From Detection to Forecasting: An Empirical Study of College Student Mental Health Using Smartphone Sensing

Jheel Kishore Gala*, Asher Sprigler†, Priyanshu Rajeshbhai Malaviya*, Elizabeth Shen‡, Yixue Zhao§, Yi Ding†, Xipeng Shen*
* North Carolina State University, Raleigh, NC, USA
† Purdue University, West Lafayette, IN, USA
‡ Green Hope High School, Cary, NC, USA
§ Yixue Research Institute, USA

***Abstract*—Mental health challenges among college students continue to rise, motivating scalable and proactive support. Passive smartphone sensing provides an unobtrusive way to capture behavioral signals related to mental well-being, enabling machine learning models that predict future symptom changes. However, most existing work targets mental health *detection*, while leaving mental health *forecasting* underexplored. Forecasting remains technically challenging, and we still lack a clear understanding of which modeling choices most effectively improve long-term forecasting accuracy in realistic deployments. In this paper, we present the first large-scale empirical study of college student mental health forecasting using the College Experience Study (CES), a multi-year longitudinal dataset with passive sensing and weekly surveys. We systematically evaluate three practical design dimensions: (1) single-user forecasting under privacy-restricted, data-scarce settings; (2) model granularity, comparing population-level generic models, similarity-based models, and personalized fine-tuned models; and (3) architecture choice, contrasting one-stage end-to-end forecasting with a two-stage decoupled pipeline. Across controlled comparisons, population-level generalization consistently delivers the highest forecasting performance, achieving up to 0.777 accuracy, while similarity-based transfer and fine-tuning provide limited gains. Two-stage pipelines often reduce accuracy due to objective mismatch between stages. These findings provide actionable baselines and guidance for deployable mental health forecasting.**

I. Introduction

Mental health challenges among college students have become increasingly severe, affecting academic performance, daily functioning, and long-term well-being [1]–[4]. In 2024, roughly 1 in 5 American college students experienced serious psychological distress; 35% of students were diagnosed with anxiety; 25% had depression; 84% faced academic challenges saying it causes them moderate or high distress; 40% of students thought their school was doing enough to support student mental health [5]. Early intervention is particularly important in this setting: when support arrives before symptoms intensify, it can improve academic outcomes, reduce dropout risk, and strengthen students' overall quality of life.

In parallel, smartphones and wearables have enabled continuous, passive sensing of behavioral signals such as mobility, sleep, and device interaction. By combining these signals with machine learning, prior work has developed models that infer mental health status from behavioral traces. Most existing studies, however, focus on *detection*, which identifies current symptoms after they emerge [6]–[17]. Detection is valuable because it can reduce reliance on self-report and expand screening coverage. However, detection is reactive, which means that it identifies problems only after they have occurred, rather than predicting them before onset.

In contrast, *forecasting* aims to predict future mental health changes (e.g., next-week anxiety severity) [18]–[23] and can enable Just-in-Time Adaptive Interventions (JITAI) that act before symptoms worsen [24], [25]. Because forecasting requires models to capture long-term temporal patterns and generalize to unseen future contexts, it is also a more challenging technical problem than detection. Despite growing interest, forecasting mental health from smartphone sensing remains underexplored in large, longitudinal college settings.

In this work, we present the first large-scale empirical study of machine learning forecasting for college student mental health using the College Experience Study (CES) dataset, the most extensive longitudinal passive sensing dataset of college student mental health released in October 2024. CES follows 215 students from 2017–2022 and provides rich passive sensing data alongside weekly surveys, enabling systematic evaluation of forecasting methods at scale. To the best of our knowledge, no prior work has explored machine learning forecasting on the CES dataset. The previous study of examining the same dataset, I-HOPE [11], focused only on detection rather than forecasting. Using CES, we focus on forecasting weekly mental health state (measured by PHQ-4 [26] from self-reports), and we adopt a sliding-window approach to model week-to-week transitions over long horizons.

