

Predicting and Understanding College Student Mental Health with Interpretable Machine Learning

Meghna Roy Chowdhury
Purdue University
West Lafayette, IN, USA
mroycho@purdue.edu

Wei Xuan
University of Southern California
Los Angeles, CA, USA
wxuan@usc.edu

Shreyas Sen
Purdue University
West Lafayette, IN, USA
shreyas@purdue.edu

Yixue Zhao
USC Information Sciences Institute
Arlington, VA, USA
yzhao@isi.edu

Yi Ding
Purdue University
West Lafayette, IN, USA
yiding@purdue.edu

Abstract

Mental health issues among college students have reached critical levels, affecting both academic performance and overall wellbeing. Predicting and understanding mental health status among college students is challenging due to three key barriers: the lack of large-scale longitudinal datasets, the prevalence of black-box machine learning models that offer little transparency, and a reliance on population-level analysis rather than personalized understanding.

To tackle these challenges, this paper presents I-HOPE, the first Interpretable Hierarchical model for Personalized mental health prediction. I-HOPE is a two-stage hierarchical model that connects raw behavioral features to mental health status through five defined behavioral categories as *interaction labels*. We evaluate I-HOPE on the College Experience Study, the longest longitudinal mobile sensing dataset. This dataset spans five years and captures data from both pre-pandemic periods and the COVID-19 pandemic. I-HOPE achieves a prediction accuracy of 91%, significantly surpassing the 60-70% accuracy of baseline methods. In addition, I-HOPE distills complex patterns into interpretable and individualized insights, enabling the future development of tailored interventions and improving mental health support.

CCS Concepts

• **Computing methodologies** → **Machine learning approaches**;
• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; • **Applied computing** → **Health informatics**.

Keywords

Mental Health, College Students, Interpretable Machine Learning, Personalized Prediction

ACM Reference Format:

Meghna Roy Chowdhury, Wei Xuan, Shreyas Sen, Yixue Zhao, and Yi Ding. 2025. Predicting and Understanding College Student Mental Health with Interpretable Machine Learning. In *ACM/IEEE International Conference*

on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), June 24–26, 2025, New York, NY, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3721201.3721372>

1 Introduction

Mental health issues among college students have escalated to critical levels, significantly affecting academic performance, social interactions, and overall wellbeing [4, 8, 9, 16]. The American College Health Association reports that 40% of students experience severe depression that disrupts daily functioning, while 60% encounter overwhelming anxiety during the 2020–2021 school year [1]. Furthermore, approximately 76% of college students report moderate to severe psychological distress, with anxiety and depression being the most prevalent diagnoses [6]. Despite increasing awareness, timely access to support remains limited for many students due to stigma, resource constraints, and challenges in detecting those at risk [36]. This situation underscores the pressing need for effective and scalable solutions to improve the understanding and prediction of mental health outcomes.

The complexity of predicting and understanding mental health status among college students arises from three primary factors. First, a comprehensive analysis requires a large-scale, longitudinal dataset that collects data through passive sensing over an extended period rather than relying on short-term data collection conducted in a lab setting. Second, although machine learning has been used to address mental health issues, many existing models utilize black-box algorithms that lack transparency and interpretability [29, 33, 37]. Third, most machine learning approaches yield aggregated insights at the population level that fail to provide individualized understanding, which is essential for personalized interventions and mental health support [5, 10].

In this paper, we address these challenges through a paradigm shift towards methodologies that not only leverage an extensive dataset but also prioritize individual variability in mental health prediction. In particular, we leverage the College Experience Study (CES) dataset [26], the longest longitudinal mobile sensing dataset for college student behaviors, released by Dartmouth College in October 2024. This dataset is especially valuable for our study because it covers a five-year period that includes pre-pandemic years, the COVID-19 pandemic, and the gradual return to normalcy as the pandemic receded. By analyzing behavioral patterns and mental health metrics over different time periods, we can investigate



This work is licensed under a Creative Commons Attribution 4.0 International License. CHASE '25, New York, NY, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1539-6/2025/06

<https://doi.org/10.1145/3721201.3721372>

insights into the pandemic’s impact on mental health and the role of behavior in shaping it.

We present I-HOPE, the first Interpretable Hierarchical mODEl for Personalized mental health prediction. I-HOPE is a two-stage hierarchical model designed to accurately predict mental health status while offering deep insights into the features contributing to various mental health conditions. Specifically, we define five interaction labels: *Leisure*, *Me Time*, *Phone Time*, *Sleep*, and *Social Time*, which categorize different daily behaviors. The key insight of I-HOPE is to connect raw behavioral features to mental health status through an intermediate layer consisting of these five interaction labels. These labels act as compact representations of complex behaviors, simplifying the input space while preserving data richness. Furthermore, I-HOPE facilitates the identification of an individual’s social type and emphasizes the interactions that most significantly impact their mental health.

We compare I-HOPE with baseline methods that lack personalized predictions or two-stage feature mappings. I-HOPE achieves an overall prediction accuracy of 91%, significantly surpassing the 60-70% accuracy of baseline methods. I-HOPE simplifies complex patterns into interpretable and personalized insights by mapping behavioral features into interaction labels. For instance, a good 7-hour sleep is linked to better mental health outcomes, while poor sleep correlates with increased anxiety and depression. Walking serves as a stress reliever associated with leisure and relaxation. Phone usage reveals contrasting patterns; interactions at home often indicate positive connections, whereas excessive use in social settings suggests stress or discomfort. The balance between social and personal time is crucial, with shared spaces that promote engagement and personal spaces that allow emotional recharge. Importantly, these behaviors affect individuals differently; some find “Me Time” in working out, while others prefer studying or spending time in their own dorms. The code is available at <https://github.com/roycmeghna/I-HOPE>.

We summarize the contributions as follows.

- **Interpretable machine learning framework development:** We present I-HOPE, the first hierarchical model that maps raw behavioral data into five interpretable interaction labels: *Leisure*, *Me Time*, *Phone Time*, *Sleep*, and *Social Time*, enhancing the transparency and prediction accuracy.
- **Personalized mental health predictions:** I-HOPE adapts to individual behaviors, enhancing the accuracy and relevance of mental health assessments.
- **Key behavioral predictors identification:** I-HOPE identifies specific behaviors, such as sleep patterns and physical activity levels, significantly influencing mental health outcomes, providing insight into targeted interventions.
- **Scalable analytical approach:** I-HOPE scales high-dimensional behavioral data analysis, linking complex datasets to real-world mental health applications.

2 College Experience Study (CES) Dataset

To thoroughly investigate college students’ mental health status, a large-scale dataset that captures daily behaviors is required. Therefore, we leverage the CES dataset [26], the longest longitudinal mobile sensing dataset for college student behaviors, released by

Table 1: PHQ-4 scores and their categories.

PHQ-4 score	Category
0-3	Normal
4-6	Mild
7-9	Moderate
10-12	Severe

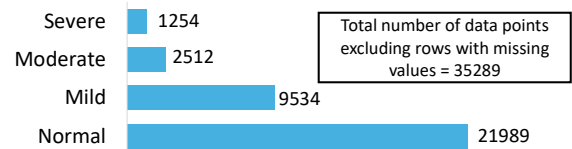


Figure 1: Data distribution across PHQ-4 categories. Normal has the highest count, whereas Severe has the lowest count.

Dartmouth College in October 2024. This dataset includes passive mobile sensing data—mobility, physical activity, sleep patterns, and phone usage—along with Ecological Momentary Assessment (EMA) surveys from 217 Dartmouth students collected between 2017 and 2022. It comprises over 210,000 data points collected across two cohorts throughout their college years on an hourly basis. The EMA surveys are delivered randomly once a week via the StudentLife mobile application [34].

