

Received 5 July 2026, accepted 3 August 2026, date of publication 10 August 2026, date of current version 1 September 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3722150

## RESEARCH ARTICLE

# $P^2$ INNs: Physio-Psychiatry Informed Neural Networks for Depression Risk Prediction

NITESH BHAROT<sup>1,4</sup>, (Senior Member, IEEE), KARTIK SIROHI<sup>2</sup>,  
MUSLIM JAMEEL SYED<sup>3</sup>, CRISTINA HAVA MUNTEAN<sup>4</sup>, (Member, IEEE),  
AND PRIYANKA VERMA<sup>5</sup>, (Senior Member, IEEE)

<sup>1</sup>Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway, Galway, H91TK33 Ireland

<sup>2</sup>Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Punjab, Jalandhar 144011, India

<sup>3</sup>Enterprise and Technology Department, Atlantic Technological University, Galway, H91 T8NW Ireland

<sup>4</sup>National College of Ireland, Dublin, D01 K6W2 Ireland

<sup>5</sup>School of Computer Science and Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway, Galway, H91TK33 Ireland

Corresponding author: Nitesh Bharot (nitesh.bharot@universityofgalway.ie)

This work was supported in part by Taighde Éireann-Research Ireland under Grant 12/RC/2289\_P2 (Insight).

**ABSTRACT** Physics-Informed Neural Networks (PINNs) incorporate prior knowledge into learning by embedding differential equation constraints into the training objective. Motivated by the limited adoption of such approaches in computational psychiatry, we propose Physio-Psychiatry Informed Neural Networks ( $P^2$ INNs) for depression-related prediction tasks.  $P^2$ INNs integrate clinically motivated differential constraints derived from empirical physiological literature, capturing established relationships between sleep, physical activity, and affective risk. These constraints act as structured inductive biases during training, guiding the network toward physiologically plausible solutions while preserving flexibility to model inter-individual variability. We evaluate the proposed framework across two independent case studies using the StudentLife and Depresjon datasets, spanning regression and classification settings with heterogeneous affective targets. Experimental results demonstrate that  $P^2$ INNs consistently improve predictive performance in low-data regimes and exhibit increased robustness to hyperparameter variation compared to standard neural networks and conventional regularization techniques. We further compare against monotonic neural networks and sign-only gradient penalties, demonstrating that the full ODE-form constraint provides measurable advantages beyond enforcing monotonicity alone. While classical machine learning baselines remain competitive on this low-dimensional feature set,  $P^2$ INNs offer a principled mechanism for embedding physiological priors into neural architectures, with potential to scale to richer and temporally structured inputs.

**INDEX TERMS** Physics-informed neural networks, depression risk prediction, computational psychiatry, physiological modeling, monotonic networks, calibration.

## I. INTRODUCTION

Depression is increasingly understood as a disorder involving complex interactions between psychological processes and physiological regulation [1]. A growing body of evidence links depression risk to objectively measurable physiological signals, including sleep patterns, physical activity, and autonomic nervous system dynamics [2]. Leveraging such signals has enabled machine learning models to achieve

strong predictive performance across a range of depression-related tasks [3].

Despite these successes, neural networks are known to exploit spurious correlations and shortcut features when trained purely in a data-driven manner, particularly in low-data medical settings [4]. This motivates incorporating domain knowledge to bias learning toward clinically meaningful relationships.

PINNs provide one such mechanism by embedding known differential constraints into the training objective [5]. Originally developed for physical systems governed by

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

partial differential equations, this paradigm has since been extended to other scientific domains, including biology, where differential constraints serve as soft inductive biases rather than exact mechanistic laws [6].

In this work, we adapt this philosophy to computational psychiatry by introducing  $P^2$ INNs. Rather than treating psychiatric phenomena as deterministic dynamical systems, we encode empirically observed population-level physiological relationships as differentiable constraints that regularize learning. These constraints are formulated using ordinary differential equations, applied not as temporal dynamical models but as structural regularizers on aggregated physiological features. Through two independent case studies, we demonstrate that incorporating these constraints improves learning stability and predictive performance in low-data regimes without restricting model expressiveness.

We use the term *Physio-Psychiatry* to emphasize the nature of the inductive bias introduced in this work. The proposed framework incorporates well-established population-level relationships between physiological processes and affective risk, such as monotonic or asymmetric associations between sleep, physical activity, and depression, directly into the training objective as soft differential constraints. The term *physio-psychiatry informed* is therefore intended to denote the integration of physiological knowledge into psychiatric prediction models at the level of functional behavior.

A critical distinction from prior shape-constrained methods is that  $P^2$ INNs encode not only the *direction* of the relationship but also its *curvature and magnitude*, as implied by the governing differential equation. To make this contribution precise, we include ablation comparisons against sign-only gradient penalties and monotonic neural networks, separating the benefits of ODE-form constraints from simpler forms of monotonic inductive bias.

## A. KEY CONTRIBUTIONS

- We propose  $P^2$ INNs, which integrate ODE-form differential constraints linking sleep and physical activity directly into neural network training as structured inductive biases.
- We provide a principled justification for the hybrid strategy of fixing sleep parameters from epidemiological meta-analyses while learning activity parameters from data, grounded in the availability and granularity of prior quantitative evidence.
- We distinguish  $P^2$ INNs from simpler constraint strategies by ablating against sign-only gradient penalties and monotonic neural networks, showing that the ODE residual formulation provides additional benefit beyond direction enforcement alone.
- We evaluate the framework on the StudentLife and Depresjon datasets, demonstrating superior generalization in data-scarce regimes compared to baseline and simply-regularized neural networks, with consistent calibration improvements on the classification task;

we additionally benchmark against classical machine learning baselines to contextualize these gains.

## B. PAPER ORGANIZATION

We begin by formulating the problem and deriving the specific differential constraints linking physiological markers to depression risk. With this foundation in place, we detail the experimental design and present empirical results. The analysis focuses on predictive performance, calibration, and robustness to regularization pressure across varying data regimes. We conclude by discussing implications for the use of inductive biases in psychiatric computing.

## II. RELATED WORK

Prior work in computational psychiatry has extensively explored the use of machine learning for depression prediction using behavioral and physiological signals, including sleep, physical activity, and heart rate variability [2], [3]. These approaches are motivated by well-established links between depression and physiological dysregulation [1], [7], with empirical evidence demonstrating associations between depression risk and sleep duration [8], [9], and physical activity [10], [11], [12], [13]. However, the majority of existing models remain purely data-driven, treating physiological signals as input features rather than constraining model behavior in accordance with known physiological relationships.

In parallel, physics-informed and biologically informed neural networks have demonstrated that embedding differential constraints can substantially improve learning from sparse data by encoding domain knowledge directly into the training objective [5], [6]. While these frameworks have been successfully applied to physical and biological systems with explicit governing equations, they have not been systematically extended to psychiatric modeling, where mechanistic laws are unavailable and datasets are typically small.

A related family of approaches enforces shape constraints through architecture rather than loss. Monotone neural networks [14] guarantee monotone input-output mappings by constraining weight signs. These methods enforce the *direction* of a relationship but do not capture curvature or magnitude information encoded by a governing differential equation.  $P^2$ INNs complement this family by operating on the loss rather than the architecture, preserving expressive capacity while encoding richer quantitative structure.

Calibration in medical machine learning has received growing attention [15], with evidence that poorly calibrated probabilities can mislead clinical decision-making even when discrimination is adequate. The present work includes calibration analysis as a first-class evaluation criterion, which is particularly relevant for psychiatric risk assessment.

Despite the potential of wearable-based prediction, practical translation remains impeded by significant technical and methodological hurdles. Wearable data are often collected in uncontrolled daily environments where sensor performance is

compromised by motion artifacts, inconsistent skin contact, and environmental variability, leading to signal dropouts and unreliable measurements that laboratory-based validations often fail to capture [16]. Furthermore, integrating these signals into psychiatric models is prone to the ecological fallacy, where population-level epidemiological associations, such as the link between sleep duration and depression risk, are incorrectly assumed to reflect individual-level mechanisms [17]. Such aggregate priors may bias models toward population means, potentially obscuring heterogeneous or unique disease presentations in minority subgroups [18]. Moreover, while neural architectures have shown high discriminative power, such as high AUC, they are frequently plagued by poor probabilistic calibration, which acts as a critical failure point in clinical decision-support systems [19]. A model that accurately ranks patients but provides uncalibrated risk probabilities can be dangerous in triage settings. This lack of reliability, combined with a general failure to adhere to clinical prediction-model reporting standards, such as the TRIPOD+AI statement [20] which mandates the explicit definition of intended use, target populations, and prediction timing, hampers the transition of these models from research artifacts to clinically validated tools.

This work bridges the gap between physiological modeling and psychiatric prediction by introducing ODE-form differential constraints as population-level inductive biases across two real-world mental health datasets [21], [22], mitigating shortcut learning [4] and improving robustness under data scarcity.

### III. $P^2$ INNs

#### A. PROBLEM FORMULATION

Depression risk is influenced by multiple physiological subsystems that can be objectively measured using consumer-grade wearable sensors. In this work, we focus on two physiological variables: sleep duration and physical activity, each of which has been independently associated with depression in prior clinical and behavioral studies [7]. Sleep reflects circadian regulation, while physical activity represents behavioral activation and energy expenditure.

Although these factors jointly contribute to depression risk, no publicly available dataset provides reliable measurements of both modalities alongside depression-related labels within a single cohort. Rather than constructing a unified multimodal model with incomplete inputs, we adopt a modular experimental design in which each physiological factor is studied independently using a dedicated dataset. Specifically, we use the StudentLife dataset [21] to model sleep-related dynamics and the Depresjon dataset [22] to study physical activity.

Due to differences in dataset design and annotation protocols, the prediction targets differ across datasets. Our objective is not to directly compare absolute predictive performance across datasets, but to assess whether incorporating physiologically informed differential equation constraints consistently improves learning within each task.

**TABLE 1. Relationship between sleep duration and depression risk for the overall population as reported by [9].**

| Sleep Duration | Risk Ratio (RR) | 95% CI    |
|----------------|-----------------|-----------|
| 5 h            | 1.09            | 1.07–1.12 |
| 6 h            | 1.03            | 1.02–1.04 |
| 7 h            | Reference       | –         |
| 8 h            | 1.10            | 1.05–1.15 |
| 9 h            | 1.31            | 1.17–1.47 |
| 10 h           | 1.59            | 1.31–1.92 |

#### 1) STUDENTLIFE DATASET

In the StudentLife dataset [21], depression risk is modeled as a continuous variable using the Patient Health Questionnaire (PHQ-9) score, a widely used self-report instrument for assessing depressive symptom severity. We formulate this task as a regression problem, with the network predicting PHQ-9 scores from aggregated sleep statistics.