Our study targets a practical question faced by any deployment: what modeling choices actually improve mental health forecasting accuracy under the same dataset and evaluation pipeline? To answer this, we systematically investigate three core dimensions of model design. ❶ We study the single-user scenario, where each model trains only on an individual's own history to reflect privacy-restricted, data-scarce deployments.

❷ When multi-user data are available, we test model granularity by comparing population-level (generic), similarity-based, and personalized (fine-tuned) models to understand the generalization–personalization tradeoff. ❸ We evaluate model architecture design choice by comparing one-stage end-to-end forecasting against a two-stage pipeline (inspired by the prior work I-HOPE [11]) that first predicts future activity scores (Leisure, Me Time, Phone Time, Sleep, Social Time) and then predicts PHQ-4 from those predicted activity behaviors.

Our results reveal distinct differences between mental health *forecasting* and *detection*, offering novel insights and actionable guidance for building robust forecasting systems. We show that forecasting is feasible even in single-user, data-scarce settings, and we identify effective modeling choices under realistic constraints. When multi-user data are available, population-level models generalize best for PHQ-4 prediction: generic XGBoost achieves a peak accuracy of 0.777, while similarity-based transfer and per-user fine-tuning yield limited and inconsistent gains. We also find that decoupling forecasting into a two-stage pipeline does not improve PHQ-4 forecasting accuracy—exhibiting the opposite pattern from prior detection work [11]—and often degrades performance due to mismatches between intermediate objectives and the final mental health states. Counterintuitively, stronger intermediate predictions in a two-stage architecture do not necessarily translate into improved end-to-end forecasting accuracy. Furthermore, we pinpoint factors that influence forecasting accuracy, such as fluctuations in PHQ-4 over time and class imbalance, directly motivating directions for future work.

In summary, we make the following contributions:

- We provide the first empirical evaluation of mental health forecasting using machine learning on the CES dataset.
- We evaluate mental health forecasting in a privacy-restricted setting where each user's model trains only on their own history, establishing a lower bound for on-device deployment.
- We compare model granularity, personalization, and adaptation strategies under a unified protocol, offering actionable guidance for future work.
- We quantify the impact of one-stage versus two-stage architectures for PHQ-4 forecasting, revealing when additional architecture complexity helps or hurts model performance.

Our broader vision is to move mobile sensing from reactive monitoring toward proactive support for student mental health. We hope this work helps the community ground future forecasting research in realistic deployment constraints – limited per-user data, strong privacy requirements, and long-term behavioral drift – while providing clear baselines and design insights for building reliable mental health forecasters at scale. Ultimately, we aim to enable forecasting models that are not only accurate in offline evaluation, but also practical for driving early, personalized interventions that improve student well-being in real-world settings.

II. Empirical Study Overview

This section provides an overview of our empirical study, including the three key research questions we address (§II-A), the longitudinal smartphone sensing dataset on college student mental health that enabled our analysis (§II-B), and experimental setup along with details of the methodologies (§II-C).

A. Research Questions

We examine mental health *forecasting* for college students from multiple perspectives, covering different real-world usage scenarios, model granularity, and architecture complexity. Through the following research questions (RQs), we aim to establish a solid foundation for future work in this space.

- **RQ**$_1$ – How accurate are the forecasts of future mental health states when each user's model is trained only on their own data?
- **RQ**$_2$ – When data from multiple users are available, what level of model granularity can achieve the best accuracy?
- **RQ**$_3$ – Can two-stage model architecture improve the model accuracy of mental health forecasting?

1) ***RQ***$_1$ **– *Feasibility of Single User Scenario*:** RQ$_1$ studies an extreme data-scarce setting where only a single user's historical data are available to train and update the forecasting model. This scenario reflects the most privacy-restricted real-world deployments, where sharing population data is infeasible or undesirable. We investigate the extent to which mental health forecasting works under this constraint, aiming to set a lower bound on performance and judge whether purely individual-centric forecasting is feasible.