This dataset is especially valuable for our study because it covers a five-year period that includes pre-pandemic years, the COVID-19 pandemic, and the gradual return to normalcy as the pandemic receded. By analyzing behavioral patterns and mental health metrics over different time periods, we can assess and predict how mobile sensing data—collected before, during, and after the COVID-19 pandemic—predicts students’ mental health. This analysis provides valuable insights into the pandemic’s impact on mental health and the role of behavior in shaping it.

PHQ-4 score. The EMA survey includes a key mental health metric, Patient Health Questionnaire-4 (PHQ-4) score [20], which serves as the focal point for our analysis and the mental health outcome we aim to predict in this paper. The PHQ-4 score is a widely recognized screening tool for assessing depressive and anxiety symptoms in clinical settings and epidemiological studies [23]. It ranges from 0 to 12, with lower scores indicating better mental health. To enhance interpretability and align with standardization practices [35], we categorize PHQ-4 scores into four levels: *normal*, *mild*, *moderate*, and *severe*. Each category corresponds to a specific PHQ-4 score range as shown in Table 1, which allows clearer insight into mental health status. The distribution of data points across the PHQ-4 categories is shown in Figure 1.

Features. The original dataset has 172 features, including smartphone usage, duration spent performing various activities, duration spent at various locations, sleep information, etc. However, not all features directly relate to mental health studies, such as the amplitude of detected audio. Additionally, some features, such as distance traveled and number of locations visited, are ambiguous.

Table 2: Pearson correlations between PHQ-4 score with smartphone usage features in five environments.

Category	Metric	Correlation
Overall	Duration of Unlock	0.0321
	# Unlocks	-0.0107
Food	Duration of Unlock	-0.0125
	# Unlocks	-0.0197
Study	Duration of Unlock	0.0014
	# Unlocks	-0.0151
Social	Duration of Unlock	-0.01
	# Unlocks	-0.0157
Dormitory	Duration of Unlock	-0.00215
	# Unlocks	-0.01505
Home	Duration of Unlock	0.0188
	# Unlocks	0.0033

After careful screening, we select 45 features for our study. These features, which include contextual information, phone usage, and sleep duration, are chosen based on the following criteria.

- (1) They provide a meaningful representation of mental health status by capturing behavioral patterns that research has consistently linked to mental health.
- (2) They demonstrated statistical significance with respect to the PHQ-4 categories, with p-values less than 0.05 and, in many cases, less than 0.01.

3 Motivation

In this section, we explore the challenges of predicting mental health status by analyzing the CES dataset and identify key insights for designing effective predictive models.

3.1 The Role of Smartphone Usage on Mental Health

In mental health research, smartphone usage has increasingly been used to facilitate real-time data collection, improve access to mental health resources, and enable personalized interventions. Prior work has shown that screen time and social media use negatively impact mental health and digital wellbeing [14, 40]. Mobile sensing and interventions have also shown promise for personalizing mental health support [39]. Therefore, we start by focusing on two smartphone usage metrics: *the number of times phones are unlocked (#Unlocks)* and *daily phone usage duration (Duration of Unlock)*.

To explore the relationship between these features and mental health, we conduct a correlation analysis using the CES dataset. The results, summarized in Table 2, indicate three main findings: (i) longer phone usage correlates with poorer mental health; (ii) higher unlock frequencies are associated with better mental health; (iii) increased phone use in social and dining settings relates to worse mental health, possibly due to social anxiety, while greater use at home suggests positive activities like socialization, leisure, or learning. These findings highlight the importance of context in understanding the impact of smartphone usage on mental health.

3.2 Smartphone Usage Features Alone Are Insufficient for Predicting Mental Health

To assess whether smartphone usage features can effectively predict mental health outcomes, we train a machine learning model on the CES dataset to predict the PHQ-4 score. Given that the PHQ-4 score has four categories, as illustrated in Table 1, this constitutes a multi-class classification problem. Therefore, we use two smartphone features mentioned above—the number of unlocks and daily phone usage duration—as input features to train a multilayer perceptron (MLP) model, a neural network widely used for classification and prediction [27]. The MLP model has 2 input nodes, 3 fully connected hidden layers, and 4 output nodes corresponding to the PHQ-4 categories. The model is trained for 50 epochs until the learning curve stabilizes. We preprocess the data by normalizing the features and ensuring balanced class distributions through oversampling for the PHQ-4 categories. We randomly split the dataset into 80% for training and 20% for testing. Our results reveal that the prediction accuracy is only 28%, which is low. These results indicate that relying solely on these smartphone usage features is insufficient to predict mental health outcomes accurately.

3.3 Improving Predictions with Additional Features

To improve prediction accuracy, we need to incorporate additional features from the CES dataset. Specifically, we use 45 manually selected features from the CES dataset related to location information, sleep duration, and activity times (see Column 2 of Table 3). We use the MLP model with 45 input nodes, 3 fully connected hidden layers, and 4 output nodes corresponding to the PHQ-4 categories. The training and test sets are the same as above. The model is trained for 50 epochs until the learning curve stabilizes. This approach improves the prediction accuracy to 60%, which is still not high. We attribute the low accuracy to two reasons: (i) treating the dataset as a whole may have obscured individual differences; (ii) a high correlation among features likely reduce the model’s ability to make accurate predictions. We discuss these below.

Addressing Individual Differences. The CES dataset includes data from 217 individuals, each contributing about 160 data points on average. Research indicates that individual differences in behavior and preferences can significantly affect predictive modeling outcomes [17, 22]. To capture these variations, we develop personalized MLP model for each student. We focus on students who contribute to at least 160 data points in the dataset, which gives us 121 students. Each model is trained on the corresponding individual’s data with 45 selected features, maintaining the same MLP architecture and train-test split as described above. This personalized approach improves the average prediction accuracy to 70%. This motivates the need for additional feature engineering to improve feature representation and prediction accuracy.

Reducing Feature Correlation. Highly correlated features introduce redundancy, leading to overfitting and reduced test accuracy while increasing computational complexity. As shown in Figure 2, many features are highly correlated. To address this, we analyze feature importance using random forest, an efficient and interpretable

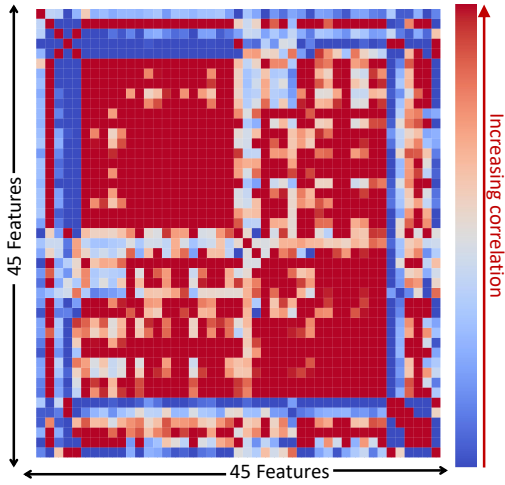


Figure 2: Heatmap for correlations among features.

ensemble learning technique [7]. This method builds multiple decision trees and combines their outputs to evaluate feature importance. Figure 3 shows the importance ranking of all 45 features, highlighting the top ones. We select the top 50% of these features to retrain personalized MLP models, adjusting the input dimensions accordingly. However, this results in slight a drop in prediction accuracy by ~5%, indicating that simply excluding certain features may remove critical information necessary for accurate predictions.

Key takeaway. These analyses suggest that the features influencing prediction accuracy vary among individuals. For instance, a higher biking time may correlate with better mental health for one person, while walking may be more relevant for another, and a combination of running and walking could be key for someone else. A global feature importance approach risks eliminating crucial features for individuals, potentially reducing accuracy. Therefore, it is vital to incorporate personalization while minimizing feature correlation, ensuring the model captures individual-specific patterns without being hindered by redundant or irrelevant features, ultimately enhancing overall prediction accuracy.