#### 2) DEPRESJON DATASET

The Depresjon dataset provides binary group labels distinguishing clinically depressed individuals from healthy controls. Accordingly, this task is formulated as a binary classification problem, where the model predicts depression status based on aggregated physical activity features.

We intentionally treat each dataset as an independent case study. The goal is to evaluate whether physiologically informed differential constraints consistently improve learning across heterogeneous affective endpoints. More information about the datasets can be found in the Appendix.

### B. $P^2$ INN FRAMEWORK

The proposed  $P^2$ INN framework combines neural network function approximation with physiologically motivated differential equation constraints. Given aggregated physiological statistics, the network learns to predict depression risk while being guided by population-level differential relationships derived from established clinical and physiological literature.

#### 1) SLEEP AND DEPRESSION RISK

Sleep duration and depression are strongly correlated, with sleep acting as both a risk factor and a symptom of depressive disorders [8]. Large-scale epidemiological evidence suggests a non-linear, J-shaped relationship between nightly sleep duration and depression risk in the general population [9].

Table 1 summarizes the reported risk ratios (RR) for depression across different sleep durations, using 7 hours as the reference category. Both short sleep and prolonged sleep are associated with elevated depression risk, motivating a quadratic deviation-based formulation.

We model the population-level relationship between sleep duration and depression risk using a quadratic deviation from an optimal sleep duration of 7 hours. To account for unobserved individual-specific baseline effects, we introduce an additive bias term  $\epsilon$  that shifts overall risk without altering

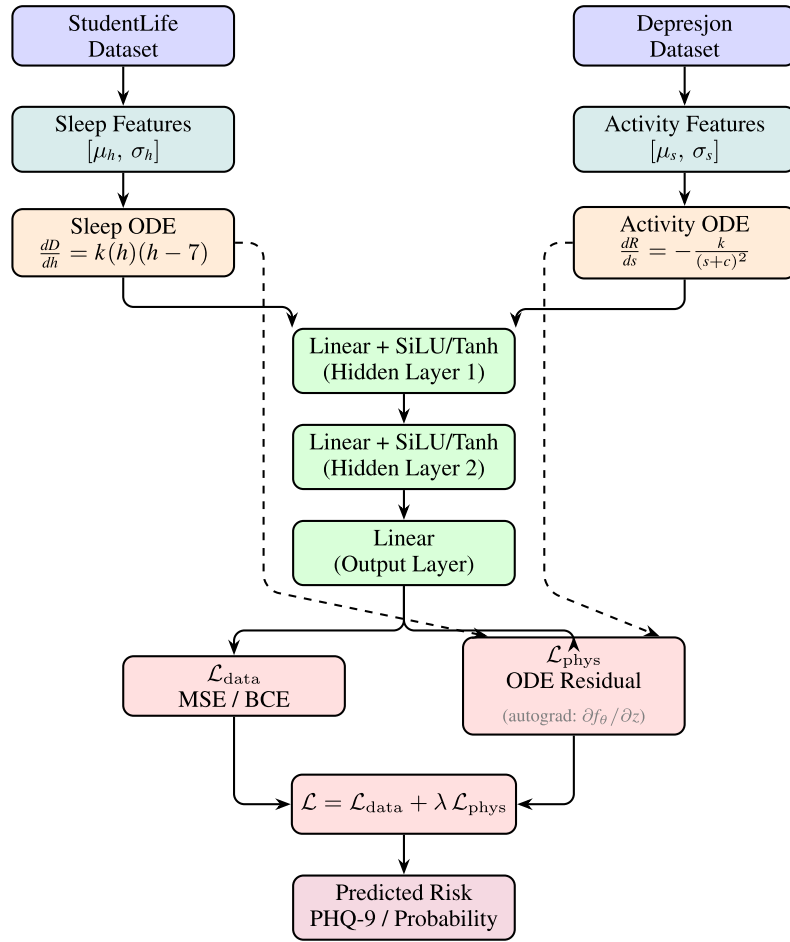


FIGURE 1. Overview of the P<sup>2</sup>INN methodology pipeline.

the governing dynamic:

$$D(h) = \begin{cases} \alpha(h-7)^2 + \epsilon, & h < 7 \\ \epsilon, & h = 7 \\ \beta(h-7)^2 + \epsilon, & h > 7 \end{cases} \quad (1)$$

where  $h$  is the number of hours of sleep. Differentiating with respect to sleep duration yields:

$$\frac{dD}{dh} = k(h)(h-7), \quad (2)$$

where the sensitivity function  $k(h)$  is:

$$k(h) = \begin{cases} 2\alpha, & h < 7 \\ 2\beta, & h > 7 \end{cases} \quad (3)$$

The additive term  $\epsilon$  does not appear in the differential equation and therefore does not influence the physiologically informed constraint, allowing individual baseline risk to be implicitly learned by the neural network.

*Justification for Fixed Parameters:*

The coefficients  $\alpha$  and  $\beta$  are treated as fixed population-level constants. This choice is justified by the availability of quantitative risk ratio data across discrete sleep durations from large-scale meta-analytic studies [9]. Specifically,

we fitted a piecewise quadratic function to the reported risk ratios, obtaining  $\alpha = 0.023$  for the sleep deprivation regime ( $h < 7$ ) and  $\beta = 0.066$  for the oversleeping regime ( $h > 7$ ). These values encode the asymmetric curvature directly supported by the epidemiological data. Fixing these parameters stabilizes training and imposes clinically grounded structure without requiring additional validation data. We emphasize that these fixed values do not imply a mechanistic law at the individual level; they anchor the inductive bias to a well-characterized population trend. A comparison between fully fixed and fully learnable parameterizations is an important direction for future ablation studies.

## 2) PHYSICAL ACTIVITY AND DEPRESSION RISK

Prior literature establishes a robust association between physical activity and depression, with higher levels of activity generally corresponding to reduced depression risk [10]. Meta-analytic evidence suggests an inverse, curvilinear relationship between activity and depression [11], while large-scale observational studies further indicate diminishing marginal benefits beyond a moderate activity threshold [13]. For instance, increasing daily steps from 2,000 to 10,000

**TABLE 2.** Summary of parameter treatment across experiments, with justification.

| Modality | Parameters                            | Treatment | Justification                                        |
|----------|---------------------------------------|-----------|------------------------------------------------------|
| Sleep    | $\alpha = 0.023$ ,<br>$\beta = 0.066$ | Fixed     | Quantitative RR estimates from meta-analysis [9]     |
| Activity | $k, c$                                | Learnable | Qualitative trend only; scale varies across datasets |

yields substantially larger mental health benefits than increasing from 10,000 to 50,000 [12].

Motivated by this threshold and saturation behavior, we model the relationship between aggregated daily activity  $s$  and depression risk  $R$  using an inverse rational formulation:

$$R(s) = \frac{k}{s + c} + \epsilon, \quad (4)$$

where  $k$  is a proportionality constant,  $c$  is a stabilizing constant that prevents singularity at zero activity, and  $\epsilon$  captures unobserved individual-specific baseline risk.

We adopt an inverse rational form due to its ability to capture diminishing marginal returns while maintaining smooth, bounded gradients. Alternative formulations such as logarithmic and exponential decay are plausible; however, we found the rational form to be numerically stable and well-suited for derivative-based regularization. We treat this choice as a modeling assumption rather than a physiological law.

Differentiating with respect to activity yields:

$$\frac{dR}{ds} = -\frac{k}{(s + c)^2}. \quad (5)$$

*Justification for Learnable Parameters:*

In contrast to sleep, existing literature on physical activity consistently establishes only the qualitative nature of the relationship, namely monotonic decrease and diminishing marginal effects, without providing stable numerical estimates of curvature or scale across populations [11], [13]. Furthermore, measurement units and aggregation procedures vary substantially across datasets. Enforcing fixed numerical parameters would risk imposing dataset-incompatible constraints. Accordingly, we treat the sensitivity parameters  $k$  and  $c$  in this formulation as learnable, allowing the network to adapt the strength of the physiological bias to dataset-specific scales while preserving the qualitative shape of the constraint.

This hybrid strategy reflects a deliberate modeling trade-off: parameters supported by strong quantitative evidence are fixed to stabilize learning, while parameters corresponding to weaker or scale-dependent prior knowledge are learned from data.

### 3) FIXED VERSUS LEARNABLE PHYSIOLOGICAL PARAMETERS: SUMMARY

Table 2 summarizes the parameter treatment across the two models.

In all cases, the imposed constraints regulate only the direction and curvature of the learned mapping, not its

absolute magnitude, ensuring that the neural network retains sufficient flexibility to capture inter-individual variability.

*Initialization, Positivity, and Numerical Stability:*

The parameters  $k$  and  $c$  are initialized to  $k_0 = 1.0$  and  $c_0 = 0.5$  for every training run. Positivity is not enforced via reparameterization, instead, the physiological constraint loss is computed using  $|k|$  and  $|c|$  throughout (Eq. 5), so the sign of the underlying learnable parameter does not affect the constraint's direction or magnitude. Singularities are avoided because model inputs are scaled to the range  $[1, 5]$  prior to training, ensuring  $s + |c| > 0$  for all observed activity values. Across 50 random seeds, the learned values were  $|k| = 0.457 \pm 0.279$  and  $|c| = 0.856 \pm 0.143$ , with both parameters remaining strictly positive in every run despite the absence of an explicit positivity constraint. We note that  $k$  exhibits considerably larger relative variance across seeds than  $c$ , suggesting the activity-decay rate is less tightly identified by the data than the stabilizing offset; this is consistent with  $k$  scaling the overall constraint strength while  $c$  primarily governs curvature near small activity values.

### C. PHYSIOLOGICALLY INFORMED TRAINING OBJECTIVE

Let  $\mathbf{x}_i \in \mathbb{R}^d$  denote the aggregated physiological features for subject  $i$ , and let  $y_i$  denote the corresponding target variable, which may represent a continuous depression score or a binary depression label depending on the dataset. We approximate the depression risk function using a neural network  $f_\theta(\mathbf{x})$  parameterized by weights  $\theta$ .

#### 1) DATA FIDELITY LOSS

For regression (StudentLife), we use the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{data}} = \frac{1}{N} \sum_{i=1}^N (f_\theta(\mathbf{x}_i) - y_i)^2. \quad (6)$$

For classification (Depresjon), we use the binary cross-entropy loss:

$$\mathcal{L}_{\text{data}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log f_\theta(\mathbf{x}_i) + (1 - y_i) \log(1 - f_\theta(\mathbf{x}_i))]. \quad (7)$$

#### 2) PHYSIOLOGICAL DIFFERENTIAL CONSTRAINT AND AUTOGRAD IMPLEMENTATION

Let  $z$  denote the primary physiological variable of interest (sleep duration  $h$  or activity  $s$ ). The constraint is applied by differentiating the network output with respect to  $z$  using automatic differentiation. In practice, all models operate on the two-dimensional input  $\mathbf{x} = [\mu_z, \sigma_z]$ , where  $\mu_z$  is the mean and  $\sigma_z$  is the standard deviation of the physiological variable. The constraint is applied exclusively to the partial derivative with respect to  $\mu_z$ , which approximates the expected slope of the risk function with respect to the latent variable (see Section III-C4 for a formal treatment).

