant-specific structure, considering both equal and heterogeneous random-effect variances. When both population-level and participant-specific effects contribute to the outcome, MERL achieves the lowest prediction error under both variance settings, with RMSEs of 1.525 and 1.469, compared with 2.470 and 2.230 for the same neural network architecture with one-hot participant identifiers and 6.166 and 6.790 without participant information. MERL remains competitive when either component is removed and estimates negligible random-effect variance when random effects are absent. Under heterogeneous random-effect variances, its estimates also reflect the relative magnitudes of the true variance components. These results demonstrate the potential of combining deep representation learning for multi- modal time series with mixed-effects modeling to exploit repeated-measures structure in longitudinal prediction.

1 INTRODUCTION

Electronic health records (EHRs) provide increasingly rich longitudinal information for modeling individual health trajectories and predicting future clinical outcomes (Tang et al., 2024). Such data are inherently heterogeneous, combining information such as demographics, diagnoses, medications, laboratory measurements and other clinical observations collected across multiple encounters (Shickel et al., 2018). Predicting future health outcomes from these longitudinal histories is important for clinical decision support and has been studied for outcomes ranging from future diagnoses to mortality and heart failure (Carrasco-Ribelles et al., 2023).

Yet structured tabular data alone may not capture all clinically relevant information. Clinical assessment often combines multiple data modalities (Kline et al., 2022), including high-dimensional examination data such as electrocardiography (ECG), photoplethysmography (PPG), respiration, or electromyography (EMG). These physiological recordings contain temporally resolved information that may complement directly collected tabular data such as demographics, lifestyle factors and other clinical measurements. Future health prediction may therefore require models that can jointly use conventional structured covariates and learn predictive representations directly from raw physiological time series.

When such multimodal information is collected repeatedly from the same individuals, two distinct forms of structure arise. Within an assessment, each physiological recording is itself a time series containing temporal patterns that may be predictive of the outcome. Across assessments, repeated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

observations of the same participant introduce longitudinal dependence and participant-specific variation. A single ECG recording, for example, contains within-recording temporal structure, while ECG recordings collected from the same individual across multiple visits additionally form repeated longitudinal observations. These two forms of structure pose different modeling problems: the former requires flexible representation learning from high-dimensional signals, whereas the latter requires accounting for the dependence induced by repeated assessments within participants.

Mixed-effects models provide a statistical framework for such longitudinal and clustered data by decomposing the outcome into fixed and random effects (Winter, 2013). Fixed effects describe population-level relationships shared across participants, whereas random effects represent systematic participant-specific deviations from these relationships. For example, a random intercept can represent differences in participant-specific baseline levels, while random slopes allow the relationship between particular covariates and the outcome to vary across participants.

Deep neural networks and conventional mixed-effects models naturally address different parts of this problem. Neural networks can learn nonlinear representations directly from physiological signals, whereas mixed-effects models explicitly account for participant-specific variation but are not naturally designed to learn representations from raw, high-dimensional multimodal observations. Although recent work has combined machine learning and neural networks with mixed-effects modeling (Xiong et al., 2019; Tran et al., 2020; Simchoni & Rosset, 2023; Mandel et al., 2023; Kilian et al., 2023), comparatively less attention has been given to longitudinal settings requiring both high-dimensional time-series representation learning and explicit repeated-measures modeling.

We study this setting through *Mixed-Effects Representation Learning* (MERL), which builds on the mixedML framework introduced by Kilian et al. (2023). MERL replaces the conventional fixed-effects function with a flexible population-level learner while retaining an explicit statistical random-effects component. We implement the population-level learner as a multimodal convolutional neural network (CNN)–Transformer that learns a nonlinear predictive relationship from physiological time series and scalar covariates. Participant-specific deviations from this population-level relationship are estimated separately through a linear mixed model. This decomposition allows the neural network to focus on representation learning from heterogeneous observations without requiring participant identifiers to be encoded directly in the learned representation.

We evaluate MERL in a controlled longitudinal simulation designed to reflect key characteristics of multimodal health data: scalar covariates should represent directly measurable information such as demographic, lifestyle, or clinical variables, while high-dimensional inputs represent physiological examinations whose predictive information is distributed over a temporal signal. Observed covariates may vary across assessments or remain constant within participants. Participant-specific variation is introduced separately through known random-effects.

Simulation is particularly useful for evaluating this setting. Access to longitudinal patient-level data containing rich multimodal measurements is constrained by privacy, governance and data-sharing considerations and limited availability of public longitudinal EHR datasets remains a barrier to reproducible research (Carrasco-Ribelles et al., 2023). More importantly, real-world data do not explicitly provide the true population function, participant-specific random-effects, or their variance components. By controlling these quantities, our simulation allows predictive performance to be evaluated together with recovery of the underlying statistical structure.

Our contributions are threefold:

1. We formulate longitudinal multimodal prediction around two distinct forms of structure: *within-assessment temporal representation learning* from high-dimensional signals and *between-assessment participant-specific variation* across repeated observations.

2. We introduce MERL for this setting, combining a multimodal CNN–Transformer population-level learner with explicit statistical random effects to model these two forms of structure separately within a unified predictive framework.

3. We develop a controlled longitudinal biosignal simulation with known population-level and participant-specific structure, enabling evaluation of future-assessment prediction, behavior under limiting fixed-effects-only and random-effects-only cases and recovery of random-effect variance components.

2 Related Work

2.1 Mixed-effects models for longitudinal data

Mixed-effects models provide a principled framework for modeling repeated and clustered observations by decomposing outcomes into population-level fixed-effects and cluster- or participant-specific random-effects (Laird & Ware, 1982). Random intercepts capture systematic differences between participants or other clusters, while random slopes allow relationships between covariates and outcomes to vary across participants. This explicit decomposition accounts for within-participant dependence while retaining interpretable population- and participant-level components. Conventional formulations, however, typically require the relationship between predictors and outcomes to be specified a priori. Nonlinear mixed-effects models relax linearity assumptions (Lindstrom & Bates, 1990), but still require a specified functional form and do not themselves provide the representation-learning machinery needed to extract predictive features directly from raw, high-dimensional observations.