4 I-HOPE

Our analysis has shown that (1) personalized models outperform a global model shared among students; and (2) effective feature engineering efforts are needed to reduce feature correlation and improve prediction accuracy. Therefore, we design I-HOPE, the first **Interpretable Hierarchical model for Personalized mental health prediction**. I-HOPE is a two-stage hierarchical model that connects behavioral features to mental health status through an intermediate layer of five interaction labels: *Leisure*, *Me Time*, *Phone Time*, *Sleep*, and *Social Time*, which categorize different daily behaviors. This model structure offers two benefits:

- (1) **Dimension Reduction.** The interaction labels serve as compact representations of complex behaviors, simplifying the input space while retaining the richness of the data. The effective feature representation contributes to high prediction accuracy.

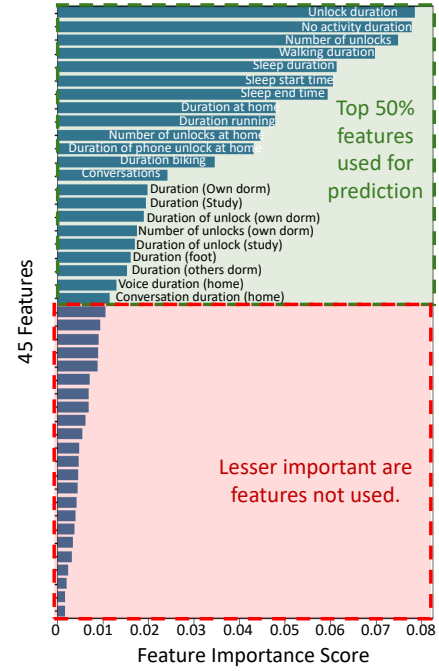


Figure 3: Feature importances using random forests.

- (2) **Enhanced Interpretability.** I-HOPE enables the identification of an individual’s social type and highlights the interactions that correlate their mental health the most.

Figure 4 presents the workflow of I-HOPE design. It starts with n input features, which are mapped to k interaction labels ($k = 5$), representing aggregated behavioral categories: *Leisure*, *Me Time*, *Phone Time*, *Sleep*, and *Social Time*. To quantify the behavioral insights, we assign a score for each interaction label by introducing a two-step rule-based method. These scores then serve as input to a neural network model that predicts one of four mental health categories: *Normal*, *Mild*, *Moderate*, or *Severe*. In summary, by leveraging the hierarchical model structure, I-HOPE improves both prediction accuracy and model interpretability. For example, I-HOPE can assess whether “walking” or “spending time with friends” in the *Leisure* category has a greater impact on mental health of an individual. I-HOPE is conducted per-user basis by training and testing on individualized data. Before introducing each stage in detail, we will first describe the feature engineering efforts we have made to prepare the final input dataset for I-HOPE.

4.1 Feature Engineering

In §3, we have demonstrated that high quality input features are critical to high prediction accuracy. Based on our previous analysis, we further conduct the following feature engineering procedure to prepare the input dataset for I-HOPE.

In our case, feature engineering is guided by domain relevance, ensuring that derived features carry meaningful interpretations in the context of mental health. For example, rather than using the number of phone unlocks and duration of phone usage as separate features, we introduce a ratio feature: $\text{number_of_unlocks} /$

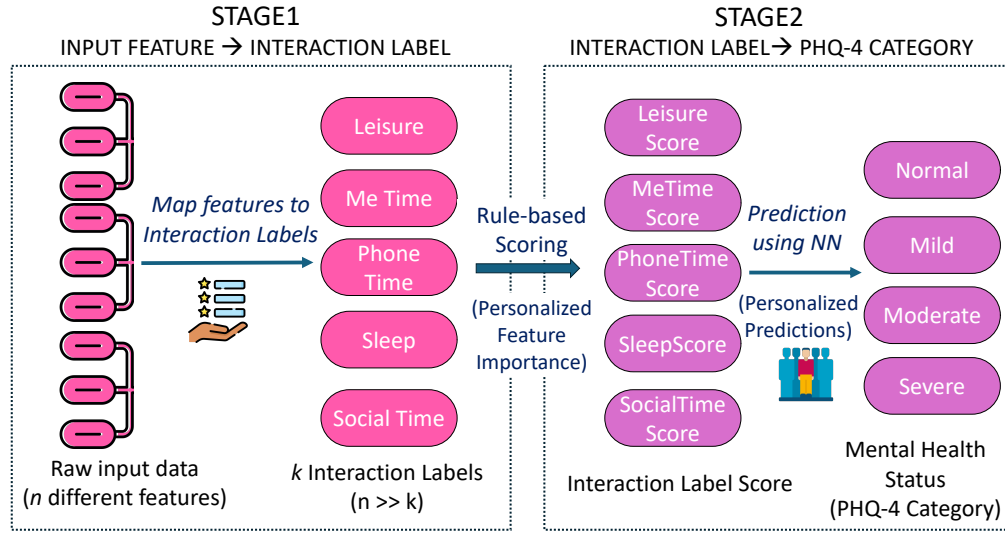


Figure 4: Workflow of the I-HOPE design.

Table 3: Interaction labels and their corresponding features.

Interaction label	Relevant features corresponding to each interaction label
Leisure	- Duration of physical activities (biking, walking, running) - Duration of conversations (in-person and phone) - Duration in locations like others' dorm, workout - Duration of phone usage in various locations
Me Time	- Duration of activities (biking, walking, running, being still, studying) - Various measures at home (conversation detected, duration) - Conversations detected at own dorm or home - Duration of phone usage in various locations
Phone Time	- Duration of phone conversations in various locations - Ratio of number of calls to duration of calls - Ratio of number of phone unlocks to duration of unlocks, in various locations
Sleep	- Duration of stillness - Audio detection at home or own dorm - Phone usage at night - Sleep duration
Social Time	- Duration of social activities (biking, walking, running, workout, study, eating food) - Ratio of number of calls to duration of calls at locations like study space, home, etc. - Amount of time spent in others' dorm

duration_of_usage. A high value of this feature suggests more frequent, brief interactions with the phone, and potentially indicative of restlessness. We apply similar transformation across multiple contextual locations, including home, own dorm, study spaces, others' dorms, social environments, and overall daily usage. In addition, instead of separately including the number and duration of incoming and outgoing calls, we combine them into a single feature:

$$\frac{\text{Number of incoming calls} + \text{outgoing calls}}{\text{Call duration of incoming} + \text{outgoing calls}}$$

This consolidation maintains the behavioral signal while reducing redundancy. Overall, we reduce the 45 features to 35 final features,

preserving domain-specific behavioral meaning while improving computational efficiency.

4.2 Stage 1: Input Features → Interaction Labels

In Stage 1, I-HOPE maps input 35 features to five interaction labels: *Leisure*, *Me Time*, *Phone Time*, *Sleep*, and *Social Time*. This approach combines domain knowledge and data-driven insights to capture relevant behavioral dimensions, reduce redundancy, and enhance interpretability while maintaining granularity for personalized analysis. The detailed process is described below.