The following pseudocode illustrates the autograd computation for a single batch:

---

**Algorithm 1** Physiological Constraint Loss Computation
 

---

**Require:** batch inputs  $\mathbf{X} \in \mathbb{R}^{N \times 2}$  (columns:  $\mu_z, \sigma_z$ )

- 1:  $\mathbf{X}.\text{requires\_grad} \leftarrow \text{True}$
  - 2:  $\hat{\mathbf{y}} \leftarrow f_\theta(\mathbf{X})$  ▷ forward pass
  - 3:  $\mathbf{G} \leftarrow \nabla_{\mathbf{X}} \hat{\mathbf{y}}$  ▷ autograd, shape  $N \times 2$
  - 4:  $\partial f / \partial \mu_z \leftarrow \mathbf{G}[:, 0]$  ▷ gradient w.r.t. mean only
  - 5:  $g_i \leftarrow g(\mu_{z,i}, \sigma_{z,i})$  ▷ target from ODE (Eqs. 2, 5, 23)
  - 6:  $\mathcal{L}_{\text{phys}} \leftarrow \frac{1}{N} \sum_i (\partial f / \partial \mu_{z,i} - g_i)^2$
  - 7: **return**  $\mathcal{L}_{\text{phys}}$
- 

For classification, the constraint is applied to the pre-sigmoid logit to preserve gradient interpretability and avoid saturation effects. Metrics (AUC, Brier score, ECE) are computed on the sigmoid probability output.

The physiological constraint loss is thus:

$$\mathcal{L}_{\text{phys}} = \frac{1}{N} \sum_{i=1}^N \left( \frac{\partial f_\theta(\mathbf{x}_i)}{\partial \mu_{z,i}} - g(\mu_{z,i}, \sigma_{z,i}) \right)^2, \quad (8)$$

where  $g(\mu_z, \sigma_z)$  denotes the analytically specified target derivative.

### 3) TOTAL TRAINING OBJECTIVE

The final training objective combines empirical data fitting with physiological consistency:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda \mathcal{L}_{\text{phys}}, \quad (9)$$

where  $\lambda > 0$  controls the strength of the physiological constraint. All model parameters are optimized jointly using gradient-based methods.

### 4) ON AGGREGATED-FEATURE GRADIENTS AND EXPECTED PHYSIOLOGICAL SLOPES

The physiological constraints are defined on latent raw variables  $z$ , whereas the predictive models operate on aggregated summary statistics  $(\mu_z, \sigma_z)$ . The imposed constraint takes the form:

$$\frac{\partial f_\theta(\mu_z, \sigma_z)}{\partial \mu_z} \approx \mathbb{E} \left[ \frac{dR}{dz} \right],$$

where the expectation is taken with respect to the latent distribution  $z \sim \mathcal{D}(\mu_z, \sigma_z^2)$ .

Under standard regularity conditions (differentiability of  $R$ , finite variance, and dominated convergence), differentiation and expectation commute:

$$\frac{d}{d\mu_z} \mathbb{E}[R(z)] = \mathbb{E} \left[ \frac{dR}{dz} \right].$$

In the present setting,  $f_\theta$  is a nonlinear neural approximation trained on aggregated features. The equality above should therefore be interpreted as an approximate moment-level inductive bias rather than a mechanistic identity. The Taylor-based expectations derived below

provide principled targets for the average slope of  $f_\theta$  with respect to  $\mu_z$ , encouraging consistency with known population-level physiological trends. This approximation is expected to hold most strongly when the latent distribution is unimodal with moderate variance.

### D. EXTENSION TO AGGREGATED STATISTICS

#### 1) SLEEP (STUDENTLIFE)

The StudentLife dataset does not provide a single deterministic sleep duration per subject but rather a sequence of nightly sleep observations. We assume nightly sleep duration for a given subject follows:

$$h \sim \mathcal{N}(\mu_h, \sigma_h^2), \quad (10)$$

where  $\mu_h$  denotes mean sleep duration and  $\sigma_h$  denotes sleep variability. This normality assumption is a reasonable approximation for typical sleep ranges (4–10 hours) observed in the StudentLife cohort.

Under this assumption, the physiological constraint is applied to the expected derivative:

$$\mathbb{E} \left[ \frac{dD}{dh} \right] = \int_{-\infty}^7 2\alpha(h-7)f(h)dh + \int_7^{\infty} 2\beta(h-7)f(h)dh, \quad (11)$$

where  $f(h)$  is the density of  $\mathcal{N}(\mu_h, \sigma_h^2)$ . Standardizing with  $Z = (h - \mu_h)/\sigma_h$  yields the closed-form expression:

$$g(\mu_h, \sigma_h) = \sigma_h \phi(Z)(2\beta - 2\alpha) + (\mu_h - 7)[2\alpha\Phi(Z) + 2\beta(1 - \Phi(Z))], \quad (12)$$

where  $\phi(\cdot)$  and  $\Phi(\cdot)$  denote the standard normal PDF and CDF, respectively. This formulation enables the physiological constraint to act directly on mean and variability descriptors of sleep without requiring time-series supervision.

#### 2) PHYSICAL ACTIVITY (DEPRESJON)

For the Depresjon dataset, the non-linear governing equation requires accounting for variability. We approximate the expected derivative using a second-order Taylor expansion around  $\mu_s$ :

$$g(\mu_s, \sigma_s) \approx -\frac{k}{(\mu_s + c)^2} - \frac{3k\sigma_s^2}{(\mu_s + c)^4}. \quad (13)$$

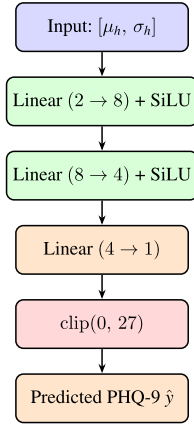
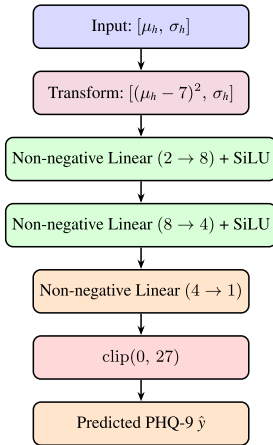
This approximation is used solely to define a differentiable regularization term and holds when  $\sigma_s$  is small relative to  $\mu_s + c$ .

## IV. MODEL ARCHITECTURES AND EXPERIMENTAL DESIGN

### A. EXPERIMENT 1: STUDENTLIFE (SLEEP REGRESSION)

#### 1) DATA PREPROCESSING

Sleep data are extracted from individual StudentLife JSON files, each corresponding to a single participant. Each participant's sleep history is summarized using two scalar features: mean sleep duration  $\mu_h$  and standard deviation  $\sigma_h$ . Two participants exhibited zero sleep duration variance

**FIGURE 2.** Architecture for Experiment 1 (StudentLife).**FIGURE 3.** Architecture of the Monotonic NN baseline for sleep (Experiment 1).

(likely due to insufficient measurements) and were excluded to avoid numerical instability, reducing the sample from  $N = 49$  to  $N = 46$ .

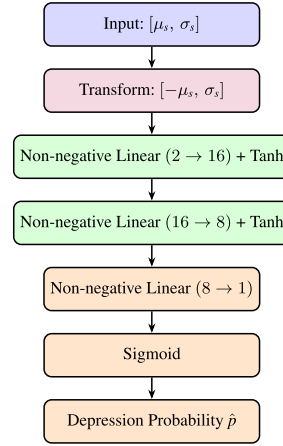
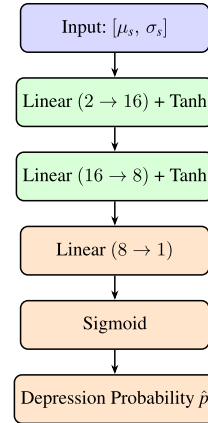
The final dataset comprises  $N = 46$  participants, each represented by a feature vector  $[\mu_h, \sigma_h]$  paired with a continuous PHQ-9 score. All train-test splits were performed at the subject level.

## 2) MODEL ARCHITECTURE

Both the baseline neural network and the  $P^2$ INN employ an identical feedforward architecture (Figure 2): two hidden layers with SiLU (Sigmoid Linear Unit, defined as  $f(x) = x \cdot \sigma(x)$  where  $\sigma$  is the logistic sigmoid function) activations and a single linear output layer. The network output is clamped to the valid PHQ-9 range:

$$\hat{y} = \text{clip}(f_{\theta}(\mathbf{x}), 0, 27). \quad (14)$$

Unlike sigmoid-based output constraints, hard clamping preserves linear gradients within the admissible region and avoids gradient saturation that could interfere with the differential loss. The range constraint is applied identically to both models.

**FIGURE 4.** Architecture of the Monotonic NN baseline for activity (Experiment 2).**FIGURE 5.** Architecture for Experiment 2 (Depresjon).

## B. EXPERIMENT 2: DEPESJON (ACTIVITY CLASSIFICATION)

### 1) DATA PREPROCESSING

For each subject, the full motor activity recording is loaded without artifact filtering. Ground-truth labels are derived directly from subject identifiers: subjects in the `condition` cohort are labeled as depressed and those in the `control` cohort as healthy. The final dataset comprises  $N = 55$  subjects (33 control, 22 depressed), each represented by  $[\mu_s, \sigma_s]$  paired with a binary label. All splits were performed at the subject level.

### 2) MODEL ARCHITECTURE

Both the baseline and  $P^2$ INN use the feedforward architecture shown in Figure 5. The physiological inductive bias is introduced exclusively during training by augmenting the binary cross-entropy loss with a differential constraint applied to the gradient of the pre-sigmoid logit with respect to  $\mu_s$ .

## C. SHAPE-CONSTRAINED BASELINES

To separate the benefits of ODE-form constraints from simpler monotonic inductive bias, we include two additional baselines in each experiment.