2.2 Representation learning from physiological signals

Deep learning instead provides a mechanism for learning task-relevant representations directly from physiological time series. CNNs and recurrent architectures have been widely applied to ECG, EMG and related biosignals (Faust et al., 2018; Rim et al., 2020). Convolutional architectures are well suited to physiological waveforms because they learn local temporal patterns directly from the raw signals and progressively combine them into more complex representations. Transformer architectures provide a complementary attention-based mechanism for modeling interactions across the encoded sequence and have increasingly been adopted for biosignal analysis (Anwar et al., 2025). In multimodal settings, modality-specific encoders can additionally learn representations appropriate to heterogeneous physiological signals before their information is fused for prediction.

These approaches primarily concern representation learning from individual observations. Repeated observations introduce a separate modeling question: how should participant-specific variation be represented? Neural network predictors can be designed either without participant identifiers or to incorporate them directly, for example through one-hot encodings or learned embeddings. Incorporating participant identifiers allows participant-specific information to influence the learned prediction function, but does not impose an explicit fixed- and random-effects decomposition. This distinction motivates approaches that combine neural network representation learning with statistical mixed-effects modeling.

2.3 Machine learning with random-effects

Early combinations of nonlinear machine learning and random-effects were predominantly tree-based. RE–EM trees alternate between regression-tree estimation and estimation of random-effects for longitudinal and clustered data (Sela & Simonoff, 2012), while mixed-effects random forests replace the tree component with a random forest to obtain a more flexible nonlinear population-level predictor (Hajjem et al., 2014). These approaches established the general principle that a nonlinear machine-learning function can replace the conventional parametric fixed-effects part in a mixed-effects model while retaining an explicit random-effects component.

Neural mixed-effects models extend this principle to learned representations, but differ substantially in how the neural network and statistical components are coupled. MeNets apply fixed and participant-specific random-effects to a shared deep representation, effectively allowing subject-specific intercepts and slopes over learned features and demonstrate this formulation for repeated person-specific image observations (Xiong et al., 2019). DeepGLMM combines representations learned by a deep feed-forward network with generalized linear mixed models and performs inference using Bayesian variational methods (Tran et al., 2020). LMMNN incorporates the covariance structure induced by random-effects directly into a neural network objective through a mixed-model negative log-likelihood and explicitly considers repeated measurements and high-cardinality group identifiers (Simchoni & Rosset, 2021; 2023). In longitudinal health prediction, Mandel et al. (2023) replace the linear fixed-effects predictor of a generalized linear mixed model with a feed-forward neural network and apply the resulting model to longitudinal mobile-health data.

These methods demonstrate that neural network representation learning and random-effects modeling are compatible, but make different choices about how the neural network and participant-specific components are coupled. In particular, some approaches place random-effects directly on learned features or incorporate the covariance structure induced by random-effects into neural network optimization. This differs from approaches that maintain separate estimation of the predictive fixed-effects function and the statistical random-effects component.

2.4 FROM MIXEDML TO MERL

The closest methodological foundation for MERL is mixedML (Kilian et al., 2023). Rather than prescribing a particular predictive architecture, mixedML treats a supervised machine-learning model as a flexible fixed-effects learner and alternates its estimation with that of a statistical mixed-effects model. This architecture-agnostic formulation allows different supervised regression models to be combined with explicit mixed-effects modeling.

MERL implements this framework for longitudinal multimodal data, where the fixed-effects learner must extract predictive representations from high-dimensional physiological time series. Specifically, we use a multimodal CNN–Transformer to learn the population-level relationship from physiological time series and scalar covariates, while a separate statistical random-effects component captures participant-specific deviations across repeated assessments. Although estimation of the two components is coupled through the alternating procedure, their roles remain explicitly separated: the neural network models population-level structure, whereas the linear mixed model represents structured participant-specific deviations.

This formulation differs from approaches such as MeNets, in which random effects operate directly on learned representations and from approaches that incorporate the mixed-effects covariance structure into the neural network objective. In MERL, high-dimensional temporal representation learning and longitudinal participant-specific variation remain distinct modeling components.

3 METHODS

3.1 PROBLEM FORMULATION

Let $i = 1, \ldots, N$ index participants and $t = 1, \ldots, T$ index assessments. Each observation consists of multimodal biosignal recordings $\boldsymbol{S}_{it}$, additional scalar covariates $\boldsymbol{c}_{it}$ and an outcome y_{it}. We model the outcome as

$$y_{it} = f_\theta(\boldsymbol{S}_{it}, \boldsymbol{c}_{it}) + \boldsymbol{q}_{it}^\top \mathbf{b}_i + \epsilon_{it}, \tag{1}$$

where f_θ represents the population-level fixed-effects function, $\boldsymbol{q}_{it}$ is the random-effects design vector, $\mathbf{b}_i$ contains participant-specific random-effects and ϵ_{it} denotes observation-level noise. The first term captures predictive relationships shared across participants, while the second represents systematic participant-specific deviations from this population-level relationship.

3.2 SIMULATED DATA

We construct a controlled longitudinal simulation in which population-level relationships and participant-specific random-effects can be varied independently. We simulate $N = 1{,}000$ participants observed over $T = 20$ assessments. At each assessment, we generate five biosignal modalities: ECG, PPG, respiration and two EMG signals. Physiologically motivated synthetic biosignals are generated using NeuroKit2 (Makowski et al., 2021). Each signal is sampled at 100 Hz for 20 seconds, resulting in 2,000 samples per biosignal. The signals are generated independently, with each modality determined by its own generative feature. Thus, the simulation does not explicitly impose cross-modal dependence between the biosignals.

The simulation is based on seven time-varying generative features $\boldsymbol{x}_{it} = (x_{it,0}, \ldots, x_{it,6})$ and three time-stable generative features $\boldsymbol{x}_{ci} = (x_{ci,0}, x_{ci,1}, x_{ci,2})$. The features determine biosignal characteristics, enter the outcome-generating process, or serve as scalar model inputs and the time-varying features exhibit heterogeneous temporal dependence. Their roles and generating processes are detailed in Appendix A and Table 3.

The random-effects component contains one participant-specific random intercept and three random slopes associated with $x_{it,4}$, $x_{it,5}$ and $x_{it,6}$. To evaluate MERL under different forms and magnitudes of participant-specific variation, we consider two random-effect variance settings. In the *equal-variance* setting, all four random effects have marginal variance

$$\sigma^2_{b_0} = \sigma^2_{b_4} = \sigma^2_{b_5} = \sigma^2_{b_6} = 0.5.$$

In the *heterogeneous-variance* setting, the random intercept has variance $\sigma^2_{b_0} = 2.0$, while the random slopes have variances $\sigma^2_{b_4} = 1.5$, $\sigma^2_{b_5} = 1.0$ and $\sigma^2_{b_6} = 0.5$, respectively. The random effects are generated independently in both settings, such that their covariance matrix is diagonal. The observation-level noise variance is fixed at $\sigma^2_\epsilon = 1$ throughout.