We train a model for each user using only their own data as it arrives over time. When data is scarce, we start with heuristic (as detailed in the process of each model); as more data is collected, the model updates and improves its forecasts. In this single-user setting with limited training data, we focus on data-efficient, sequence-based forecasters. We use Markov models (on value/delta) [27]–[29] as a lightweight baseline that captures short-term state transitions, and the Patch Time Series Transformer (PatchTST) [30] as a transformer that learns long-range temporal dependencies. We do not test classical machine learning models such as linear regression or XGBoost (tested in the data-rich setting in RQ$_2$), which often need handcrafted feature engineering and careful tuning, hence may overfit and behave unstably with only one user's history. Next, we describe the three models used for RQ$_1$.

Markov Model on Value: This model employs a Markov process [27], [28] over observed values to capture the short-term state transition in a user's mental status. The core component consists of transition probabilities $p(y_t \mid y_{t-1}, y_{t-2}, \ldots, y_{t-k})$, where y_i denotes the value at time i. The probabilities are continuously updated as more user data becomes available. The parameter k determines the context length; we set it to be the sliding window size N, as visualized in Figure 1 (detailed in Sliding Window Approach in §II-C1). Predictions are only produced once a full history window of N observations is available. If no matching patterns exist in the prior user history, the model uses the previous time step $(t-1)$ to make the prediction.

Markov Model on Delta: This model is similar to the previous Markov model but operates on the *differences* between

consecutive values rather than the raw values themselves [29]. It models the dynamics of change, where the i-th entry is the difference between observations at times $i+1$ and i. At time t, the model predicts the expected change relative to the previous time step and derives the corresponding predicted value accordingly. If it cannot find a matching pattern in the user's past data, it falls back to using the value at time step $t-1$ as the prediction.

PatchTST: This model has gained attention in mental health research recently [31], [32]. The idea is to turn a long time series into "chunks" (patches), then use a Transformer to learn patterns across those chunks [30], [33], [34]. By using patches, PatchTST captures long-range patterns more effectively. In our experiments, we use the past 8-week data as input and divide it into overlapping 2-week patches, shifting by 1 week each time. We start model updates and evaluation only after a 15-step warm-up period (15 weeks, about $\frac{1}{4}$ of a year) to make online adaptation more stable.

*2) **RQ_2 – Model Granularity**:* RQ_2 considers that we have a large-scale dataset with data from many users for training. In this setting, the key question is not only how much data to use, but also how to build the model at what level of granularity. Prior work shows that more data does not always improve accuracy, especially when user behaviors vary across contexts [35]. To address this gap, we study how model granularity affects mental health forecasting using smartphone sensing data. Specifically, we evaluate three representative levels of granularity:

1) a generic *population-level* model trained on data from all users in the training set (**Generic Model**);
2) a *similarity-based* model that leverages data from behaviorally similar users in the training set (**Similarity-based Model**);
3) a personalized *individual-level* model that is fine-tuned on each user's own data (**Personalized Model**).

Intuitively, a model customized for each user has the potential to outperform a generic model trained on data from many users. However, personalization can also reduce generality: the model may overfit to a user's past data and become less accurate at predicting the user's future states. This tradeoff between generalization and personalization appears in many domains [36], [37], but it can look different in mental health forecasting. We therefore need to study it carefully in this setting to identify the most effective strategy. Since many personalization methods rely on fine-tuning with newly collected user data, we also analyze how fine-tuning interacts with this tradeoff. Next, we describe each model in detail.

❶ **Generic Model (Population-Level):** The Generic Models are trained offline using the data from all users in the training set (detailed in §II-C). Such models capture collective behavioral patterns and global data regularities across the entire cohort. They thus lie at the *generalization* end of the spectrum. Within this class, we constructed seven models: six using classical machine learning methods and one using a transformer approach. We now explain the seven Generic Models at a high level, and the detailed configurations of the models are explained in §II-C.