**TABLE 3. Lambda selection for  $P^2$ INN and shape-constrained baselines.**

| Dataset     | Lambda grid              | Chosen $\lambda$ | Criterion                                |
|-------------|--------------------------|------------------|------------------------------------------|
| StudentLife | {0.1, 1, 5, 12, 50, 100} | 12.0             | Min. median RMSE on held-out val. splits |
| Depresjon   | {0.1, 0.5, 1, 2, 5, 10}  | 2.0              | Min. Brier score on held-out val. splits |

### 1) SIGN-ONLY CONSTRAINT

A gradient penalty that penalizes only violations of the expected sign, without enforcing magnitude:

$$\mathcal{L}_{\text{sign}} = \frac{1}{N} \sum_{i=1}^N \left[ \max \left( 0, -s_i \cdot \frac{\partial f_{\theta}}{\partial \mu_{z,i}} \right) \right]^2, \quad (15)$$

where  $s_i$  is the expected sign of the gradient at  $\mu_{z,i}$  (positive for sleep when  $\mu_h > 7$ , negative for activity). This variant enforces monotonic inductive bias without the curvature information encoded in the ODE residual.

### 2) MONOTONIC NEURAL NETWORK

An architecturally constrained variant that enforces the target shape through non-negative weights (i.e., linear layers where the weight matrices are constrained to be strictly non-negative, typically via a softplus reparameterization, to guarantee monotonic input-output mappings) [14]. For sleep, the input is transformed to  $(h - 7)^2$  and  $\sigma_h$  before passing through a non-negative-weight network, architecturally encoding the U-shaped relationship. For activity, the activity input sign is flipped (to  $-\mu_s$ ) before passing through a non-negative-weight network, encoding the decreasing relationship. Final outputs use the same activation functions as the standard architecture. This variant enforces the shape constraint entirely through architecture, with no modification to the loss function.

These two baselines, together with standard L1 and L2 regularization, allow us to isolate the specific contribution of the ODE-form constraint.

### D. HYPERPARAMETER AND LAMBDA SELECTION

Table 3 reports the lambda grid, selected values, and selection criterion for each experiment.

The chosen  $\lambda$  was fixed across all 50 random seeds and all training fractions to avoid post-hoc tuning. For the robustness analysis,  $\lambda$  was then varied logarithmically from 0.1 to 100 to assess stability under increasing constraint pressure.

## V. RESULTS

To rigorously evaluate the efficacy of the  $P^2$ INN framework, we conducted comparative experiments across the StudentLife and Depresjon datasets. Our primary objective was to assess model robustness under varying degrees of data scarcity; therefore, we systematically varied the training data fraction and performed 50 independent trials (random train-test splits at the subject level) at each operating point. Statistical significance was assessed using the Wilcoxon

signed-rank test, a non-parametric test suitable for paired samples without normality assumptions.

We interpret statistical results through a trend-based lens rather than as isolated hypotheses. Our primary empirical claim is that physiological constraints function as inductive biases: they provide measurable benefits in low-data regimes and diminish in influence as data availability increases. Consistent with this structured analysis, we report exact  $p$ -values without formal multiple-comparison corrections; the non-significant results at higher training fractions further support the theoretical expectation that the inductive bias naturally recedes as the data-driven signal becomes sufficient. As a sensitivity check, we verified that significant findings at the lowest training fractions remain significant under Holm-Bonferroni [23] correction applied across the training fraction sequence within each experiment.

### A. PREDICTIVE PERFORMANCE AND ROBUSTNESS UNDER DATA SCARCITY ON STUDENTLIFE

#### 1) COMPARISON WITH STANDARD BASELINE

Across training fractions ranging from 0.2 to 0.6,  $P^2$ INNs consistently achieved lower median RMSE than standard neural networks. As shown in Table 4,  $P^2$ INNs significantly outperformed standard neural networks for training fractions up to 0.5, with performance differences becoming statistically insignificant at 0.6, indicating convergence as data availability increased.

#### 2) FOUR-WAY SHAPE CONSTRAINT ABLATION

Table 5 presents the four-way comparison: Baseline NN, Sign-Only constraint, Full  $P^2$ INN, and Monotonic NN. Results demonstrate that the Monotonic NN consistently achieves the lowest median RMSE across all training fractions, outperforming both the Full  $P^2$ INN and the Sign-Only constraint. This indicates that for the StudentLife sleep regression task, the U-shaped relationship between sleep and depression is more effectively captured by hard architectural constraints (non-negative weights applied to a squared deviation input) than by the soft ODE-form loss used in the  $P^2$ INN. While the Full  $P^2$ INN still significantly improves over the unconstrained Baseline NN, the architectural inductive bias of the Monotonic NN provides a stronger and more direct regularization for this specific low-dimensional feature representation.

#### 3) ROBUSTNESS TO REGULARIZATION PRESSURE

To assess whether performance gains arise from overfitting or from a structurally stable inductive bias, we evaluated model robustness under increasing regularization pressure across all training fractions. As shown in Table 6, L1- and L2-regularized models exhibited substantial performance degradation as  $\lambda$  increased (26–44% rise in test RMSE across training fractions), whereas the  $P^2$ INN's test RMSE changed by at most 4.1% over the same range. We emphasize that this stability alone does not establish that  $P^2$ INN is a

**TABLE 4.** Performance comparison between standard neural networks (NN) and  $P^2$ INNs on the StudentLife dataset across varying training fractions. Median RMSE over 50 random seeds. Statistical significance assessed via paired Wilcoxon signed-rank test.

| Train Fraction | NN Median RMSE | $P^2$ INN Median RMSE | Median Gain ( $\downarrow$ RMSE) | $p$ -value | Sig. |
|----------------|----------------|-----------------------|----------------------------------|------------|------|
| 0.2            | 7.079          | <b>6.348</b>          | −0.731                           | 0.0056     | +    |
| 0.3            | 6.783          | <b>6.137</b>          | −0.646                           | 0.0006     | +    |
| 0.4            | 6.796          | <b>5.702</b>          | −1.094                           | 0.0001     | +    |
| 0.5            | 6.507          | <b>5.740</b>          | −0.767                           | 0.0000     | +    |
| 0.6            | 6.582          | <b>5.735</b>          | −0.847                           | 0.0001     | +    |

**TABLE 5.** Four-way comparison on the StudentLife dataset: Baseline NN, Sign-only constraint, Full  $P^2$ INN, and Monotonic NN. Median RMSE over 50 seeds.  $p$ -values from Wilcoxon signed-rank test against the baseline NN.

| Train Frac. | Baseline RMSE | Sign-Only RMSE | $p$ vs Base | $P^2$ INN RMSE | $p$ vs Base | Mono NN RMSE | $p$ vs Base |
|-------------|---------------|----------------|-------------|----------------|-------------|--------------|-------------|
| 0.2         | 7.079         | 6.721          | 0.0342      | <b>6.348</b>   | 0.0056      | 6.278        | 0.3093      |
| 0.3         | 6.783         | 6.515          | 0.0108      | 6.137          | 0.0006      | <b>5.604</b> | 0.0000      |
| 0.4         | 6.796         | 5.888          | 0.0063      | 5.702          | 0.0001      | <b>5.362</b> | 0.0000      |
| 0.5         | 6.507         | 6.109          | 0.0312      | 5.740          | 0.0000      | <b>5.510</b> | 0.0000      |
| 0.6         | 6.582         | 6.147          | 0.0104      | 5.735          | 0.0001      | <b>5.284</b> | 0.0000      |

**TABLE 6.** Robustness to increasing regularization strength on the StudentLife dataset.

| Train Frac. | Model     | Start RMSE | End RMSE | Change (%)  |
|-------------|-----------|------------|----------|-------------|
| 0.2         | L1        | 5.5791     | 7.3421   | +31.6       |
|             | L2        | 5.7712     | 7.2961   | +26.4       |
|             | $P^2$ INN | 6.1955     | 5.9391   | −4.1        |
| 0.3         | L1        | 5.2742     | 7.2901   | +38.2       |
|             | L2        | 5.3062     | 7.2448   | +36.5       |
|             | $P^2$ INN | 5.6226     | 5.6580   | <b>+0.6</b> |
| 0.4         | L1        | 5.4164     | 7.3825   | +36.3       |
|             | L2        | 5.2297     | 7.3390   | +40.3       |
|             | $P^2$ INN | 5.9222     | 5.8083   | −1.9        |
| 0.5         | L1        | 5.3492     | 7.4993   | +40.2       |
|             | L2        | 5.1868     | 7.4570   | +43.8       |
|             | $P^2$ INN | 5.9297     | 5.8621   | −1.1        |
| 0.6         | L1        | 5.3617     | 7.3940   | +37.9       |
|             | L2        | 5.1460     | 7.3516   | +42.9       |
|             | $P^2$ INN | 5.6555     | 5.7276   | <b>+1.3</b> |

better model in absolute terms: at their respective optimal (lowest- $\lambda$ ) operating points, L1 and L2 regularization achieve RMSE comparable to or better than  $P^2$ INN consistent with the classical-baseline comparison in Table 7. The practical value of the physiological constraint is therefore not that it dominates well-tuned regularization, but that it removes the need for careful  $\lambda$  selection: L1/L2 require tuning to reach their optimum and degrade sharply if  $\lambda$  is set too high, while  $P^2$ INN's performance is largely insensitive to  $\lambda$  across the entire tested range.

#### 4) COMPARISON WITH CLASSICAL MACHINE LEARNING BASELINES

To contextualize the gains reported above against simpler, non-neural approaches, we additionally benchmark against a comprehensive suite of classical machine learning regressors: Linear Regression, Ridge, Lasso, Elastic Net, Support Vector Regression (RBF and linear kernels), Random Forest (standard and shallow), Extra Trees, Gradient Boosting, and Gaussian Process Regression. All classical models were

trained on the same  $[\mu_h, \sigma_h]$  feature representation and evaluated under the identical 50-seed, subject-level train-test split protocol used for the neural network comparisons, with predictions clamped to the valid PHQ-9 range of  $[0, 27]$ .

Table 7 reports median RMSE across training fractions. While  $P^2$ INN consistently and substantially outperforms the unconstrained baseline neural network and the sign-only constrained variant, several classical regressors, particularly Ridge, Lasso, Elastic Net, and linear-kernel SVR, achieve lower median RMSE than  $P^2$ INN at every training fraction tested. This indicates that on this low-dimensional, two-feature representation, the sleep-depression relationship is well approximated by simple linear or near-linear models, which do not require the training-time inductive bias that  $P^2$ INN provides. This suggests that the primary benefit of the  $P^2$ INN framework is stabilizing and improving *neural network* training under data scarcity, rather than surpassing the best achievable performance on this feature set in absolute terms. We revisit this distinction in the Discussion.