Within each random-effect variance setting, we generate three outcomes corresponding to different data-generating processes. The *mixed-effects* outcome combines the population-level fixed-effects function with participant-specific random effects and observation-level noise. The *fixed-effects-only* outcome contains the population-level fixed-effects function and observation-level noise but no participant-specific random effects, whereas the *random-effects-only* outcome contains the participant-specific random intercept and random slopes together with observation-level noise but no population-level fixed-effects signal. The latter two serve as limiting cases for evaluating MERL when either participant-specific random-effects variation or the population-level fixed-effects signal is absent. Consequently, the two variance settings affect the mixed-effects and random-effects-only outcomes, whereas the fixed-effects-only outcome contains no random effects in either setting. Full details of the feature-generating processes, biosignal construction, outcome equations and variance settings are provided in Appendix A.

3.3 Mixed-Effects Representation Learning

MERL builds on the alternating fixed- and random-effects estimation strategy of mixedML (Kilian et al., 2023), implementing the fixed-effects function f_θ with a multimodal CNN–Transformer.

At outer iteration k, the fixed-effects part of the model produces

$$\hat{y}_{it}^{(k,\text{fix})} = f_{\theta^{(k)}}(\boldsymbol{S}_{it}, \boldsymbol{c}_{it}), \tag{2}$$

where $\theta^{(k)}$ denotes its model parameters at iteration k. The corresponding training residuals are

$$r_{it}^{(k)} = y_{it} - \hat{y}_{it}^{(k,\text{fix})}. \tag{3}$$

A linear mixed-effects model is fitted to these residuals:

$$r_{it}^{(k)} = \boldsymbol{q}_{it}^\top \boldsymbol{b}_i^{(k)} + e_{it}^{(k)}, \tag{4}$$

where $\boldsymbol{q}_{it}$ specifies the random-effect components. In our experiments,

$$\boldsymbol{q}_{it} = [1, x_{it,4}, x_{it,5}, x_{it,6}]^\top, \tag{5}$$

these correspond to a random intercept and three random slopes. The random-effects covariance matrix is modeled as diagonal, residual errors are assumed independent and model parameters are estimated using restricted maximum likelihood (REML).

The estimated random-effects contribution is then removed from the original training target:

$$\tilde{y}_{it}^{(k)} = y_{it} - \boldsymbol{q}_{it}^\top \hat{\boldsymbol{b}}_i^{(k)}. \tag{6}$$

The neural network is subsequently trained on the adjusted target $\tilde{y}_{it}^{(k)}$. For prediction, the fixed- and random-effects components are combined:

$$\hat{y}_{it}^{(k)} = \hat{y}_{it}^{(k,\text{fix})} + \boldsymbol{q}_{it}^\top \hat{\boldsymbol{b}}_i^{(k)}. \tag{7}$$

MERL therefore alternates between learning the population-level fixed-effects function from the multimodal inputs and estimating systematic participant-specific deviations from its residuals.

The neural network is a multimodal CNN–Transformer that processes the five biosignal modalities together with five scalar covariates. Each biosignal modality is processed by a separate convolutional

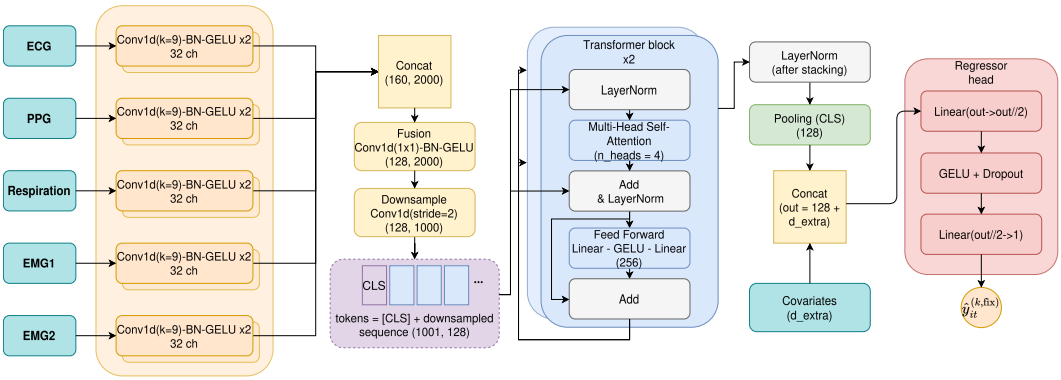


Figure 1: Architecture of the neural network used in MERL and the neural network baselines. Five biosignal modalities are independently encoded by convolutional subnetworks, whose representations are fused and processed by a Transformer encoder. The resulting representation is combined with five scalar covariates and passed through a regression head to produce the population-level prediction.

subnetwork to obtain a modality-specific representation. These representations are fused along the feature dimension and processed by a Transformer encoder to model temporal interactions across the fused multimodal sequence. The resulting representation is concatenated with the scalar covariates and passed through a regression head to produce the population-level fixed-effects prediction. An overview of the architecture is shown in Figure 1.

Participant identifiers are not provided to the neural network. Participant-specific deviations are instead represented explicitly through the random-effects component of MERL.

In each outer iteration, the neural network is reinitialized. MERL is run for at most 10 outer iterations and terminates when the absolute change in full training RMSE between successive iterations falls below 10^{-4}. Further architectural and optimization details are provided in Appendix B.

3.4 Experiments and Evaluation

We compare MERL with two neural network baselines using the same multimodal CNN–Transformer architecture. NN receives the biosignals and scalar covariates without participant identifiers, whereas $\mathrm{NN}_{\mathrm{OHE}}$ additionally receives a one-hot participant identifier concatenated before the regression head. We also consider global-mean, participant-mean and last-outcome baselines, which uses as prediction the training-set mean, participant-specific training mean and assessment-19 outcome, respectively.

All neural network models are trained by minimizing mean squared error. The training procedure, optimizer settings, architectural hyperparameters and MERL-specific optimization parameters are provided in Appendix B. To assess robustness to stochastic optimization and initialization, each experiment is repeated using five random seeds.

The primary task is future-assessment prediction for previously observed participants. Assessments 1–19 are used for training. At assessment 20, the 1,000 participant observations are divided equally into validation ($n = 500$) and test ($n = 500$) sets. Thus, all participants are observed during training and the evaluation concerns prediction of a future observation for a previously observed participant rather than cold-start generalization to unseen participants. Participant-specific random-effects used for assessment-20 prediction are estimated exclusively from assessments 1–19.