(1) **Initial mapping:** We manually map each feature to different each interaction label based on their semantic and behavioral

Algorithm 1 Rule-based initialization of *SleepScore*

```

1: Input: Thresholds of each feature in Sleep label, which are
   population means of the feature distributions.
2: Initialize: SleepScore = 0
3: if Sleep duration > Threshold (where Threshold = Mean of
   Sleep duration distribution) then
4:   SleepScore += 1
5: end if
6: if Duration at own dorm > Threshold (where Threshold =
   Mean of Duration at own dorm distribution) then
7:   SleepScore += 1
8: end if
9: if Duration of stillness > Threshold (where Threshold =
   Mean of Duration of stillness distribution) then
10:  SleepScore += 1
11: end if
12: if Conversation detected at home < Threshold (where
   Threshold = Mean of Conversation detected at home dis-
   tribution) then
13:  SleepScore += 1
14: end if
15: Output: SleepScore

```

relevance. For instance, features like *unlock duration* and *phone conversations* belong to *Phone Time*, while *duration of social activities* and *time spent with friends* belong to *Social Time*. The initial mapping is detailed in Table 3. Notably, some features are assigned to multiple labels, reflecting their multidimensional impact on various behavioral categories. We further employ K-means clustering with $k=5$ to perform a soft validation, assessing whether five interaction labels are a reasonable choice.

- (2) **Rule-based scoring initialization:** Since interaction labels are manually created, we need to assign them numerical values for downstream quantitative analysis. I-HOPE uses a two-step rule-based scoring method to compute an *[InteractionLabel]Score* for each label. The first step is to initialize the scores. Specifically, for each feature within a label, we compare its value to the population mean value of that feature; if it exceeds the mean, the score increases by 1. The intuition is that higher feature values suggest stronger relevance to the label and thus contribute more to its score. We illustrate the first step of this method using the *Sleep* label in Algorithm 1, although the same procedure applies to all interaction labels.
- (3) **Rule-based scoring using feature importance:** The second step of the rule-based scoring method is to use a random forest model on the features and initialized scores to assess the contribution of individual input features to each interaction label. We choose random forest due to its ability to handle high-dimensional data and capture nonlinear relationships [7]. I-HOPE then uses feature importance generated from random forest to refine the final scores. For each feature within a label, we compare its value to the population mean; if its value exceeds the mean, we increase the score by its Normalized Weighted Feature Impact (NWFI), which is a weighted measure combining the feature’s random forest-derived importance and

Algorithm 2 Rule-based final scoring of *SleepScore*

```

1: Input: Thresholds of each feature in Sleep label, which are
   population means of the feature distributions.
2: Initialize: SleepScore = 0
3: if Sleep duration > Threshold (where Threshold = Mean of
   Sleep duration distribution) then
4:   SleepScore += NWFI of Sleep duration
5: end if
6: if Duration at own dorm > Threshold (where Threshold =
   Mean of Duration at own dorm distribution) then
7:   SleepScore += NWFI of Duration at own dorm
8: end if
9: if Duration of stillness > Threshold (where Threshold =
   Mean of Duration of stillness distribution) then
10:  SleepScore += NWFI of Duration of stillness
11: end if
12: if Conversation detected at home < Threshold (where
   Threshold = Mean of Conversation detected at home dis-
   tribution) then
13:  SleepScore += NWFI of Conversation detected at home
14: end if
15: Output: SleepScore

```

its normalized value. The normalization reflects the scaling of feature values, and the weighted indicates the incorporation of random forest feature importance. We illustrate the second step of this method using the *Sleep* label in Algorithm 2, although the same procedure applies to all interaction labels.

4.3 Stage 2: Interaction Labels → PHQ-4 Category

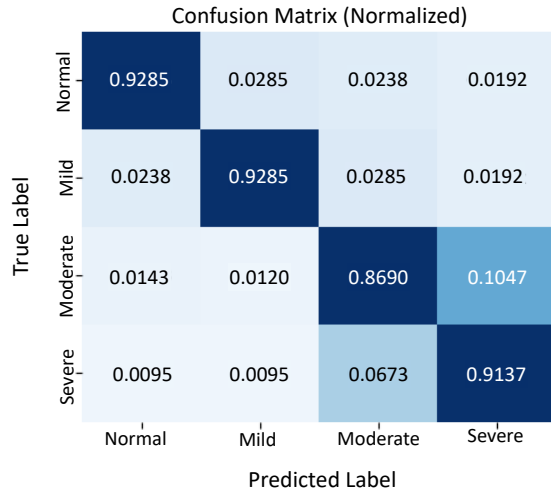
In Stage 2, I-HOPE uses the five interaction label scores (*LeisureScore*, *MeScore*, *PhoneScore*, *SleepScore*, *SocialScore*) as inputs to predict PHQ-4 categories. This stage simplifies the predictive process by using these labels as compact and interpretable representations of the raw data. Specifically, I-HOPE builds a neural network model to predict PHQ-4 categories using the five interaction label scores as inputs. The neural network includes an input layer with five nodes, three hidden layers for pattern extraction, and an output layer with four nodes for the PHQ-4 categories (*Normal*, *Mild*, *Moderate*, or *Severe*). I-HOPE uses the ReLU activation function for hidden layers and the Softmax activation function for the output layer for multi-class classification. The neural network model is trained for 50 epochs using the Adam optimizer [19] with a learning rate of 0.001 and categorical cross-entropy as the loss function.

4.4 Implementation

We implement I-HOPE in Python. We use Pandas and NumPy for data preprocessing (cleaning, transformation, and feature extraction) and feature engineering. We implement neural network models using TensorFlow. Evaluation metrics such as accuracy are computed with Scikit-learn.

Table 4: Comparison of different methods for predicting PHQ-4 categories. Evaluation metrics include precision (Pre.), recall (Rec.), F1-score (F1), and overall accuracy for each PHQ-4 category.

PHQ-4 Category	Baseline 1			Baseline 2			Baseline 3			I-HOPE		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Normal	0.6453	0.627	0.6360	0.7149	0.677	0.6954	0.6781	0.654	0.6659	0.9513	0.9285	0.9398
Mild	0.6135	0.617	0.6152	0.6432	0.667	0.6549	0.6297	0.667	0.6478	0.9489	0.9285	0.9386
Moderate	0.5908	0.600	0.5953	0.6371	0.660	0.6483	0.6083	0.660	0.6332	0.8791	0.8690	0.8740
Severe	0.5831	0.559	0.5708	0.6265	0.614	0.6202	0.5982	0.614	0.6060	0.8645	0.9138	0.8884
Overall Accuracy	0.60			0.70			0.65			0.91		

**Figure 5: Normalized confusion matrix from I-HOPE.**

5 Evaluation

In this section, we evaluate the I-HOPE to address the following research questions (RQs):

RQ1. Prediction accuracy: How accurately does I-HOPE predict the mental health status of college students?

RQ2. Interpretability from hierarchical mapping: What new insights into mental health can we gain from hierarchical feature mapping of I-HOPE?

RQ3. Scoring for interaction labels: How much do features contribute to the scoring of interaction labels?

The rest of this section is divided into three parts, each addressing a research question.

5.1 Prediction Accuracy

We compare I-HOPE against three baseline methods that differ in the training approach (aggregated vs. personalized), feature extraction (all raw features vs. selected features) and model structure (single stage vs. hierarchical). We evaluate 121 students who contribute at least 160 data points in the dataset. We report results from the best-performing model from 5-fold cross validation for all baselines. The methods are summarized as follows, all using the

previously described neural network architecture with adjusted input dimensions:

- **Baseline 1:** An aggregated single model trained with selected 45 raw features across 121 students.
- **Baseline 2:** Personalized models for each individual using selected 45 raw features.
- **Baseline 3:** Personalized models for each individual using the top 50% of 45 features based on feature importance from random forest models.
- **I-HOPE:** The Interpretable Hierarchical model for Personalized mental health prediction presented in this paper.