- **Classical Machine Learning Model:** We include six commonly used regression models: Linear Regression [38], [39], ElasticNet [40], [41], RandomForest Regressor [42], [43], XGBoost Regressor [44]–[46], Ridge Regressor [47], [48], and a two-layer fully connected Neural Network [49], [50]. The input vector concatenates all target-feature values from the previous $N-1$ time steps. Each model then predicts the target-feature values for the next time step.
- **Transformer Model:** To capture long-term temporal dependencies, we use the same PatchTST Transformer architecture, as in RQ_1. We train the model on data from all users in the training set, so it can learn shared temporal trends across users.

❷ **Similarity-based Model Trained on Similar Users' Data:** The Similarity-based Models train on data from the users in the training set who are most similar to the target user. This setup assumes that users with similar behavior may follow similar mental-state patterns, so it keeps a certain degree of personalization while still benefiting from cross-user learning. We use the same models as above here, namely, six classical machine learning models (Linear Regression, ElasticNet, RandomForest Regressor, XGBoost Regressor, Ridge Regressor, two-layer fully connected Neural Network), and one transformer model (PatchTST). §II-C2 will detail how we identify the most similar training users for each target user.

❸ **Personalized Model (Individual-Level):** The Personalized Models are initialized from either the Generic Model or the Similarity-based Model mentioned above. We then fine-tune the models separately for each target user using their own historical data, so it can adapt to individual behavioral patterns. Training runs online: the update occurs after each new observation is collected. These models provide stronger personalization than the earlier classes. For the transformer model (PatchTST), we fine-tune it online with the RMSprop optimizer [51] after each new observation. For the six classical machine learning models, we switch from batch training to incremental training. Specifically, we replace the batch regressors with their SGD regressor counterparts, which are mathematically equivalent but support incremental updates (via the `partial_fit` method in `scikit-learn` [52]).

*3) **RQ_3 – Model Architecture Complexity**:* The standard model uses a *one-stage* design: it directly predicts a user's future mental health state. As illustrated in Figure 2, the recent I-HOPE work [11] introduced a new *two-stage* model that first predicts the user's future behaviors (i.e., activity scores) at the target time, then uses those predicted behaviors as inputs to a second model to estimate the mental health outcome. I-HOPE demonstrated improvement over the one-stage model on the same CES dataset (§II-B) we study. However, as discussed in §I, *detection* and *forecasting* are distinct tasks. It is therefore unclear whether the two-stage design will also benefit mental health forecasting. RQ_3 tests this question and examines whether increased architectural complexity is promising for forecasting. To do so, we build several two-stage frameworks with different model combinations and compare

them against matched one-stage counterparts (see §II-C3).

B. College Experience Study (CES) Dataset

To characterize college student daily behavior and mental health changes, we use the College Experience Study (CES) dataset [53], the longest-running smartphone sensing study for college student mental health so far by following 215 Dartmouth students of two cohorts from 2017 to 2022. Released in October 2024, CES contains over 210,000 hourly data points. Below, we highlight two key parts of the CES dataset that enable our large-scale analysis.

*1) **Smartphone Sensing Features**:* The CES dataset includes rich smartphone sensing data continuously collected from students' phones. It includes 172 features that capture phone use, sleep, mobility, and physical activity. However, prior work I-HOPE found that many of these features do not help mental health prediction, and thus manually selected 35 features that keep domain-specific behavioral meaning while reducing computation cost [11]. I-HOPE further maps these 35 features into five high-level categories: *Leisure*, *Me Time*, *Phone Time*, *Sleep*, and *Social Time*. For each category, it computes a single numeric value, called an **activity score**, that ranges from $0-1$. In I-HOPE's experiments, these five activity scores consistently outperformed the original 35 raw features. We see the same trend when we run our classical machine learning models with the 35 features and compare them with the activity scores. Thus, following I-HOPE's best practice, we use only the five derived activity scores in all experiments in this paper.

*2) **Self-Reported PHQ-4 Score**:* To measure mental health outcomes, Ecological Momentary Assessment (EMA) surveys were sent to students once a week on their phones [54]. The survey includes the Patient Health Questionnaire-4 (PHQ-4) [26], which captures anxiety and depression by asking about students' recent internal feelings. Psychologists have widely used PHQ-4 as a reliable self-report tool for tracking anxiety and depression [55], [56]. PHQ-4 scores range from 0 to 12, where higher values indicate more severe symptoms. Following prior work I-HOPE [11], we group the raw scores into four categories to reduce noise from individual differences in reporting: **Normal** (0–3), **Mild** (4–6), **Moderate** (7–9), and **Severe** (10–12). We use these category labels for model training and evaluation.