We evaluate all methods under the three outcomes and two random-effect variance settings described in Section 3.2.

Predictive performance is measured using root mean squared error (RMSE). For NN, $\mathrm{NN}_{\mathrm{OHE}}$ and MERL, we report mean test RMSE and its standard deviation across the five random seeds. To compare the three neural network based architecture, we additionally conduct pairwise paired t-tests on test RMSE across matched random seeds within each outcome and random-effect variance setting.

The three pairwise comparisons are NN versus $\mathrm{NN}_{\mathrm{OHE}}$, NN versus MERL and $\mathrm{NN}_{\mathrm{OHE}}$ versus MERL. We apply a Bonferroni correction across these three comparisons and use a significance level of $\alpha = 0.05$.

Because the simulation provides known ground-truth random-effects variance components, we additionally evaluate random-effects recovery by comparing the estimated variances of the participant-specific intercept and slopes with the true variance. The estimated residual variance is likewise compared with the true residual variance $\sigma_\epsilon^2 = 1$.

4 RESULTS

Table 1: Test RMSE across outcomes, prediction methods and random-effect variance settings. For NN, $\mathrm{NN}_{\mathrm{OHE}}$ and MERL, values are reported as mean (standard deviation) across five seeds. Best performance within each outcome and variance setting is shown in bold. Superscripts indicate significant pairwise differences based on paired t-tests with Bonferroni correction ($\alpha = 0.05$): ^{a} significantly different from NN; ^{b} significantly different from $\mathrm{NN}_{\mathrm{OHE}}$.

RMSE ↓	Global	Group	Last	NN	$\mathrm{NN}_{\mathrm{OHE}}$	MERL
Equal random-effect variances: $\sigma^2_{b_0} = \sigma^2_{b_4} = \sigma^2_{b_5} = \sigma^2_{b_6} = 0.5$						
Fixed only	9.502	5.389	3.785	**1.179 (0.038)**	1.299 (0.053)	1.230 (0.033)^b
Random only	1.699	1.590	2.294	1.705 (0.028)	**1.196 (0.020)**^{a}	1.268 (0.222)^a
Mixed	11.210	6.487	3.908	6.166 (0.101)	2.470 (0.158)^a	**1.525 (0.155)**^{a,b}
Heterogeneous random-effect variances: $\sigma^2_{b_0} = 2.0$, $\sigma^2_{b_4} = 1.5$, $\sigma^2_{b_5} = 1.0$, $\sigma^2_{b_6} = 0.5$						
Fixed only	9.502	5.389	3.785	**1.179 (0.038)**	1.299 (0.053)	1.230 (0.033)^b
Random only	6.631	3.825	2.155	6.809 (0.353)	1.687 (0.226)^a	**1.345 (0.111)**^{a}
Mixed	11.703	6.747	4.055	6.790 (0.289)	2.230 (0.090)^a	**1.469 (0.075)**^{a,b}

Table 2: True and estimated random-effect variances across outcomes and random-effect variance settings. MERL estimates are reported as mean (standard deviation) across five seeds. The true residual variance is 1.

	Fixed only		Random only		Mixed	
Random effect	True	Estimated	True	Estimated	True	Estimated
Equal random-effect variances: $\sigma^2_{b_0} = \sigma^2_{b_4} = \sigma^2_{b_5} = \sigma^2_{b_6} = 0.5$						
(Intercept)	0.000	0.028 (0.022)	0.500	0.444 (0.001)	0.500	0.521 (0.107)
x_4	0.000	0.000 (0.001)	0.500	0.434 (0.000)	0.500	0.317 (0.012)
x_5	0.000	0.001 (0.000)	0.500	0.402 (0.000)	0.500	0.435 (0.002)
x_6	0.000	0.000 (0.000)	0.500	0.416 (0.000)	0.500	0.427 (0.014)
Residual	1.000	0.938 (0.096)	1.000	1.002 (0.000)	1.000	0.792 (0.119)
Heterogeneous random-effect variances: $\sigma^2_{b_0} = 2.0$, $\sigma^2_{b_4} = 1.5$, $\sigma^2_{b_5} = 1.0$, $\sigma^2_{b_6} = 0.5$						
(Intercept)	0.000	0.028 (0.022)	2.000	1.751 (0.069)	2.000	1.914 (0.182)
x_4	0.000	0.000 (0.001)	1.500	1.173 (0.102)	1.500	1.153 (0.042)
x_5	0.000	0.001 (0.000)	1.000	0.911 (0.011)	1.000	0.906 (0.009)
x_6	0.000	0.000 (0.000)	0.500	0.415 (0.015)	0.500	0.419 (0.009)
Residual	1.000	0.938 (0.096)	1.000	0.784 (0.274)	1.000	0.817 (0.127)

4.1 FUTURE-ASSESSMENT PREDICTION

Table 1 reports future-assessment prediction performance across the three outcomes and two random-effect variance settings. MERL shows its clearest advantage in the mixed-effects outcome.

It achieves the lowest RMSE under both the equal- and heterogeneous-variance settings, outperforming both NN and $\mathrm{NN}_{\mathrm{OHE}}$. These differences are statistically significant under the Bonferroni-corrected paired t-tests in both variance settings. The outcome-history baselines also yield substantially higher RMSE than MERL in both mixed-effects settings.

The structural edge cases show a different pattern. In the fixed-effects-only outcome, all three neural network based approaches achieve similar predictive performance. NN has the lowest numerical RMSE, while MERL differs significantly from $\mathrm{NN}_{\mathrm{OHE}}$.

In the random-effects-only outcome, methods that incorporate participant-specific information outperform NN in both variance settings. Under equal random-effect variances, $\mathrm{NN}_{\mathrm{OHE}}$ has the lowest numerical RMSE, whereas MERL has the lowest numerical RMSE under heterogeneous random-effect variances. Both MERL and $\mathrm{NN}_{\mathrm{OHE}}$ differ significantly from NN in both settings, while the difference between MERL and $\mathrm{NN}_{\mathrm{OHE}}$ is not significant after Bonferroni correction.

Overall, these results show that MERL provides its clearest predictive advantage when population-level and participant-specific effects jointly contribute to the outcome. At the same time, its performance remains competitive in the fixed-effects-only and random-effects-only edge cases, suggesting robustness to differences in the underlying effect structure and to the magnitude and distribution of participant-specific variation.