The evaluation results are summarized in Table 4. Baseline 1 achieves 60% accuracy due to feature redundancy and lack of personalization. Baseline 2 improves to 70% but still faces feature correlation and redundancy issues. Baseline 3 reduces feature correlation using feature importance but drops to 65% accuracy by excluding significant features. In contrast, I-HOPE reaches 91% overall prediction accuracy by using interaction labels to capture distinct user behaviors while balancing interpretability and predictive power. We can see that the baseline methods struggle with feature redundancy and adapting to user-specific patterns, while I-HOPE effectively integrates personalized interaction labels and feature importance. We also breakdown I-HOPE's predictions against the true labels for each category using a confusion matrix. As shown in Figure 5, I-HOPE achieves high accuracy across all four categories. Overall, the evaluation using multiple metrics highlights the strong predictive performance of I-HOPE.

5.2 Interpretability

To gain insights into mental health and improve prediction accuracy, it is crucial to extract meaningful and interpretable features from raw behavioral data. I-HOPE simplifies and organizes raw data into actionable constructs, enabling the capture of individual behavioral patterns and their connection to mental health outcomes.

5.2.1 Evaluation of Stage 1: Input Features → Interaction Labels.

The first stage of I-HOPE maps the raw input features into five interaction labels: *Leisure*, *Me Time*, *Phone Time*, *Sleep*, and *Social Time*. This simplifies the dataset while preserving key behavioral patterns relevant to mental health. To enhance interpretability, we conduct feature importance analysis for all selected features within each label. This analysis identifies which features most influence the model's predictions, revealing meaningful behavioral patterns related to mental health. We use a random forest model to compute

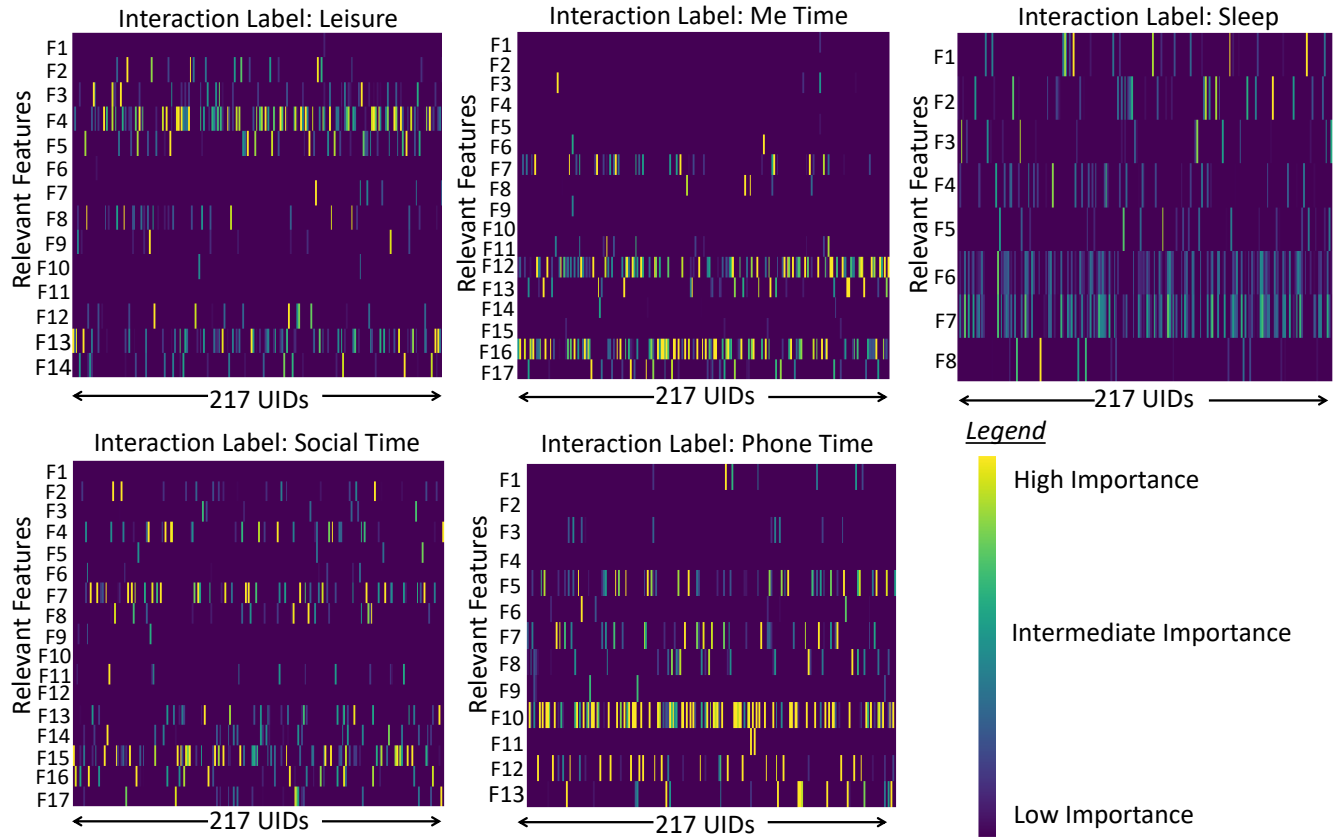


Figure 6: The individual-level heatmaps that highlight feature importance for each interaction label across 217 individuals (UIDs). The y-axis represents label-specific features, which vary across different interaction labels.

feature importance scores, providing an interpretable ranking of features based on their contributions. This analysis is performed individually for personalized insights into user behavior.

Figure 6 shows personalized feature importance across five labels using heatmaps for 217 students. The analysis reveals distinct patterns of feature relevance for each label, highlighting both shared and unique predictors. Below, we summarize the key findings and insights for each label.

- Label: Leisure** - Feature F4 (Duration of walking) emerged as the most important predictor for 90.36% of individuals, emphasizing walking as a key leisure activity. Other features, such as F13 (Phone usage at home), F3 (Duration of running), F5 (Duration of conversations), F8 (Voice detection at others' dorms), and F14 (Phone usage at others' dorms), are moderately important for 20–25% of individuals, indicating that leisure also encompasses social interactions and phone use. In contrast, features like F1 (Biking duration), F2 (Footsteps), F6 (Duration of calls), F9 (Conversations at others' dorms), F10 (Audio detections at social places), F11 (Duration at "leisure"), and F12 (Workout duration) consistently show lower relevance. F11 has low importance despite its direct link to the label, likely due to unclear dataset definitions, underscoring the need for improved data clarity. *These*

findings illustrate how behaviors like walking and social interactions influence leisure and provide insights into broader themes related to mental health and lifestyle habits.

- Label: Me Time** - Features such as F16 (Phone usage at own dorm), F12 (Time spent at own dorm), F7 (Voice detections at own dorm), F13 (Study duration), and F17 (Phone usage at study locations) are important for 89.5% of individuals, with F12 and F16 being the most significant. These findings suggest that activities in dorm spaces, such as studying and socializing, are key indicators of "Me Time" and contribute to a supportive environment. Conversely, features like F1 (Biking duration), F2 (Footsteps), F3 (Running duration), F5 (Walking duration), and F14 (Workout duration) show low importance, indicating that individual physical activities are less relevant to this label. Other features display varying levels of significance across individuals, reflecting the diverse experiences of "Me Time". *These results underscore the role of personal environments like dorms in fostering meaningful routines and behaviors, offering insights into how individuals balance solitude, productivity, and emotional wellbeing.*
- Label: Sleep** - Features F6 (Sleep duration) and F7 (Duration of being idle) are key predictors, highlighting the importance of inactivity and sleep in identifying sleep patterns. Features like F1 (Conversation detection at dorm) and F4 (Location at dorm) show

moderate importance, suggesting that reduced social interactions and staying in the dorm contribute to recognizing sleep behaviors, likely linked to quiet nighttime routines. In contrast, features such as F2 (Voice detections at home), F3 (Conversations at home), F5 (Phone usage at home), and F8 (Health fitness) are minimally relevant. *These results emphasize the role of inactivity and dorm contexts in understanding sleep patterns while activities like fitness or phone use have a limited impact on sleep.*