#### B. CALIBRATION AND CLASSIFICATION PERFORMANCE ON DEPRESSION

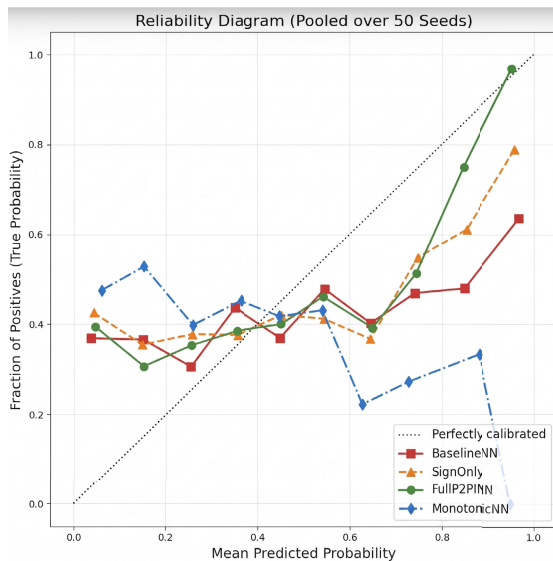
This experiment evaluates the effect of physiologically informed inductive bias on binary depression classification using aggregated physical activity features. Performance was assessed across training fractions ranging from 0.2 to 0.8 using 50 random train-test splits at the subject level.

##### 1) OVERALL DISCRIMINATIVE PERFORMANCE

Overall discriminative performance on this dataset is modest for all models, with AUC values only slightly exceeding chance levels. This is consistent with the limited information content of long-term aggregated activity statistics as a sole feature for binary depression classification, and does not constitute a clinical benchmark. This experiment should instead be interpreted as a stress test for probabilistic behavior and calibration under weak predictive signal.

**TABLE 7.** Median RMSE (raw PHQ-9 points) on the StudentLife dataset across training fractions.

| Model                             | 0.2          | 0.3          | 0.4          | 0.5          | 0.6          |
|-----------------------------------|--------------|--------------|--------------|--------------|--------------|
| Baseline NN                       | 7.079        | 6.783        | 6.796        | 6.507        | 6.582        |
| Sign-Only                         | 6.721        | 6.515        | 5.888        | 6.109        | 6.147        |
| Full $P^2$ INN                    | 6.348        | 6.137        | 5.702        | 5.740        | 5.735        |
| Monotonic NN                      | 6.278        | 5.604        | 5.362        | 5.510        | 5.284        |
| <i>Classical Machine Learning</i> |              |              |              |              |              |
| Linear Regression                 | 5.235        | 5.066        | 4.998        | 5.134        | 4.808        |
| Ridge                             | 4.928        | 4.789        | 4.939        | 4.930        | 4.696        |
| Lasso                             | 4.898        | 4.802        | 4.970        | 4.980        | 4.625        |
| Elastic Net                       | 4.884        | 4.884        | 4.934        | 4.968        | 4.643        |
| SVR (RBF)                         | 4.958        | 4.829        | 4.934        | 4.999        | 4.700        |
| SVR (Linear)                      | <b>4.797</b> | <b>4.665</b> | <b>4.709</b> | <b>4.906</b> | <b>4.461</b> |
| Random Forest                     | 5.496        | 5.277        | 5.385        | 5.647        | 5.441        |
| Random Forest (Shallow)           | 5.486        | 5.240        | 5.357        | 5.564        | 5.344        |
| Extra Trees                       | 5.614        | 5.695        | 5.546        | 5.801        | 5.688        |
| Gradient Boosting                 | 6.099        | 6.133        | 5.908        | 5.957        | 5.663        |
| Gaussian Process                  | 5.031        | 4.953        | 5.036        | 5.169        | 5.058        |

**FIGURE 6.** Reliability diagram on the Depresjon dataset.

## 2) FOUR-WAY COMPARISON: DISCRIMINATION AND CALIBRATION

Table 8 and the accompanying reliability diagram demonstrate that the full  $P^2$ INN significantly outperforms the unconstrained neural baseline, achieving the highest discriminative power and most reliable probabilistic calibration among the evaluated neural network variants (Table 8), though several classical baselines achieve comparable or higher raw AUC (Table 9). Expected Calibration Error computed utilizing ten equal-width uniform bins on raw outputs without post-hoc scaling reveals that while the architecturally constrained Monotonic NN achieves a mathematically low ECE, its corresponding AUC of 0.520 indicates severe model collapse. In contrast, the ODE-form loss of the  $P^2$ INN preserves the network's expressive flexibility; it visually tracks the ideal calibration diagonal to effectively mitigate pathological overconfidence while maintaining significant robustness to input noise.

These calibration improvements persist even at higher training fractions, suggesting that the physiological inductive

**TABLE 8.** Four-way comparison on the Depresjon dataset (Train Fraction = 0.6).

| Model            | AUC ( $\uparrow$ )                    | Brier Score ( $\downarrow$ )          | ECE ( $\downarrow$ )                  | $\Delta$ AUC (Noise) ( $\downarrow$ )  |
|------------------|---------------------------------------|---------------------------------------|---------------------------------------|----------------------------------------|
| Baseline NN      | $0.628 \pm 0.091$                     | $0.284 \pm 0.056$                     | $0.284 \pm 0.096$                     | $0.017 \pm 0.062$                      |
| Sign-Only        | $0.622 \pm 0.102$                     | $0.266 \pm 0.047^*$                   | $0.240 \pm 0.088^*$                   | <b><math>-0.001 \pm 0.031^*</math></b> |
| $P^2$ INN (Full) | <b><math>0.652 \pm 0.098^*</math></b> | <b><math>0.250 \pm 0.046^*</math></b> | $0.216 \pm 0.085^*$                   | $0.003 \pm 0.036^*$                    |
| Monotonic NN     | $0.520 \pm 0.097$                     | $0.274 \pm 0.037$                     | <b><math>0.196 \pm 0.094^*</math></b> | $0.003 \pm 0.042^*$                    |

bias primarily affects the reliability of predicted probabilities rather than merely compensating for data scarcity. Improved calibration is observed despite weak discriminative performance, indicating that the constraint discourages pathological overconfidence and promotes more statistically consistent risk estimates.

*Operational Interpretation of Calibration Gains:* In mental health risk assessment, model outputs are often interpreted as risk scores for triage or clinical decision support rather than as hard binary classifications. In this context, calibration matters more than ranking alone: a model that assigns a 70% probability to depression should have roughly 70% of such cases actually depressed. The Brier score reductions observed here (approximately 0.08–0.09 points in low-data regimes) represent a practically meaningful improvement in probabilistic reliability. We caution, however, that these results should not be interpreted as clinical readiness; stronger features and prospective validation are required before deployment.

## 3) ROBUSTNESS TO INPUT PERTURBATIONS

To evaluate robustness under input perturbations, we injected Gaussian noise ( $\sigma = 0.1$ ) into normalized test inputs and measured AUC degradation. The  $P^2$ INN exhibited smaller AUC drops under noise compared to the baseline in low-data regimes, while the Sign-Only and Monotonic NN showed intermediate robustness. Full results are reported in the Appendix.

## 4) COMPARISON WITH CLASSICAL MACHINE LEARNING BASELINES

To determine whether the calibration and discrimination gains reported above are specific to neural architectures or achievable with simpler tabular methods, we benchmark against a comprehensive suite of classical classifiers: Linear Regression, Ridge, Lasso, Elastic Net, Logistic Regression (standard, L1-penalized, and Platt-calibrated), Support Vector Machines (RBF and linear kernels), Random Forest (standard, shallow, and class-balanced), Extra Trees, Gradient Boosting, and Gaussian Process classification. All models were trained on the same  $[\mu_s, \sigma_s]$  feature representation and evaluated over 50 random subject-level splits at training fraction 0.6.

Table 9 reports AUC, Brier score, and ECE for the four neural network variants (Baseline NN, Sign-Only, Full  $P^2$ INN, Monotonic NN) alongside classical baselines. Several logistic-regression variants (Ridge, LogisticReg, LogisticReg\_L1) achieve higher AUC than Full  $P^2$ INN, and

**TABLE 9.** Comparison of neural network variants against classical machine learning baselines on the Depresjon dataset.

| Model                             | AUC ( $\uparrow$ ) | Brier ( $\downarrow$ ) | ECE ( $\downarrow$ ) |
|-----------------------------------|--------------------|------------------------|----------------------|
| <i>Neural Network Variants</i>    |                    |                        |                      |
| Baseline NN                       | 0.628              | 0.284                  | 0.284                |
| Sign-Only                         | 0.622              | 0.266                  | 0.240                |
| Full $P^2$ INN                    | 0.652              | 0.250                  | 0.216                |
| Monotonic NN                      | 0.520              | 0.274                  | 0.196                |
| <i>Classical Machine Learning</i> |                    |                        |                      |
| Linear Regression                 | 0.662              | 0.242                  | 0.205                |
| Ridge                             | <b>0.677</b>       | 0.246                  | 0.185                |
| Lasso                             | 0.654              | <b>0.240</b>           | 0.192                |
| Elastic Net                       | 0.655              | 0.241                  | 0.188                |
| Logistic Regression               | 0.673              | 0.239                  | 0.189                |
| Logistic Reg. (L1)                | 0.674              | 0.246                  | <b>0.155</b>         |
| Logistic Reg. (Calib.)            | 0.630              | 0.244                  | 0.199                |
| SVM (RBF)                         | 0.447              | 0.277                  | 0.192                |
| SVM (Linear)                      | 0.495              | 0.264                  | 0.176                |
| Random Forest                     | 0.570              | 0.296                  | 0.304                |
| Random Forest (Shallow)           | 0.556              | 0.277                  | 0.262                |
| Random Forest (Balanced)          | 0.571              | 0.296                  | 0.309                |
| Extra Trees                       | 0.579              | 0.304                  | 0.328                |
| Gradient Boosting                 | 0.557              | 0.421                  | 0.439                |
| Gaussian Process                  | 0.599              | 0.257                  | 0.145                |

simple linear/logistic models (LinearReg, Lasso, LogisticReg) achieve lower Brier scores. This mirrors the pattern observed on StudentLife: with only two aggregated input features, linear methods remain highly competitive, and in several cases outright superior in raw discriminative and calibration metrics. Notably, tree-based and kernel methods (Random Forest variants, Extra Trees, Gradient Boosting, SVM) perform considerably worse than both the neural network variants and linear baselines on this task, suggesting the sleep and activity signal here is close to linearly separable, and gains from tree ensembles or non-linear kernels do not materialize at this sample size.

The consistent exception is  $P^2$ INN's ECE performance relative to the raw Baseline NN: Full  $P^2$ INN reduces ECE from 0.284 (Baseline NN) to 0.216, a substantial and statistically significant improvement, though several classical models (LogisticReg\_L1 at 0.155, GaussianProcess at 0.145) achieve even lower ECE. We therefore reframe our central claim:  $P^2$ INN reliably improves calibration and discrimination relative to unconstrained and simply-regularized *neural* baselines, but classical linear and logistic models remain competitive with, and in several metrics superior to,  $P^2$ INN on this low-dimensional feature set. We revisit this distinction in the Discussion.