4.2 Recovery of Random-Effects Structure

Table 2 compares the variance components estimated by MERL with their known simulation values. When random effects are absent, MERL estimates the random-slope variances at or near zero, while the estimated random-intercept variance is also small. Thus, the model does not introduce substantial participant-specific variance when the data-generating process contains no random effects.

In the random-effects-only outcome, the estimated variance components reflect the underlying random-effects structure in both settings. Under equal random-effect variances, the four estimated random-effect variances are close to the common true variance of 0.5 and the estimated residual variance closely matches its true value of 1. Under heterogeneous random-effect variances, MERL captures the ordering and differing magnitudes of the variance components, although the random-effect variances are generally underestimated.

A similar pattern is observed in the mixed-effects outcome. MERL reflects the heterogeneous variance structure, including the substantially larger random-intercept variance, but recovery is less exact for several random slopes and the residual variance. Under equal random-effect variances, recovery is likewise less exact than in the random-effects-only outcome, particularly for the random slope associated with x_4. Overall, the estimated variance components reflect the presence, absence and broad structure of participant-specific variation, while exact variance recovery becomes more difficult when population-level and participant-specific components must be estimated jointly.

5 Discussion

Longitudinal multimodal data combine two distinct modeling challenges: high-dimensional temporal structure within each assessment and participant-specific dependence across repeated assessments. MERL addresses these through separate components: a multimodal CNN–Transformer learns the population-level predictive relationship from raw biosignals and scalar covariates, while explicit random effects represent systematic participant-specific deviations. The results suggest that this separation is particularly useful when both sources of variation contribute to the outcome.

The mixed-effects outcome provides the clearest evidence for this behavior. As shown in Table 1, MERL achieves the lowest prediction error under both the equal and heterogeneous random-effect variance settings, significantly outperforming both NN and $\mathrm{NN}_{\mathrm{OHE}}$ in the Bonferroni-corrected paired comparisons. The comparison with $\mathrm{NN}_{\mathrm{OHE}}$ is particularly informative because providing participant identifiers is not equivalent to explicitly modeling participant-specific deviations. One-hot identifiers allow participant-specific information to enter the neural network prediction function directly, whereas MERL maintains an explicit decomposition between the shared population-level relationship and structured participant-specific variation.

The fixed-effects-only and random-effects-only outcomes serve as edge cases for evaluating the behavior of MERL when one component of the model is absent from the data-generating process. In the fixed-effects-only outcome, MERL remains close to the standalone neural network and estimates random-effect variances at or near zero (Table 2), indicating that the random-effects component does not introduce substantial participant-specific variation when none is present. In the random-effects-only outcome, MERL remains competitive with direct participant encoding despite the absence of a population-level signal. These edge cases are primarily intended to probe the behavior of the proposed decomposition rather than to represent typical longitudinal cohort data. In practice, repeated-measures data will generally contain both population-level relationships and participant-specific variation, making the mixed-effects outcome the more representative setting for the intended application of MERL.

The explicit separation of the population-level learner and participant-specific random effects also provides flexibility for prediction across seen and unseen participants. For a previously observed participant, the population-level neural network prediction can be combined with the participant's estimated random effects. For an unseen participant, the trained neural network can instead provide a population-level prediction without requiring participant identifiers or estimated random effects as inputs. If observations from that participant subsequently become available, participant-specific random effects could be estimated and incorporated into future predictions. Thus, the same decomposition naturally accommodates different levels of available participant-specific information, although the present experiments evaluate only previously observed participants.

The variance-recovery results provide further evidence for the value of the explicit decomposition. As shown in Table 2, MERL estimates negligible random-effect variances when random effects are absent and substantial variance when they are present. Importantly, under the heterogeneous-variance setting, the estimated components also preserve the ordering of the true variances in both the random-effects-only and mixed-effects outcomes. MERL therefore captures not only the presence of participant-specific variation but also differences in its magnitude. Recovery is not exact, however, with several random-effect and residual variances underestimated. This highlights the difficulty of allocating variation among a flexible neural network population-level model, random effects and residual noise and motivates considering predictive accuracy and variance recovery as complementary evaluation criteria.

MERL builds on the architecture-agnostic mixedML formulation by implementing its fixed-effects learner with a CNN–Transformer operating directly on multimodal physiological waveforms. This separates within-assessment temporal representation learning from between-assessment participant-specific variation, allowing each source of structure to be modeled explicitly rather than requiring a single neural network representation to account for both.

Several limitations remain. The current evidence is simulation-based, which enables evaluation against known population-level and participant-specific structure but does not establish performance on empirical health data. The experiments consider continuous outcomes, regularly observed assessments, diagonal random-effects covariance structures and independently generated biosignal modalities. Furthermore, although the model formulation supports population-level prediction for unseen participants, this setting is not evaluated here. Future work should evaluate MERL on empirical longitudinal multimodal datasets and investigate unseen-participant prediction, sequential estimation of random effects for newly observed participants, more complex covariance structures, irregular observations, non-Gaussian outcomes and cross-modal dependencies.

Overall, MERL shows its clearest predictive advantage when population-level and participant-specific effects coexist, corresponding to the primary setting considered in this work. In the limiting cases, MERL also behaves appropriately when either component is absent, remaining competitive in predictive performance and estimating negligible random-effect variance when participant-specific variation is absent. Across the two variance settings, its estimated random effects further reflect the presence and relative magnitude of participant-specific variation. These results support the explicit separation of neural network representation learning and statistical random effects as a promising approach for longitudinal multimodal prediction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods through Python code and to provide feedback on the research methodology and experimental design. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, translate content, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results.

Additionally, we used generative AI tools to suggest experimental parameters, identify relevant literature and edit parts of the manuscript to improve readability. We reviewed and verified all AI-assisted work, including testing AI-assisted code for correctness, independently evaluating methodological and experimental suggestions and reviewing all AI-assisted edits to the manuscript.

We take responsibility for the final content of this work, including text, claims, code and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code and data for reproducing the experiments in this work are available at `https://anonymous.4open.science/r/mixed_effects_representation_learning-A84C/`. The repository contains the implementation of MERL, including the mixed-effects estimation procedure and the multimodal CNN–Transformer. It also provides the simulated datasets used for the reported experiments and scripts for running the experimental pipeline presented in this paper. Additional details on the simulation procedure, model architecture, training configuration and experimental setup are provided in the appendix.