C. Experiment Setup

This section describes the setup for RQ_2 and RQ_3 in our empirical study. We first explain how we preprocess the data, split it into training and test sets, and build input sequences using a sliding window (§II-C1). We then present our methods for finding the most similar users (§II-C2). Finally, we introduce the two-stage architecture to study the effect of model complexity (§II-C3).

*1) **Dataset Preprocessing**:* Before training, we preprocess the dataset by filtering users and splitting the data into training and test sets at the user level. We also adopt a sliding window method to turn the longitudinal data into sequences for forecasting.

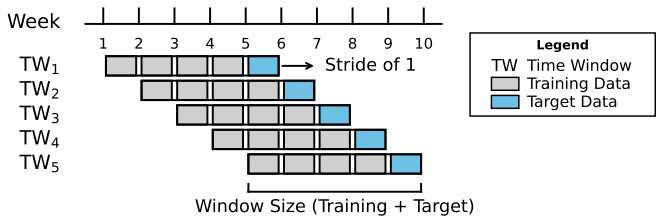

Fig. 1: The sliding window approach for an individual user with window size 5 and training ratio 0.8.

User Selection and Training/Test Split: We keep only participants with at least 160 records to ensure we have enough data for training and forecasting. With window size $N = 5$ (explained below in Sliding Window Approach), each user will have at least 156 samples after windowing. This matches the number of weeks in three years, confirming that the user contributed data for at least three out of the four study years during college. Having multiple years of data captures a wider range of behaviors and helps reveal longer-term patterns. After filtering, we obtain 121 users. We evaluate models using five-fold cross validation with a fixed random seed in our experiments for consistent comparisons. For each model, we first average test results within each user, then average across users in a fold, and finally average across folds. This keeps the evaluation fair regardless of how many observations each user has. In each fold, we test on 32 target users and train Generic Models (as described in §II-A2) on the remaining 89 users. These 89 users produce about 20,000 training samples per fold.

Sliding Window Approach: For all models except for PatchTST (which uses a separate process detailed in §II-A1), we convert each user's longitudinal data into fixed-length time windows based on timestamps. Let N be the window size. After testing various values, we choose $N = 5$ to balance forecasting accuracy and responsiveness. Figure 1 illustrates the case with $N = 5$. With a training ratio of 0.8, we form each input sample from four consecutive observations, where each observation contains five activity scores and one PHQ-4 score. The fifth observation of five activity scores and the PHQ-4 score is the prediction target. This design keeps the input (gray in Figure 1) and target (blue in Figure 1) sharing the same six features. We generate samples using a one-week stride. We impute missing values with the median and normalize all input features using statistics computed across all users, so the scaling reflects population-level patterns.

*2) **Similar User Identification**:* To build the *Similarity-based* Models discussed in §II-A2, we explore similarity-based transfer learning, where knowledge is transferred from the most similar training user to initialize the model for the target test user. This helps address data scarcity for new users by leveraging the richer histories of behaviorally similar

individuals.

Behavior Vector: To make similarity comparison efficient, we compress each user's first 15-week data into a 15-feature behavior vector. This is done by computing three key statistics for each of the five activity scores: mean (average level), standard deviation (variability), and lag-1 autocorrelation (temporal consistency). Using 15 weeks (about one quarter of a year) reflects a realistic deployment setting: it provides sufficient data to capture stable patterns while reducing the noise seen in smaller samples.