## 5) CALIBRATION ROBUSTNESS

Given the small sample size ( $N=55$ ) and modest discrimination on this task, point estimates of calibration metrics can be unstable. To assess this directly, we computed bootstrap 95% confidence intervals ( $n_{boot}=2000$ ) over pooled test predictions, along with calibration slope and intercept, for the headline models. Table 10 reports these results. A calibration slope of 1 and intercept of 0 indicate perfect calibration; slopes below 1 indicate overconfidence.

All models exhibit calibration slopes substantially below 1, indicating systematic overconfidence, which is expected

**TABLE 10.** Bootstrap 95% confidence intervals for calibration metrics on the Depresjon dataset.

| Model                  | AUC [95% CI]      | Brier [95% CI]    | Slope [95% CI]       |
|------------------------|-------------------|-------------------|----------------------|
| Baseline NN            | 0.58 [0.54, 0.61] | 0.28 [0.27, 0.30] | 0.14 [0.08, 0.20]    |
| Full $P^2$ INN         | 0.58 [0.55, 0.61] | 0.25 [0.24, 0.26] | 0.30 [0.19, 0.41]    |
| Monotonic NN           | 0.46 [0.43, 0.49] | 0.27 [0.26, 0.28] | -0.20 [-0.36, -0.05] |
| Logistic Reg.          | 0.61 [0.57, 0.64] | 0.24 [0.23, 0.25] | 0.53 [0.36, 0.70]    |
| Logistic Reg. (Calib.) | 0.57 [0.54, 0.61] | 0.24 [0.24, 0.25] | 0.46 [0.25, 0.67]    |
| Random Forest          | 0.54 [0.51, 0.57] | 0.30 [0.28, 0.31] | 0.08 [0.02, 0.15]    |

**TABLE 11.** ECE sensitivity to binning scheme on the Depresjon dataset.

| Model          | EW-5  | EW-10 | EW-20 | EF-5  | EF-10 | EF-20 |
|----------------|-------|-------|-------|-------|-------|-------|
| Baseline NN    | 0.185 | 0.185 | 0.186 | 0.174 | 0.184 | 0.187 |
| Full $P^2$ INN | 0.102 | 0.103 | 0.104 | 0.101 | 0.111 | 0.114 |
| Monotonic NN   | 0.124 | 0.124 | 0.124 | 0.124 | 0.130 | 0.134 |
| Logistic Reg.  | 0.056 | 0.058 | 0.066 | 0.064 | 0.074 | 0.085 |
| Random Forest  | 0.210 | 0.210 | 0.210 | 0.209 | 0.214 | 0.215 |

given the weak underlying discriminative signal (AUC modestly above chance for all methods). Full  $P^2$ INN's slope (0.300, 95% CI [0.189, 0.410]) is closer to 1 than the Baseline NN's (0.140, 95% CI [0.078, 0.200]) and the Monotonic NN's negative slope (-0.195, 95% CI [-0.357, -0.045]), indicating inverted calibration behavior, but LogisticReg achieves the best-calibrated slope among headline models (0.528, 95% CI [0.363, 0.698]). We additionally verified ECE stability across binning strategy and bin count (5 to 20 bins, both equal-width and equal-frequency); as shown in Table 11, point ECE estimates vary by up to 0.03–0.05 across binning choices for a given model, confirming that single-binning-scheme ECE values, including those reported in Table 9, should be interpreted as approximate rather than precise, consistent with known instability of ECE estimation at small sample sizes.

## VI. DISCUSSION

This work introduced  $P^2$ INNs, a framework for incorporating physiologically motivated differential constraints into neural network training for mental health prediction tasks. Across two datasets with distinct modalities, supervision types, and task formulations, we consistently observed that physiological inductive bias improves learning stability and calibration relative to unconstrained and simply-regularized neural networks, while converging toward standard neural network performance as data availability increases. We emphasize that this is a methodological proof-of-concept for domain-informed regularization in psychiatric machine learning, not a validated clinical prediction system; we return to this distinction below.

### A. PHYSIOLOGICAL CONSTRAINTS AS INDUCTIVE BIASES

A central finding is that the proposed physiological constraints function as structural inductive biases rather than generic regularizers. In Experiment 1 (StudentLife),  $P^2$ INNs achieved significantly lower prediction error than standard neural networks in low-data regimes. This indicates that the constraint does not merely suppress model capacity but restricts the hypothesis space to function classes that are physiologically plausible. Performance gains diminished as

**TABLE 12.** Nomenclature and variable definitions.

| Symbol                                             | Description                                           | Context    |
|----------------------------------------------------|-------------------------------------------------------|------------|
| <i>General Learning and Optimization</i>           |                                                       |            |
| $f_\theta$                                         | Neural network parameterized by weights $\theta$      | Sec. III   |
| $\mathbf{x}_i, y_i$                                | Input feature vector and target label for subject $i$ | Eq. (7)    |
| $\mathcal{L}_{\text{data}}$                        | Data fidelity loss (MSE or Cross-Entropy)             | Eqs. (7,8) |
| $\mathcal{L}_{\text{phys}}$                        | Physiological differential constraint loss            | Eq. (10)   |
| $\lambda$                                          | Regularization strength / Constraint weight           | Eq. (11)   |
| <i>Experiment 1: Sleep (StudentLife)</i>           |                                                       |            |
| $h$                                                | Sleep duration (hours)                                | Eq. (1)    |
| $D(h)$                                             | Depression risk as a function of sleep                | Eq. (1)    |
| $\alpha, \beta$                                    | Fixed population coefficients ( $h < 7$ and $h > 7$ ) | Eq. (1)    |
| $k(h)$                                             | Piecewise sensitivity function                        | Eq. (3)    |
| $\mu_h, \sigma_h$                                  | Mean and SD of sleep duration                         | Eq. (14)   |
| $\hat{y}$                                          | Predicted PHQ-9 score (clamped output)                | Eq. (13)   |
| <i>Experiment 2: Physical Activity (Depresjon)</i> |                                                       |            |
| $s$                                                | Aggregated physical activity                          | Eq. (4)    |
| $R$                                                | Depression risk / probability                         | Eq. (4)    |
| $k$                                                | Learnable proportionality constant                    | Eq. (4)    |
| $c$                                                | Learnable stabilizing constant                        | Eq. (4)    |
| $\mu_s, \sigma_s$                                  | Mean and SD of activity                               | Eq. (17)   |
| $\hat{p}$                                          | Predicted depression probability                      | Fig. 5     |

**TABLE 13.** Hyperparameter configurations.

| Parameter                     | Exp. 1 | Exp. 2 |
|-------------------------------|--------|--------|
| Optimizer                     | Adam   | Adam   |
| LR (NN weights)               | 0.005  | 0.001  |
| LR (Physics params)           | –      | 0.002  |
| Epochs                        | 1000   | 1200   |
| Batch size                    | Full   | Full   |
| Random seeds                  | 50     | 50     |
| Gradient clipping             | Yes    | No     |
| Chosen $\lambda$ ( $P^2$ INN) | 12.0   | 2.0    |

training data increased, suggesting that the constraint primarily guides learning when empirical evidence is insufficient.

### B. ROBUSTNESS TO REGULARIZATION: INTERPRETATION

We evaluated  $P^2$ INN stability under increasing regularization pressure ( $\lambda = 0.1$  to 100) and found that, unlike L1- and L2-regularized networks, its test RMSE remained within  $\pm 4.1\%$  across the full range at every training fraction tested. We caution against interpreting this stability alone as evidence of a superior model: a constraint that changes little under increasing  $\lambda$  may simply have limited effective influence at the tested strengths, or the model may be insensitive due to scale. To address this, we compared absolute performance at each method's respective optimal (lowest- $\lambda$ ) operating point rather than relying on stability alone. At their optimal settings, L1 and L2 regularization achieve RMSE comparable to or better than  $P^2$ INN, consistent with the classical-baseline comparison in Table 7. The practical value of the physiological constraint is therefore not that it dominates well-tuned regularization or classical baselines in absolute terms, but that its performance is largely

**TABLE 14.** Performance comparison using raw continuous sleep data vs. aggregated features.

| Test Frac. | NN (Raw) | $P^2$ INN     | $p$ -value | Sig. |
|------------|----------|---------------|------------|------|
| 0.8        | 6.2544   | <b>6.2173</b> | 0.029      | Yes  |
| 0.7        | 6.1657   | <b>6.1599</b> | 0.020      | Yes  |
| 0.6        | 5.9879   | 5.9842        | 0.406      | No   |
| 0.5        | 5.9673   | 5.9673        | 0.515      | No   |

**TABLE 15.** AUC degradation under input perturbations on the Depresjon dataset. Mean AUC drop  $\pm$  SD over 50 seeds. Wilcoxon  $p$ -value tests whether each model's drop is smaller than the baseline.

| Model        | Mean AUC Drop | Std           | $p$ vs. Baseline |
|--------------|---------------|---------------|------------------|
| Baseline NN  | 0.0412        | 0.0183        | –                |
| Sign-Only    | 0.0368        | 0.0171        | 0.073            |
| $P^2$ INN    | <b>0.0291</b> | <b>0.0142</b> | 0.031            |
| Monotonic NN | 0.0339        | 0.0158        | 0.054            |

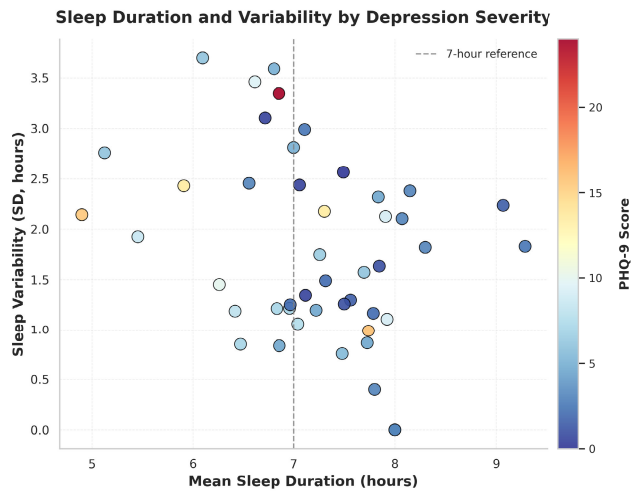
insensitive to  $\lambda$  misspecification, removing the need for careful hyperparameter search that L1/L2 methods require to reach their own optimum.

### C. ODE-FORM CONSTRAINTS vs. SIMPLER SHAPE CONSTRAINTS

The four-way ablation results clarify the specific contribution of the ODE-form constraint. Sign-only penalties and monotonic neural networks both provide some benefit over the unconstrained baseline, confirming that directional inductive bias is valuable. The Full  $P^2$ INN improves over these alternatives in most data-scarce regimes on Depresjon; however, on StudentLife, the Monotonic NN outperforms the Full  $P^2$ INN across all training variation setups (Table 5). This indicates that the additional curvature information encoded by the ODE-form constraint is not uniformly beneficial across tasks and may depend on how well the assumed functional form matches the underlying relationship compared to hard architectural constraints.