REFERENCES

Ayman Anwar, Yassin Khalifa, James L. Coyle, and Ervin Sejdic. Transformers in biosignal analysis: A review. *Information Fusion*, 114:102697, 2025. doi: 10.1016/j.inffus.2024.102697.

Lucía A. Carrasco-Ribelles, José Llanes-Jurado, Carlos Gallego-Moll, Margarita Cabrera-Bean, Mònica Monteagudo-Zaragoza, Concepción Violán, and Edurne Zabaleta-del Olmo. Prediction models using artificial intelligence and longitudinal data from electronic health records: A systematic methodological review. *Journal of the American Medical Informatics Association*, 30 (12):2072–2082, 2023. doi: 10.1093/jamia/ocad168.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2013. ISBN 9781134742707. doi: 10.4324/9780203771587.

Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U. Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, 161:1–13, 2018. doi: 10.1016/j.cmpb.2018.04.005.

Ahlem Hajjem, François Bellavance, and Denis Larocque. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328, 2014. doi: 10.1080/00949655.2012.741599.

Pascal Kilian, Sangbeak Ye, and Augustin Kelava. Mixed effects in machine learning — a flexible mixedML framework to add random effects to supervised machine learning regression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=MKZyHtmfwH`.

Adrienne Kline, Hanyin Wang, Yikuan Li, Stephen Dennis, Michelle Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5:171, 2022. doi: 10.1038/s41746-022-00712-8.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, 1982.

Mary J. Lindstrom and Douglas M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687, 1990.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2021. doi: 10.3758/s13428-020-01516-y.

F. Mandel, R. P. Ghosh, and I. Barnett. Neural networks for clustered and longitudinal data using mixed effects models. *Biometrics*, 79(2):711–721, 2023. doi: 10.1111/biom.13615.

B. Rim, N. J. Sung, S. Min, and M. Hong. Deep learning in physiological signal data: A survey. *Sensors*, 20(4):969, 2020. doi: 10.3390/s20040969.

Rebecca J. Sela and Jeffrey S. Simonoff. RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2):169–207, 2012. doi: 10.1007/s10994-011-5258-3.

Benjamin Shickel, Patrick J. Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2018. doi: 10.1109/JBHI.2017.2767063.

Giora Simchoni and Saharon Rosset. Using random effects to account for high-cardinality categorical features and repeated measures in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25111–25122. Curran Associates, Inc., 2021.

Giora Simchoni and Saharon Rosset. Integrating random effects in deep neural networks. *Journal of Machine Learning Research*, 24(156):1–57, 2023. URL `http://jmlr.org/papers/v24/22-0501.html`.

A. S. Tang, S. R. Woldemariam, S. Miramontes, et al. Harnessing EHR data for health research. *Nature Medicine*, 30:1847–1855, 2024. doi: 10.1038/s41591-024-03074-8.

M.-N. Tran, N. Nguyen, D. Nott, and R. Kohn. Bayesian deep net GLM and GLMM. *Journal of Computational and Graphical Statistics*, 29(1):97–113, 2020. doi: 10.1080/10618600.2019.1637747.

Bodo Winter. Linear models and linear mixed effects models in R with linguistic applications, 2013. URL `https://arxiv.org/abs/1308.5499`.

Yunyang Xiong, Hyunwoo J. Kim, and Vikas Singh. Mixed effects neural networks (MeNets) with applications to gaze estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7735–7744, 2019. doi: 10.1109/CVPR.2019.00793.

A Simulation Details

A.1 Longitudinal Covariates

We simulate seven time-varying features, $\boldsymbol{x}_{it} = [x_{it,0}, \ldots, x_{it,6}]^\top$ and three participant-level features, $\boldsymbol{x}_{ci} = [x_{ci,0}, x_{ci,1}, x_{ci,2}]^\top$. Table 3 summarizes their roles in biosignal generation, the random-effects design and the predictive models. A summary of the simulation parameters is provided in Table 4.

The time-varying features are generated using a partially persistent process,

$$\boldsymbol{z}_{it} = \boldsymbol{A}(\boldsymbol{T}\mathbf{z}_{i,t-1}) + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{z}_{it}^{\text{new}} + \boldsymbol{\epsilon}_{it}, \tag{8}$$

where $\boldsymbol{A}$ controls feature persistence, $\boldsymbol{T}$ is a diagonal transition matrix and

$$\boldsymbol{\epsilon}_{it} \sim \mathcal{N}(0, 0.2^2\boldsymbol{I}).$$

The transition coefficients in $\boldsymbol{T}$ are centered at one with standard deviation 0.05. The diagonal entries of $\boldsymbol{A}$ determine the temporal behavior of each feature: an entry of 1 yields persistent dynamics driven by the previous-assessment value, an entry of 0 yields a new draw at each assessment and

intermediate values combine the previous-assessment value with a new draw, producing partially persistent dynamics. Participant-level features are sampled once per participant and remain constant across all assessments.

The covariance matrices have unit diagonal entries, with off-diagonal correlations sampled from 0.1 or 0.3 with equal probability. These two covariance values were chosen deliberately to create realistic low-to-moderate dependence between generative variables (e.g., Cohen, 2013). The primary simulation uses random seed 42.

Consequently, the simulated features exhibit heterogeneous longitudinal dependence, ranging from persistent to independently redrawn dynamics, with partially persistent features between these extremes and additional noise at each assessment.

Table 3: Roles of the generative features in the simulation.

Feature	**Role**	**Use**
$x_{it,0}$	Biosignal generation	ECG heart rate
$x_{it,1}$	Biosignal generation	PPG heart rate
$x_{it,2}$	Biosignal generation	Respiratory rate
$x_{it,3}$	Biosignal generation	EMG_1 burst duration
$x_{ci,0}$	Biosignal generation	EMG_2 burst duration
$x_{it,4}$	Random-effects design + model input	Random slope; covariate
$x_{it,5}$	Random-effects design + model input	Random slope; covariate
$x_{it,6}$	Random-effects design + model input	Random slope; covariate
$x_{ci,1}$	Model input	Covariate
$x_{ci,2}$	Model input	Covariate

A.2 Biosignal Generation

Five biosignal channels are generated for every participant-assessment using NeuroKit2 (Makowski et al., 2021): ECG, PPG, respiration, EMG1 and EMG2. Example realizations of the five simulated biosignals are shown in Figure 2. Each signal contains 2,000 samples, corresponding to 20 seconds sampled at 100 Hz. The mapping from generative features to physiological simulation parameters is summarized in Table 5.