- **Label: Social Time** - Feature F15 (Duration at study location) is the key predictor, indicating a strong link between study locations and social activities. Other features include F2 (Walking duration), F7 (Voice detections at study areas), F13 (Time spent in others' dorms), and F16 (Phone usage in others' dorms), which highlight the importance of mobility, conversations, and phone use. In contrast, features like F1 (Footsteps), F9 (Conversations in others' dorms), F10 (Conversations at social places), and F12 (Duration of leisure) show limited importance, suggesting they are less directly related to social time or relevant only for a few individuals. *These findings emphasize the role of specific environments and interaction patterns in shaping social time while indicating that factors like leisure duration or general mobility have a minor impact.*
- **Label: Phone Time** - Feature F10 (Phone usage at own dorm) is the primary predictor, indicating that students primarily use their phones in personal spaces where they feel comfortable engaging in calls and messaging. Other features include F5 (Audio detected at own dorm), F7 (Audio detected at home), F8 (Phone usage at home), F12 (Phone usage at study locations), and F13 (Phone usage throughout the day), which highlight variations in phone use across personal, home, and study contexts. Conversely, features like F1 (Audio conversations detected throughout the day), F2 (Number of SMS), F3 (Voice detection at home), F4 (Voice detection in social settings), F6 (Conversations detected at home), F9 (Phone usage in others' dorms), and F10 (Phone usage in workout areas) show minimal relevance, likely due to ambiguous definitions or differences in data collection between iOS and Android devices. *These findings emphasize the role of personal and study spaces in shaping phone usage while highlighting the need for clearer feature definitions and standardized data collection to enhance the reliability of less relevant features.*

In summary, the findings reveal personalized variability among students in feature importance, with key behavioral patterns linked to mental health and activity contexts:

- **Behavioral patterns:** Walking, phone usage, and time in personal spaces are key to labels like *Leisure*, *Phone Time*, and *Me Time*. Walking supports relaxation, while dorm-based phone use reflects connection and self-reflection, linking physical spaces to emotional wellbeing.
- **Activity contexts:** Study locations and others' dorms strongly predict *Social Time*, emphasizing collaborative settings. In contrast, private spaces dominate *Me Time* and *Phone Time*, illustrating the balance between shared and personal spaces for social engagement and self-care.
- **Sleep and rest:** Sleep duration and idle time are central to *Sleep*, with dorm-related contexts suggesting restful routines tied to spatial habits and reduced interactions, reinforcing sleep's role in mental health.

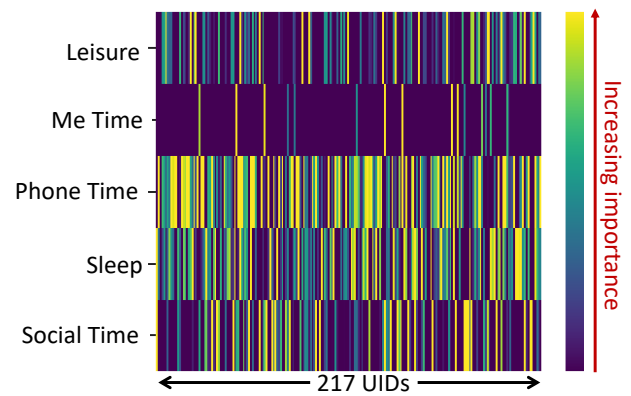


Figure 7: The individual-level heatmap that shows the importance of interaction label scores in predicting PHQ-4 categories. Each row represents an interaction label across 217 individuals.

- **Personalized variability as a key factor:** The varying importance of features like walking or conversation detection reflects the individual nature of behavior. Tailored analyses can enhance mental health predictions, making them more accurate and context-aware.

These results emphasize the importance of selecting features that capture distinct behavioral dimensions for interaction labels. They also demonstrate that personalized analysis is key for accurate mental health prediction, as feature importance varies across users. This forms the basis for our label scoring and predictions, explained in the following subsections.

5.2.2 Evaluation of Stage 2: Interaction Labels → PHQ-4 Categories. Figure 7 shows the importance of interaction labels (*Leisure*, *Me Time*, *Phone Time*, *Sleep*, *Social Time*) in predicting PHQ-4 mental health categories across 217 users. Each row corresponds to a label, and each column represents an individual user, with colors ranging from dark purple (low importance) to bright yellow (high importance). The figure highlights how the relevance of these labels varies significantly across individuals, revealing the personalized nature of mental health predictors.

Among the labels, *Sleep* is the most consistently important, with high relevance for 95% of students, reflecting the strong link between healthy sleep patterns and stable mental health. *Phone Time* also shows significant importance across many users, likely indicating stress or coping behaviors related to screen use. In contrast, labels like *Social Time* and *Leisure* exhibit more variability; *Social Time* is crucial for some but less so for others, reflecting individual differences in social interactions' impact on mental health. Similarly, *Leisure* is an important predictor for a subset of users, emphasizing the role of relaxation activities. However, *Me Time* serves as a predictor for an even smaller group, highlighting the challenges of finding solitude in a school environment with shared spaces.

In all, Figure 7 underscores a critical insight – mental health predictors are not one-size-fits-all. Each label contributes uniquely, reflecting the diverse and personalized ways behavioral patterns

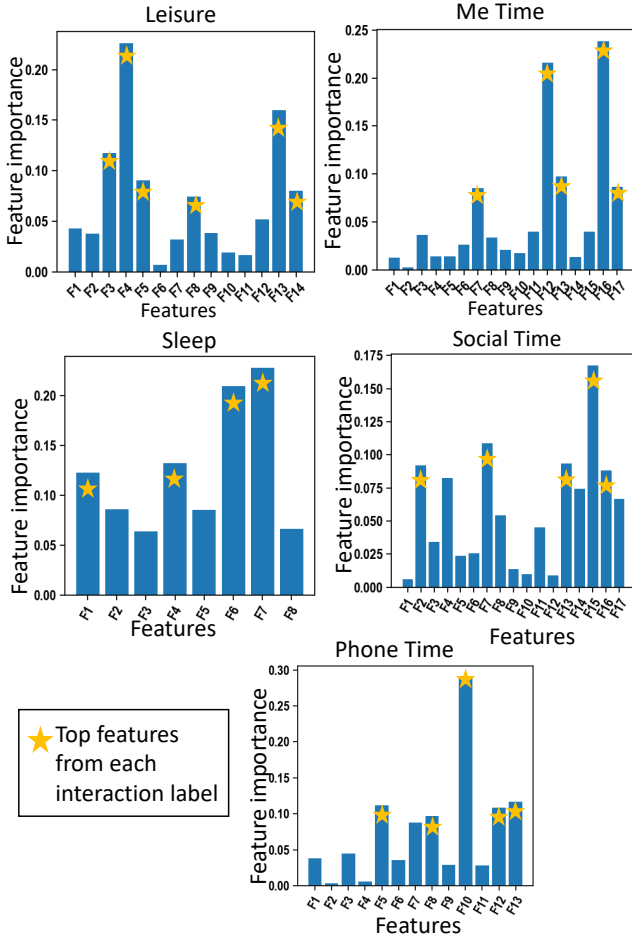


Figure 8: Mean feature importance across all individuals for each interaction label.

influence mental health outcomes. While *Sleep* and *Phone Time* act as general predictors across most users, *Social Time*, *Me Time*, and *Leisure* provide insights tailored to specific individuals. By combining these labels, we can build a comprehensive understanding of behavioral patterns and their impact on mental health, supporting the need for personalized modeling approaches to predict PHQ-4 categories effectively.