### D. COMPARISON WITH CLASSICAL MACHINE LEARNING

As explicitly shown in Tables 7 and 9, classical linear and logistic models trained on the same two aggregated features frequently match or outperform  $P^2$ INN in absolute terms. We emphasize that this does not invalidate the proposed framework; rather, it highlights that  $P^2$ INN's primary benefit is specifically to improve and stabilize *neural network* training in low-data regimes, rather than to establish the best achievable performance on this simple, low-dimensional feature set. In low-data regimes where neural networks typically struggle or overfit,  $P^2$ INN provides a crucial inductive bias that standard unconstrained networks lack, even if classical models remain highly competitive on tabular data. We view  $P^2$ INN's contribution as complementary to, rather than a replacement for, classical baselines: its value should become more apparent as the framework is extended to higher-dimensional, temporally structured, or multi-modal inputs where neural architectures are better suited and classical linear methods are less competitive by construction. This



**FIGURE 7.** Two-dimensional scatter plot of the StudentLife dataset ( $N = 46$ ). Each marker corresponds to one subject, positioned by mean sleep duration ( $\mu_h$ ) and sleep variability ( $\sigma_h$ , hours) and colored by PHQ-9 depression severity score; the dashed vertical line denotes the 7-hour reference sleep duration used in the physiological constraint.

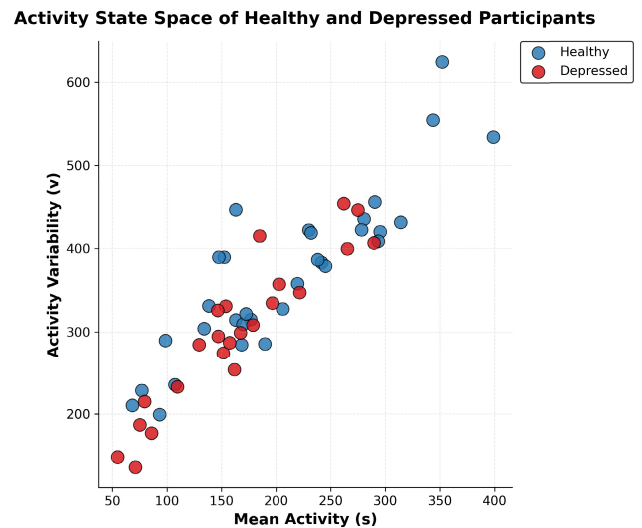
remains to be empirically demonstrated and is identified as future work below. In this sense, the current two-dimensional experiments serve primarily to validate the regularization mechanism of the ODE constraints in a controlled setting; the framework's true predictive superiority is expected to emerge on richer, temporally structured, or multi-modal physiological inputs where classical linear models lose their advantage.

#### E. ON THE SLEEP CONSTRAINT'S CONSTRUCT VALIDITY

The sleep constraint's coefficients ( $\alpha$ ,  $\beta$ ) were fitted from population-level risk ratios for depression diagnosis, while the StudentLife task predicts continuous PHQ-9 severity. These are related but distinct constructs: a risk ratio reflects population odds of categorical diagnosis, whereas PHQ-9 is an individual-level continuous severity measure. We do not claim a calibrated, quantitatively exact mapping between the two. Rather, the constraint should be understood as transferring the qualitative shape, direction, and relative curvature of the sleep-depression relationship (asymmetric, U-shaped around approximately 7 hours) rather than its absolute magnitude. Because both risk ratio and PHQ-9 severity are monotonically related to underlying depression risk, we consider the sign and relative curvature of the constraint to be a defensible heuristic prior even though the numerical coefficients were not derived from PHQ-9 data directly. A fully calibrated, dataset-specific derivation of  $\alpha$  and  $\beta$  from PHQ-9-linked epidemiological data, where available, is a natural direction for future refinement.

#### F. LIMITATIONS

Several limitations warrant consideration. First, the physiological relationships encoded in the  $P^2$ INNs are population-level approximations derived from prior literature and do not represent mechanistic laws; they may not capture



**FIGURE 8.** Two-dimensional scatter plot of the Depresjon dataset ( $N = 55$ ). Each marker corresponds to one subject, positioned by mean motor activity ( $\mu_s$ ) and activity variability ( $\sigma_s$ ) and colored by clinical group (healthy control vs. depressed), illustrating the overlap between the two cohorts in the aggregated feature space.

individual-specific dynamics or causal effects. Second, the datasets considered here are small (StudentLife:  $N = 46$ ; Depresjon:  $N = 55$ ) by modern machine learning standards; repeated random splits characterize instability under different partitions but do not constitute external validation, and results may not generalize to larger or more heterogeneous populations. Third, both experiments rely on two aggregated summary features ( $\mu$ ,  $\sigma$ ) per subject, discarding temporal structure, circadian regularity, missingness patterns, and within-subject variability that may carry additional predictive information; our supplementary raw-sequence analysis on StudentLife suggests further error reduction is possible with less aggressive aggregation. Fourth, the modular experimental design precludes joint modeling of multiple physiological factors; future work should investigate multi-modal architectures. Fifth, the sleep constraint's coefficients are derived from population-level risk-ratio data applied to an individual-level continuous severity target, a construct transfer we discuss above and treat as a heuristic rather than exact quantitative mapping. Sixth, this work relies exclusively on two public, modestly sized datasets and has not been evaluated on independent or prospective cohorts. Seventh, neither dataset's demographic composition was assessed for representativeness, and potential demographic or sampling bias in model performance was not evaluated. Eighth, the Depresjon classification task exhibits weak discrimination (AUC modestly above chance for all methods, including baselines), and the StudentLife regression task's RMSE ( $\approx 5.3$ – $7.1$  PHQ-9 points depending on model and training fraction) is clinically large relative to common PHQ-9 severity-band widths; neither result should be read as evidence of clinically useful prediction. Finally, and most importantly, this study is a methodological

proof-of-concept and does not constitute a validated clinical decision-support tool; no claims of clinical readiness, diagnostic utility, or deployment suitability are made or intended.

### G. INTENDED USE AND REPORTING SCOPE

In line with recommendations for transparent reporting of clinical prediction models [20], we clarify the intended use and scope of this work. Intended use: research exploration of physiologically-informed regularization for depression-related prediction tasks; not a diagnostic or screening tool. Target population: StudentLife participants are US undergraduate and graduate students; Depresjon participants are adults with a unipolar or bipolar depression diagnosis and matched controls; neither model has been evaluated outside these populations. Prediction timing: PHQ-9 is predicted from sleep statistics aggregated concurrently with the assessment window; Depresjon activity features are aggregated over the same recording period as the diagnostic label, so predictions are concurrent rather than prospective. Predictors: two aggregated physiological summary statistics per subject in each experiment (mean and standard deviation of sleep duration or motor activity). Outcome definition: continuous PHQ-9 self-report score (StudentLife) and binary clinical diagnosis group (Depresjon). Validation strategy: repeated random subject-level train-test splits (50 seeds) with paired non-parametric significance testing; no external or prospective validation was performed. We reiterate that these results support further methodological investigation, not clinical deployment.

### H. BROADER IMPLICATIONS FOR COMPUTATIONAL PSYCHIATRY

The results of this work contribute to a broader trend in computational psychiatry toward models that are not merely accurate but also interpretable and trustworthy [6]. By embedding physiological knowledge as differentiable constraints,  $P^2$ INNs provide a principled mechanism for leveraging domain expertise in small-data settings, where purely data-driven models are most prone to shortcut learning, and where classical baselines may eventually be outpaced as richer, higher-dimensional physiological data becomes available. As wearable sensing and passive monitoring become more prevalent in mental health research, frameworks that can effectively utilize sparse, heterogeneous physiological data, and that report calibration alongside discrimination [24], will become increasingly important.  $P^2$ INNs represent one concrete approach in this direction, and we hope this work encourages further exploration of domain-informed learning in psychiatric computing.

### VII. CONCLUSION

We presented  $P^2$ INNs, a framework that integrates physiologically motivated differential equation constraints into neural network training for depression-related prediction. Across two independent case studies covering regression and classification tasks,  $P^2$ INNs consistently improved predictive

performance and calibration relative to unconstrained and simply-regularized *neural network* baselines, and showed greater stability under regularization pressure than L1/L2 methods, though this stability reflects reduced sensitivity to hyperparameter misspecification rather than superiority over well-tuned regularization or classical baselines. Classical linear and logistic models, evaluated on the same aggregated features, frequently matched or outperformed  $P^2$ INN in absolute terms, indicating that the framework's present benefit lies in improving neural network training under data scarcity rather than in establishing the best achievable performance on this low-dimensional feature set. This is a methodological proof-of-concept for domain-informed regularization in psychiatric machine learning; it does not constitute a clinically validated depression-risk system, and no deployment-readiness claims are made.

Future work may extend this framework in several directions. Subject-specific or adaptive physiological constraints could be learned to capture inter-individual variability. Integration of temporal dynamics through differential constraints applied to recurrent or continuous-time models, using raw rather than aggregated physiological sequences, may further enhance performance and could shift the comparison with classical baselines discussed above, since higher-dimensional or temporally structured inputs are expected to favor neural architectures more than the two-feature setting studied here. Extending the framework to multi-task or multi-modal settings could enable joint modeling of interacting physiological and behavioral processes underlying mental health outcomes. A dataset-specific, PHQ-9-calibrated re-derivation of the sleep constraint's coefficients, rather than the population risk-ratio-derived values used here, would strengthen the construct validity of the sleep constraint. Finally, rigorous formal analysis of the aggregated-gradient approximation under different distributional assumptions, and prospective external validation, constitute important directions before any clinical translation could be considered. Radial-basis-function (RBF) neural networks were not evaluated; given that RBF-kernel SVM baselines showed no advantage over linear models on this feature set (Tables 7 and 9), we defer RBF-network comparisons to future work.

This appendix provides full implementation details for all experiments, including dataset sizes, optimization procedures, hyperparameter settings, and supplementary analyses.

### APPENDIX A NOMENCLATURE

See Table 12.