The features determine the physiological parameters of the signals. Specifically, $x_{it,0}$ controls ECG heart rate in $[30, 200]$ bpm, $x_{it,1}$ controls PPG heart rate in $[30, 160]$ bpm and $x_{it,2}$ controls respiratory rate in $[12, 30]$ breaths/min. The features $x_{it,3}$ and $x_{ci,0}$ control the EMG burst durations for EMG1 and EMG2, respectively, each in the range $[0.5, 3]$ s.

ECG and PPG signals use fixed heart-rate-variability parameters, while each EMG signal contains five simulated bursts. The signals are generated independently, with each modality determined by its own generative feature. Thus, the simulation does not explicitly impose cross-modal dependence between the biosignals.

A.3 Outcome Generation

The outcome intercept is $\beta_0 = 5$ and observation-level noise follows

$$\epsilon_{it} \sim \mathcal{N}(0, 1).$$

The population-level fixed-effects component is defined as

$$f(\boldsymbol{x}_{it}, \boldsymbol{x}_{ci}) = \sum_{j=0}^{6} x_{it,j} + \sum_{j=0}^{2} x_{ci,j}. \tag{9}$$

Participant-specific effects are generated as

$$g_i(\boldsymbol{x}_{it}) = b_{0i} + \sum_{j \in \{4,5,6\}} b_{ji} x_{it,j}, \tag{10}$$

Table 4: Summary of simulation parameters. Random-effect variances are shown for the equal-variance and heterogeneous-variance settings, respectively.

Parameter	Value
Participants (N)	1,000
Assessments (T)	20
Time-varying generative features (p)	7
Participant-level generative features (p_c)	3
Feature covariance values	0.1 or 0.3 (each w.p. 0.5)
Transition coefficient, $\boldsymbol{T}_j$	$\mathcal{N}(1, 0.05^2)$
Temporal behaviors	$1/3$ persistent, $1/3$ resampled, $1/3$ mixed
Feature transition noise std.	0.2
Biosignal duration / sampling rate	20 s / 100 Hz
Biosignal samples	2,000
Biosignal channels (F)	5
Outcome intercept, β_0	5
Observation-level noise std.	1
Equal random-effect variances	
Random intercept variance, $\sigma^2_{b_0}$	0.5
Random slope variance, $\sigma^2_{b_4}$	0.5
Random slope variance, $\sigma^2_{b_5}$	0.5
Random slope variance, $\sigma^2_{b_6}$	0.5
Heterogeneous random-effect variances	
Random intercept variance, $\sigma^2_{b_0}$	2.0
Random slope variance, $\sigma^2_{b_4}$	1.5
Random slope variance, $\sigma^2_{b_5}$	1.0
Random slope variance, $\sigma^2_{b_6}$	0.5
Training assessments	Assessments 1–19 ($n = 19{,}000$)
Validation assessments	Assessment 20 ($n = 500$)
Test assessments	Assessment 20 ($n = 500$)
Primary simulation seed	42

Table 5: Mapping from generative features to physiological simulation parameters.

Generative feature	Physiological parameter	Target range
$x_{it,0}$	ECG heart rate	$[30, 200]$ bpm
$x_{it,1}$	PPG heart rate	$[30, 160]$ bpm
$x_{it,2}$	Respiratory rate	$[12, 30]$ breaths/min
$x_{it,3}$	EMG_1 burst duration	$[0.5, 3]$ s
$x_{ci,0}$	EMG_2 burst duration	$[0.5, 3]$ s

where

$$\mathbf{b}_i = [b_{0i} \quad b_{4i} \quad b_{5i} \quad b_{6i}]^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{D}). \tag{11}$$

The random effects are generated independently, so $\mathbf{D}$ is diagonal.

We consider two random-effect variance settings. In the *equal-variance* setting,

$$\mathbf{D}_{\text{equal}} = \text{diag}(0.5,\ 0.5,\ 0.5,\ 0.5), \tag{12}$$

such that the random intercept and all three random slopes have marginal variance 0.5. In the *heterogeneous-variance* setting,

$$\mathbf{D}_{\text{heterogeneous}} = \text{diag}(2.0,\ 1.5,\ 1.0,\ 0.5), \tag{13}$$

corresponding to $\sigma^2_{b_0} = 2.0$, $\sigma^2_{b_4} = 1.5$, $\sigma^2_{b_5} = 1.0$ and $\sigma^2_{b_6} = 0.5$. The observation-level noise variance remains $\sigma^2_\epsilon = 1$ in both settings.

For each variance setting, three outcomes are generated under different data-generating processes:

$$y_{it}^{\text{fixed}} = \beta_0 + f(\boldsymbol{x}_{it}, \boldsymbol{x}_{ci}) + \epsilon_{it}, \tag{14}$$

$$y_{it}^{\text{random}} = \beta_0 + g_i(\boldsymbol{x}_{it}) + \epsilon_{it}, \tag{15}$$

$$y_{it}^{\text{mixed}} = \beta_0 + f(\boldsymbol{x}_{it}, \boldsymbol{x}_{ci}) + g_i(\boldsymbol{x}_{it}) + \epsilon_{it}. \tag{16}$$

The fixed-effects-only outcome contains no participant-specific random effects, the random-effects-only outcome contains no population-level fixed-effects signal and the mixed-effects outcome contains both components. The equal- and heterogeneous-variance settings therefore modify the participant-specific component in the random-effects-only and mixed-effects outcomes. The fixed-effects-only outcome does not contain random effects and is generated according to the same structural specification in both settings.

B Model and Training Specification

B.1 Neural Network Architecture

The neural network processes the five biosignal channels using independent two-layer one-dimensional convolutional encoders. Each encoder maps one input channel to 32 features using convolutional kernels of size 9, batch normalization and GELU activations.

The five modality representations are concatenated to obtain 160 channels and are then mapped to 128 channels using a 1×1 convolution. A stride-2 convolution reduces the temporal dimension from 2,000 to 1,000 samples.

A learned CLS token and learned positional embeddings are added before the Transformer encoder. The Transformer contains two layers, four attention heads, model dimension 128, feed-forward dimension 256, dropout 0.1 and GELU activation.

The final CLS representation is concatenated with the five scalar covariates

$$[x_{it,4}, x_{it,5}, x_{it,6}, x_{ci,1}, x_{ci,2}],$$

which are the model-input covariates listed in Table 3. This produces 133 input features for the regression head. The regression head contains a 66-unit hidden layer with GELU activation and dropout 0.1, followed by a scalar output.