5.3 Scoring for Interaction Labels

To analyze how each feature contributes to scoring for each interaction label, we visualize the feature importances of relevant features derived from random forest models across all individuals. Figure 8 displays the top features (marked with a star) for each interaction label:

- **Leisure:** F3 (running duration), F4 (walking duration), F5 (conversation duration), F8 (voice detection at others' dorms), F13 (phone usage at home), and F14 (phone usage at others' dorms).
- **Me Time:** F7 (voice detections at own dorm), F12 (time spent in own dorm), F13 (study duration), F16 (phone usage at own dorm), and F17 (phone usage at study locations).

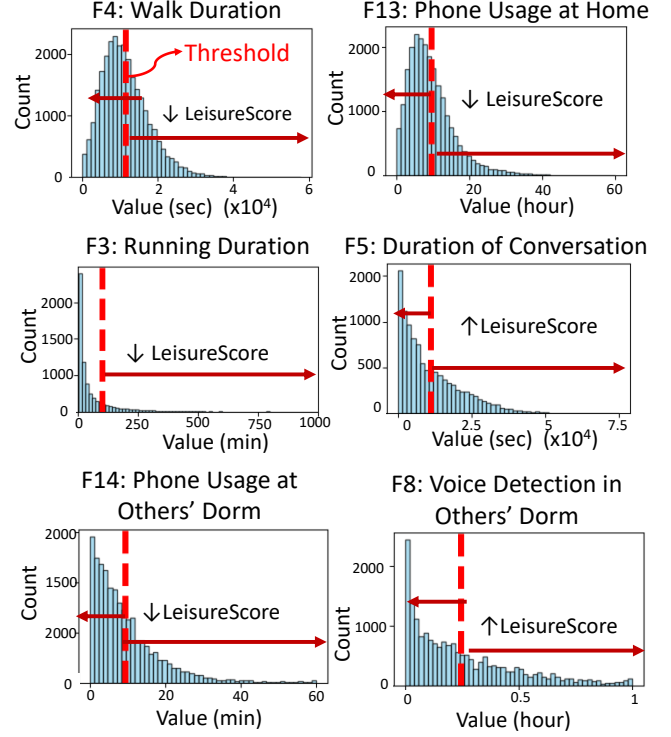


Figure 9: Thresholds for LeisureScore calculation based on six features. The vertical dashed red line is the threshold.

- **Sleep:** F1 (conversations detected at dorm), F4 (duration at own dorm), F6 (sleep duration), and F7 (duration of being idle).
- **Social Time:** F2 (walking duration), F7 (voice detected at study location), F13 (time in others' dorms), F15 (duration at study location), and F16 (phone usage at others' dorm).
- **Phone Time:** F5 (audio detected at own dorm), F8 (phone usage at home), F10 (phone usage at own dorm), F12 (phone usage at study locations), and F13 (phone usage throughout the day). Phone usage corresponds to the ratio of the number of phone unlocks to usage duration.

As illustrated in Algorithm 2, the scores for each interaction label are calculated based on the weighted importance of various features (i.e., NWFI in Algorithm 2), measured against a threshold value derived from the mean of distribution across all individuals. Next, we show the thresholds for *Leisure Time* and *Sleep* to demonstrate how feature contributions influence the final label scoring.

In Figure 9, the dotted lines highlight the threshold values for the most important features (F4, F13, F3, F5, F14, and F8) for *Leisure Time*. F4 reflects walking duration: if a student's walking duration is less than the threshold (3 hours), it indicates leisure activity. Consequently, the *LeisureScore* is incremented proportionally to the weighted feature importance of walking for the respective student. Similarly, F8 considers voice detection in others' dormitories. If the duration exceeds the threshold (15 minutes), it suggests that the student is having leisure time with friends, causing an increase in *LeisureScore* by the corresponding weighted feature importance.

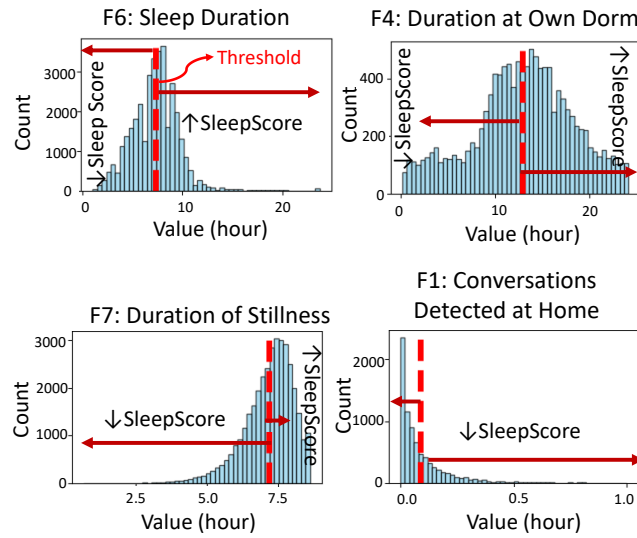


Figure 10: Thresholds for SleepScore calculation based on four features. The vertical dashed red line is the threshold.

In Figure 10, the dotted lines highlight the threshold values for the most important features (F6, F4, F7, and F1) for *Sleep*. F6 indicates total sleep duration: if a student’s sleep exceeds 7 hours, representing the standard benchmark for healthy sleep, the *SleepScore* is incremented based on the student’s weighted feature importance for sleep. F4, which measures the time spent in the room, shows that durations greater than 12.9 hours typically signify overnight dorm occupancy, often indicative of sleep rather than other activities. F7 corroborates healthy sleep with thresholds above 7.5 hours, while F1 reflects quietness, with values below 6 minutes suggesting an environment conducive to rest.

6 Related Work

Machine learning increasingly tackles various challenges in mental health research [2, 12, 18, 29, 30, 33]. Using data from mobile devices that monitor smartphone use, social media activity, and medical records, researchers have developed machine learning techniques to identify and predict conditions such as depression, anxiety, and stress [15, 28, 40]. Furthermore, natural language processing has opened new pathways for analyzing text data, including social media posts and therapy transcripts, offering valuable insights into emotional and psychological states [3, 11, 13, 21, 31].

However, despite these advances, many machine learning techniques struggle to capture and understand mental health characteristics at the individual level. Researchers often focus on applying techniques without addressing the issues of variability, personalization, and interpretability [5, 24, 25, 32]. This limitation restricts the practical application of these methods, particularly in treatment recommendations and interventions.

The release of the College Experience Study dataset offers a unique opportunity to analyze the longest longitudinal passive mobile sensing dataset on the mental health of college students at the

individual level [26]. Our prior work was the first attempt to analyze this dataset [38]. This paper goes a step further by addressing a key research gap with I-HOPE, the first interpretable hierarchical model designed to accurately predict and offer new insights into the mental health of college students.

7 Conclusion and Future Work

This paper addresses the challenge of personalized predictions and understanding mental health among college students. Leveraging a cutting-edge longitudinal dataset, we present I-HOPE, the first hierarchical model designed to predict and interpret mental health status at the individual level. I-HOPE bridges the gap between black-box machine learning and actionable insights, offering both accurate and interpretable, personalized predictions. These contributions are especially important in the mental health domain, where individualization is critical.

Our evaluation shows that I-HOPE has a better performance than baseline methods and offers insights into how behavioral patterns, such as sleep, phone usage, and social time, relate to mental health outcomes on a personalized level. However, like any data-driven model, there are several areas that offer opportunities for further improvement, which we describe as follows.

From a data perspective, our analysis relied on features aggregated at a daily level. Incorporating higher-resolution hourly data could reveal more granular behavioral patterns, potentially enhancing prediction accuracy. Additionally, certain data processing decisions—such as replacing missing values with zeros—may oversimplify real-world behaviors. Future work could explore more sophisticated imputation techniques to better capture underlying trends.