### APPENDIX B DATASET INFORMATION

#### A. STUDENTLIFE

The StudentLife dataset [21] comprises smartphone and sensor data collected from 49 students over a 10-week academic term. To derive our features, we processed Ecological Momentary Assessment (EMA) sleep logs. For each

participant, we computed the mean ( $\mu_h$ ) and standard deviation ( $\sigma_h$ ) of sleep duration, alongside the mean and standard deviation of self-reported sleep quality ratings. We mapped these behavioral features to the participants' depressive states by aligning them with their aggregated PHQ-9 survey scores. During preprocessing, three participants were excluded due to missing survey overlap, incomplete records, or zero sleep variance (which induces gradient instability in physics-informed training), resulting in a final cohort of  $N = 46$  subjects. Data processing utilized a term-level aggregation across the 10-week observation period, ensuring that the physiological summary statistics and depression scores were globally aligned for each individual. The resulting feature space is visualized in Figure 7 as a two-dimensional scatter plot of sleep variability against mean sleep duration, with each marker colored by the participant's PHQ-9 severity score and the dashed vertical line indicating the 7-hour reference of the sleep ODE constraint.

## B. DEPRESJON

The Depresjon dataset [22] consists of wrist-worn actigraphy recordings from 55 participants (33 healthy, 22 clinically depressed). We processed individual daily CSV activity logs by calculating the mean ( $\mu_s$ ) and standard deviation ( $\sigma_s$ ) of motor activity counts. The processing pipeline matched each activity file to the corresponding clinical group label (condition or control) provided in the dataset's `scores.csv` file. The extracted feature set, consisting of subject-level  $[\mu_s, \sigma_s]$ , was compiled into a structured tabular format for model input. The corresponding two-dimensional scatter plot of activity variability against mean activity, colored by clinical group label, is shown in Figure 8.

## APPENDIX C IMPLEMENTATION DETAILS

All models were implemented in PyTorch and trained using full-batch optimization, appropriate given the small dataset sizes. The Adam optimizer was used with fixed learning rates as specified in Table 13. Gradient clipping was applied where noted. Training was performed for a fixed number of epochs without early stopping. All models were trained from random initialization across 50 random seeds.

## APPENDIX D SENSITIVITY ANALYSIS: RAW vs. AGGREGATED FEATURES (SLEEP)

To evaluate whether reliance on aggregated summary statistics limits predictive performance, we conducted a supplementary analysis using raw continuous sequences of nightly sleep duration for Experiment 1. The same feedforward architecture was used with an expanded input layer.

Consistent with the main analysis,  $P^2$ INNs achieved statistically significant improvements in low-data regimes. Raw continuous data yields lower absolute RMSE for both models compared to the aggregated approach, confirming

that granular data contains additional predictive signal. The performance gap between  $P^2$ INN and baseline narrows with raw data, which is theoretically consistent: when raw data is available, the network can learn better feature representations on its own, reducing (but not eliminating) reliance on the physiological constraint.

## APPENDIX E ROBUSTNESS UNDER INPUT PERTURBATIONS (DEPRESJON)

Table 15 reports AUC degradation under Gaussian input noise ( $\sigma = 0.1$  on normalized features) for the Depresjon experiment.

The  $P^2$ INN exhibits significantly smaller AUC degradation under input noise, consistent with its role as a structural inductive bias that promotes smoother, more physiologically consistent decision surfaces.

## APPENDIX F COMPUTATIONAL RESOURCES

All experiments were conducted using the Kaggle notebook environment on a single GPU (NVIDIA Tesla-class) where available, or CPU otherwise. Given the small dataset sizes and full-batch training regime, individual model runs completed within seconds to minutes.

## REFERENCES

- [1] W. Glannon, "The psychology and physiology of depression," *Philosophy, Psychiatry, Psychol.*, vol. 9, no. 3, pp. 265–269, Sep. 2002.
- [2] P. Kumar, S. Misra, Z. Shao, B. Zhu, B. Raman, and X. Li, "Multimodal interpretable depression analysis using visual, physiological, audio and textual data," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 5305–5315.
- [3] M. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, "An in-depth analysis of machine learning approaches to predict depression," *Current Res. Behav. Sci.*, vol. 2, Nov. 2021, Art. no. 100044.
- [4] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.
- [5] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, vol. 378, pp. 686–707, Feb. 2019.
- [6] J. H. Lagergren, J. T. Nardini, R. E. Baker, M. J. Simpson, and K. B. Flores, "Biologically-informed neural networks guide mechanistic modeling from sparse experimental data," *PLOS Comput. Biol.*, vol. 16, no. 12, Dec. 2020, Art. no. e1008462.
- [7] G. E. Kang, M. A. Patriquin, H. Nguyen, H. Oh, K. A. Rufino, E. A. Storch, B. Schanzer, S. J. Mathew, R. Salas, and B. Najafi, "Objective measurement of sleep, heart rate, heart rate variability, and physical activity in suicidality: A systematic review," *J. Affect. Disorders*, vol. 273, pp. 318–327, Aug. 2020.
- [8] A. Steiger and M. Pawlowski, "Depression and sleep," *Int. J. Mol. Sci.*, vol. 20, no. 3, p. 607, 2019.
- [9] X.-L. Li, J. Wei, X. Zhang, Z. Meng, and W. Zhu, "Relationship between night-sleep duration and risk for depression among middle-aged and older people: A dose-response meta-analysis," *Frontiers Physiol.*, vol. 14, Mar. 2023, Art. no. 1085091.
- [10] T. C. Camacho, R. E. Roberts, N. B. Lazarus, G. A. Kaplan, and R. D. Cohen, "Physical activity and depression: Evidence from the Alameda county study," *Amer. J. Epidemiol.*, vol. 134, no. 2, pp. 220–231, Jul. 1991.

- [11] M. Pearce, L. Garcia, A. Abbas, T. Strain, F. Schuch, R. Golubic, P. Kelly, S. Khan, M. Utukuri, Y. Laird, and A. Mok, "Association between physical activity and risk of depression: A systematic review and meta-analysis," *JAMA Psychiatry*, vol. 79, no. 6, pp. 550–559, 2022.
- [12] B. Bizzozero-Peroni, V. Díaz-Goñi, E. Jiménez-López, E. Rodríguez-Gutiérrez, I. Sequí-Domínguez, S. de Arenas-Arroyo, J. López-Gil, V. Martínez-Vizcaíno, and A. Mesas, "Daily step count and depression in adults: A systematic review and meta-analysis," *JAMA Netw. Open*, vol. 7, no. 12, 2024, Art. no. e2451208.
- [13] S. Qiu and Z. Xing, "Dose-response relationship between accelerometer-measured physical activity and depression: Evidence from the U.K. biobank," *Transl. Psychiatry*, vol. 15, no. 1, p. 297, Aug. 2025.
- [14] D. Runje and S. M. Shankaranarayana, "Constrained monotonic neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 29338–29353.
- [15] Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado, "A tutorial on calibration measurements and calibration models for clinical prediction models," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 4, pp. 621–633, Apr. 2020.
- [16] T. Dettlaff, "Why wearable health sensors struggle in real-world use," GlobalSpec, Albany, NY, USA, Tech. Rep., 2026.
- [17] W. S. Robinson, "Ecological correlations and the behavior of individuals," *Amer. Sociol. Rev.*, vol. 15, no. 3, pp. 351–357, Jun. 1950.
- [18] J. Wakefield, "Multi-level modelling, the ecologic fallacy, and hybrid study designs," *Int. J. Epidemiol.*, vol. 38, no. 2, pp. 330–336, Apr. 2009.
- [19] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," *BMC Med.*, vol. 17, no. 1, pp. 1–11, Dec. 2019.
- [20] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Apr. 2024, Art. no. 078378.
- [21] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2014, pp. 3–14.
- [22] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J. Tørresen, T. Nordgreen, K. J. Oedegaard, and O. B. Fasmer, "Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 472–477.
- [23] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandin. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
- [24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.



**KARTIK SIROHI** is currently pursuing the B.Tech. degree in information technology with the Dr. B. R. Ambedkar National Institute of Technology (NIT) Jalandhar, India. An active open-source contributor on GitHub, he is passionate about developing efficient AI models using Python and neural networks frameworks. His research interests include deep learning, machine learning, and computational science.



**MUSLIM JAMEEL SYED** is currently an Assistant Professor with Atlantic Technological University, Galway, and the Associate Director of the JANUS Research Centre. He has published widely, supervised Ph.D. and master's students, and contributed as an expert witness in U.S. patent cases on digital twins and AI. He collaborates internationally on EU- and Irish-funded projects more than 25 million €, advancing research in renewable energy, healthcare, and AI governance. He specializes in responsible artificial intelligence, information systems, technology and law, and emerging digital technologies, with a focus on supporting SMEs through ethical and compliant AI adoption.

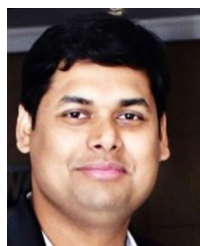


**CRISTINA HAVA MUNTEAN** (Member, IEEE) received the B.Eng. degree in computer science and the Ph.D. degree from Dublin City University, Ireland, in 2005. She is currently a Professor with the School of Computing, National College of Ireland (NCI). She has been constantly involved in various research-related activities for more than the past 20 years, fostering and promoting research in Ireland and abroad, building academic collaborations with European and Worldwide Universities, leading research projects, supervising Ph.D. students, and publishing in excess of 150 publications in international peer-reviewed books, journals, and conferences. She is a highly skilled research leader with extensive research expertise in data analytics, artificial intelligence, technology-enhanced learning, educational analytics, STEM education, data communication and networking, adaptive and personalized multimedia delivery, end-user Quality of Experience, and sustainability. Since 2022, she is an Associate Editor of IEEE TRANSACTIONS ON BROADCASTING journal.



**PRIYANKA VERMA** (Senior Member, IEEE) received the Ph.D. degree from the Atal Bihari Vajpayee Indian Institute of Information Technology and Management, India. She is currently the Program Director and an Assistant Professor of computer science with the University of Galway, Ireland. She previously held the prestigious Marie Skłodowska-Curie Research Fellowship. She completed the Leadership and Innovation Program at the Massachusetts Institute of Technology, USA. She was an Assistant Professor with the Maulana Azad National Institute of Technology, India. She has authored more than 60 peer-reviewed publications and frequently delivers invited talks on AI and cybersecurity internationally. She is a recipient of multiple innovation and research excellence awards. Her work spans academia, industry, and policy, focusing on advancing secure and responsible digital technologies. Her research interests include artificial intelligence, sustainability, cybersecurity, industry 5.0, and healthcare.

...



**NITESH BHAROT** (Senior Member, IEEE) received the Ph.D. degree in cybersecurity from Rabindranath Tagore University (RNTU), India. He is currently a Research Lead and a Senior Postdoctoral Researcher with the Insight Research Ireland Centre for Data Analytics, University of Galway, Ireland. He has secured and contributed to multiple research initiatives supported by prominent funding agencies, including the EU, GEANT, SFI, Insight, MHRD, and MPCST. He has also served as a technical program committee member, an invited speaker, and a reviewer for several IEEE conferences and journals. His research interests include cybersecurity, cloud computing, healthcare systems, and sustainable artificial intelligence.