The participant identifier is not provided to the FE. In the NN_{OHE} baseline, a one-hot encoding of the participant identifier is additionally concatenated with the five scalar covariates before the regression head.

B.2 Mixed-Effects Specification

The residual mixed-effects model uses a participant identifier as the grouping variable:

$$r_{it} = b_{0i} + b_{4i}x_{it,4} + b_{5i}x_{it,5} + b_{6i}x_{it,6} + e_{it}. \tag{17}$$

Equivalently, the random-effects design vector is

$$\boldsymbol{q}_{it} = [1, x_{it,4}, x_{it,5}, x_{it,6}]^\top.$$

The random-effects covariance matrix is specified as diagonal and the residual errors are specified as independent. The mixed-effects model is estimated using restricted maximum likelihood (REML).

B.3 MERL Optimization

The neural network is trained by minimizing mean squared error. The outer MERL procedure monitors the full-training RMSE and terminates when the absolute change in training RMSE between successive outer iterations falls below 10^{-4} or when 10 outer iterations have been completed.

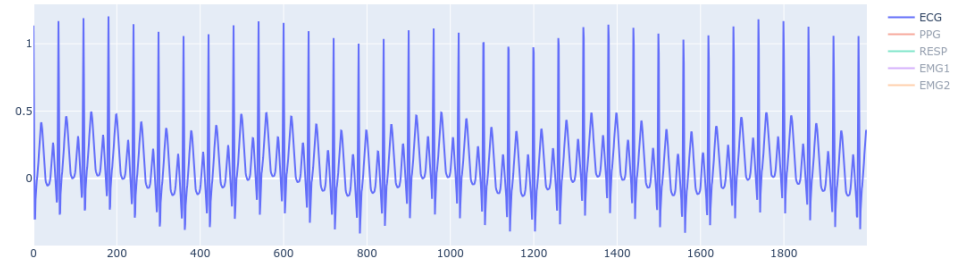
Table 6: Main network architecture and optimization settings.

Parameter	Value
CNN channels	32
CNN kernel size	9
Downsampling factor	2
Transformer dimension	128
Attention heads	4
Transformer layers	2
Feed-forward dimension	256
Dropout	0.1
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	10^{-4}
Batch size	20
Maximum epochs	500
Early-stopping patience	10
Maximum MERL iterations	10

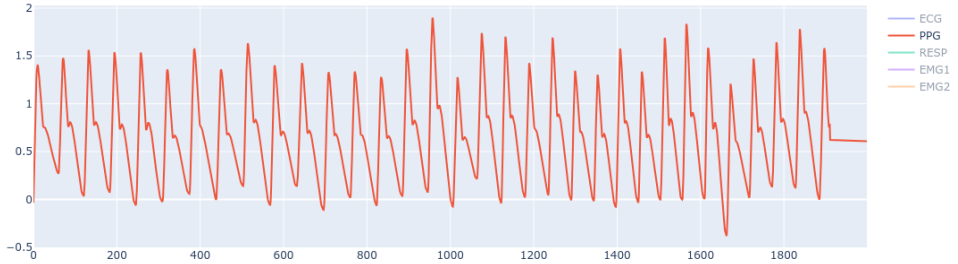
The neural network is reinitialized at each outer iteration rather than fine-tuned from the preceding iteration. Participant-specific random effects are estimated from training residuals only. Validation data are used for neural-network early stopping but are not used to estimate random effects.

The five experimental random seeds are

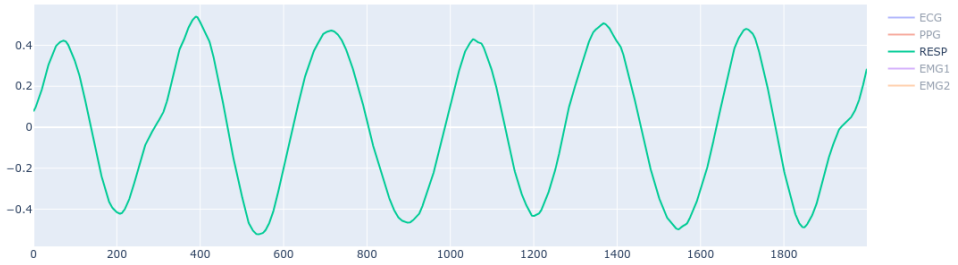
$$\{42, 123, 456, 789, 101112\}.$$



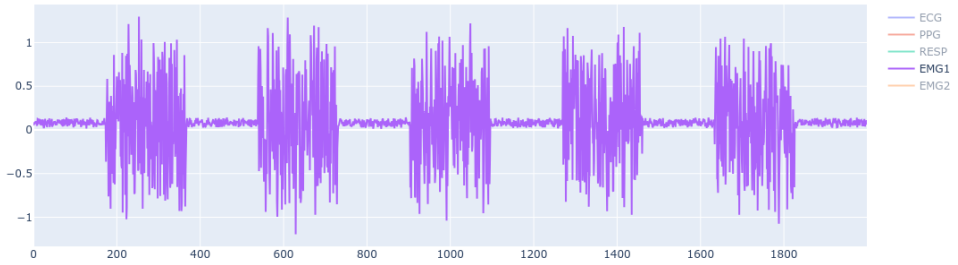
(a) Electrocardiogram (ECG)



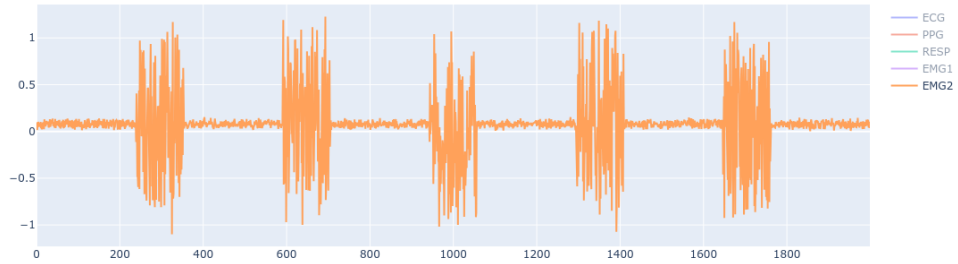
(b) Photoplethysmogram (PPG)



(c) Respiration



(d) Electromyogram 1 (EMG_1)



(e) Electromyogram 2 (EMG_2)

Figure 2: Example simulated biosignals generated using NeuroKit2. Each panel shows one of the five biosignal modalities used in the simulation.