From a modeling perspective, while the current framework uses five interaction labels—empirically motivated and preliminarily validated—future work could systematically explore how varying the number of labels impacts both predictive performance and interpretability. Additionally, incorporating aggregate metrics (e.g., mean performance) and confidence intervals during evaluation would offer a more robust assessment of model reliability.

Furthermore, the lack of a fixed random seed introduced unintended variability in reported results; standardizing this in future experiments will improve reproducibility and consistency.

These refinements present a pathway for advancing mental health research, enabling the development of more interpretable, personalized, and actionable models.

Acknowledgments

We thank anonymous reviewers for their helpful feedback to improve the final version of the paper. The work of Meghna Roy Chowdhury and Shreyas Sen is supported by the National Science Foundation under Grant # 1944602.

References

- [1] Zara Abrams. 2022. Student mental health is in crisis. Campuses are rethinking their approach. *Monitor on Psychology* 53, 7 (2022), 60.
- [2] Daniel A Adler, Fei Wang, David C Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *Plos one* 17, 4 (2022), e0266516.
- [3] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to

- mental health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476.
- [4] Georgia Barbayannis, Mahindra Bandari, Xiang Zheng, Humberto Baquerizo, Keith W Pecor, and Xue Ming. 2022. Academic stress and mental well-being in college students: correlations, affected groups, and COVID-19. *Frontiers in psychology* 13 (2022), 886344.
 - [5] Laura Batstra and Sami Timimi. 2024. Is more psychotherapy a dead horse? An essay on the (in) effectiveness of individual treatment for mental suffering. *PLOS Mental Health* 1, 7 (2024), e0000194.
 - [6] BestColleges. 2024. College Student Mental Health Statistics. <https://www.bestcolleges.com/research/college-student-mental-health-statistics/>. Accessed: 2024-03-22.
 - [7] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
 - [8] Tianshu Chu, Xin Liu, Shigemi Takayanagi, Tomoko Matsushita, and Hiro Kishimoto. 2023. Association between mental health and academic performance among university undergraduates: The interacting role of lifestyle behaviors. *International Journal of Methods in Psychiatric Research* 32, 1 (2023), e1938.
 - [9] Jessica Colarossi. 2022. Mental health of college students is getting worse. *The Brink* (2022).
 - [10] Mark Conley. 2024. Finding personalized approaches to treating mental illness. <https://stanmed.stanford.edu/precision-mental-health-promise/> [Online; accessed 2025-01-04].
 - [11] Benjamin L Cook, Ana M Progovac, Pei Chen, Brian Mullin, Sherry Hou, and Enrique Baca-Garcia. 2016. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Computational and mathematical methods in medicine* 2016, 1 (2016), 8708434.
 - [12] Simon D'Alfonso. 2020. AI in mental health. *Current opinion in psychology* 36 (2020), 112–117.
 - [13] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preotiu-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* 115, 44 (2018), 11203–11208.
 - [14] Nadia O Elamin, Juman M Almasaad, Reem B Busaeed, Daniah A Aljafari, and Muhammed A Khan. 2024. Smartphone addiction, stress, and depression among university students. *Clinical Epidemiology and Global Health* 25 (2024), 101487.
 - [15] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.
 - [16] Dalal Hammoudi Halat, Abderrezzaq Soltani, Roua Dalli, Lama Alsarraj, and Ahmed Malki. 2023. Understanding and fostering mental health and well-being among university faculty: A narrative review. *Journal of clinical medicine* 12, 13 (2023), 4425.
 - [17] Zengyou He, Yifan Tang, Lianyu Hu, Mudi Jiang, and Yan Liu. 2023. Personalized Interpretable Classification. *arXiv preprint arXiv:2302.02528* (2023).
 - [18] Jina Kim, Daewun Lee, Eunil Park, et al. 2021. Machine learning for mental health in social media: bibliometric study. *Journal of Medical Internet Research* 23, 3 (2021), e24870.
 - [19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
 - [20] Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. 2009. An ultra-brief screening scale for anxiety and depression: the PHQ-4. *Psychosomatics* 50, 6 (2009), 613–621.
 - [21] Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727* (2023).
 - [22] Ben Lengerich, Bryon Aragam, and Eric P Xing. 2019. Learning sample-specific models with low-rank personalized regression. *Advances in Neural Information Processing Systems* 32 (2019).
 - [23] Bernd Löwe, Inka Wahl, Matthias Rose, Carsten Spitzer, Heide Glaesmer, Katja Wingenfeld, Antonius Schneider, and Elmar Brähler. 2010. A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of affective disorders* 122, 1-2 (2010), 86–95.
 - [24] Danilo Moggia, Wolfgang Lutz, Eva-Lotta Brakemeier, and Leonard Bickman. 2024. Treatment Personalization and Precision Mental Health Care: Where are we and where do we want to go? *Administration and Policy in Mental Health and Mental Health Services Research* 51, 5 (2024), 611–616.
 - [25] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* (2017).
 - [26] Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Courtney Rogers, Meghan L Meyer, and Andrew T Campbell. 2024. Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–37.
 - [27] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastrokakis. 2009. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems* 8, 7 (2009), 579–588.
 - [28] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, David C Mohr, et al. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e4273.
 - [29] Elena Sajno, Sabrina Bartolotta, Cosimo Tuena, Pietro Cipresso, Elisa Pedrol, and Giuseppe Riva. 2023. Machine learning in biosignals processing for mental health: A narrative review. *Frontiers in Psychology* 13 (2023), 1066317.
 - [30] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine* 49, 9 (2019), 1426–1448.
 - [31] Robert Stewart and Sumithra Velupillai. 2021. Applied natural language processing in mental health big data. *Neuropsychopharmacology* 46, 1 (2021), 252.
 - [32] John Torous, Mathew V Kiang, Jeanette Lorme, Jukka-Pekka Onnela, et al. 2016. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR mental health* 3, 2 (2016), e5165.
 - [33] Salih Tutun, Marina E Johnson, Abdulaziz Ahmed, Abdullah Albizri, Sedat Irgil, Ilker Yesilkaya, Esma Nur Ucar, Tanalp Sengun, and Antoine Harfouche. 2023. An AI-based decision support system for predicting mental health disorders. *Information Systems Frontiers* 25, 3 (2023), 1261–1276.
 - [34] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, 3–14.
 - [35] Felix Sebastian Wicke, Lina Krakau, Bernd Löwe, Manfred E Beutel, and Elmar Brähler. 2022. Update of the standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders* 312 (2022), 310–314.
 - [36] World Health Organization. 2024. The Overwhelming Case for Ending Stigma and Discrimination in Mental Health. <https://www.who.int/europe/news/item/26-06-2024-the-overwhelming-case-for-ending-stigma-and-discrimination-in-mental-health> Accessed: 2025-03-05.
 - [37] Hanhui Xu and Kyle Michael James Shuttleworth. 2024. Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm”. *Intelligent Medicine* 4, 1 (2024), 52–57.
 - [38] Wei Xuan, Meghna Roy Chowdhury, Yi Ding, and Yixue Zhao. 2025. Unlocking Mental Health: Exploring College Students' Well-being through Smartphone Behaviors. *arXiv preprint arXiv:2502.08766* (2025).
 - [39] Sheng Yu, Narjes Nourzad, Randy J. Semple, Yixue Zhao, Emily Zhou, and Bhaskar Krishnamachari. 2024. CAREForMe: Contextual Multi-Armed Bandit Recommendation Framework for Mental Health. In *Proceedings of the 11th IEEE/ACM International Conference on Mobile Software Engineering and Systems (Portugal) (MOBILESoft '24)*.
 - [40] Yixue Zhao, Tianyi Li, and Michael Sobolev. 2024. Digital Wellbeing Redefined: Toward User-Centric Approach for Positive Social Media Engagement. In *Proceedings of the IEEE/ACM 11th International Conference on Mobile Software Engineering and Systems*, 95–98.