Similarity Metrics: Transfer learning works best when accurately finding users with similar behavior. To address this, we compare users using both similarity metrics (Cosine similarity, Pearson correlation) and distance metrics (Euclidean, FastDTW). For behavior vector comparison, we use: ❶ Cosine similarity [57] (maximize), which focuses on the angle of behavior vectors; ❷ Pearson correlation [58] (maximize), which captures linear co-variation; and ❸ Euclidean distance [59] (minimize), which focuses on absolute differences in magnitude. For raw-sequence (temporal) comparison, we use ❹ FastDTW [60] (minimize), which aligns the full activity sequences and accounts for small time shifts.

Transfer and Fine-Tuning: For each test user, we pick the most relevant training user by choosing the highest similarity score (for similarity metrics) or the smallest distance (for distance metrics). By choosing a single user, we ensure a personalized transfer of user-specific behavioral metrics. We then train an SGD-based MultiOutput Regressor (Linear Regression, Ridge, or ElasticNet) on that source user and transfer the learned parameters to the target test user. Then, we fine-tune the transferred model on the test user's data. This step adapts the model to the user's local behavior patterns before we evaluate it on unseen data.

*3) **Two-Stage Architecture Design**:* To answer RQ_3 as discussed in §II-A3, we design a two-stage architecture for mental health forecasting based on prior work I-HOPE [11] to compare with the one-stage model, as illustrated in Figure 2. The architectures and tradeoffs of these two approaches are described below.

One-Stage Design: The one-stage design represents the most straightforward setup. Models in this category (including all the models from §II-A1) directly map past inputs to all future targets at once. In one step, the model predicts all six outputs: five activity scores and one PHQ-4 score. This simplicity comes with a potential drawback: the model must simultaneously optimize for both the continuous activity scores and the categorical PHQ-4 score, which can create competing loss goals and hurt performance.

Two-Stage Design: The two-stage design mitigates this complexity by separating activity score forecast from classifying the categorical PHQ-4 score: ❶ **Stage 1: Activity Score Forecasting:** The initial stage focuses exclusively on predicting only the five continuous activity scores. The output of Stage 1 is the activity score forecast for the next time point. ❷ **Stage 2: PHQ-4 Forecasting:** The forecast activity scores from Stage 1 are the input features to a second model that predicts the PHQ-4 score of that user. This Stage 2 model is the same pre-trained model used in I-Hope [11].

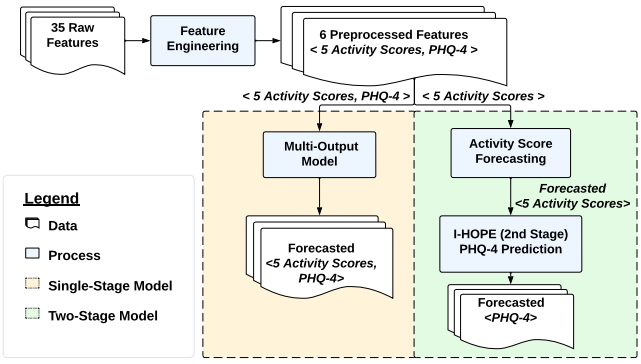

Fig. 2: Comparison of the architectures of One-Stage and Two-Stage designs of the forecasting schemes.

By comparing PHQ-4 accuracy from the one-stage model to the two-stage model, we test whether the two-stage design improves forecast performance. This also allows us to quantify the extent to which improvements in Stage 1 activity score forecasting translate into higher PHQ-4 accuracy, since the Stage 2 classifier remains the same as in I-HOPE.

III. Results and Lessons Learned

This section presents the results of our empirical study and the lessons learned in each of the research questions.

A. RQ_1 – *Feasibility of Single User Scenario*

As discussed in §II-A1, RQ_1 explores the scenario where only a single user's data is available to train the model, simulating the most privacy-restricted real-world deployments.

*1) **Results on Single User Scenario**:* Figure 3 shows the forecast accuracy of the three models described in §II-A1. We train each model using only a single user's own data and evaluate them across 121 users. The violin plot shows the accuracy distributions across users, and we also mark the mean, max, and min values. Among the three models, the Transformer-based PatchTST achieves the highest mean accuracy (0.744), followed closely by the Markov Model on Value (0.708). The Markov Model on Delta performs worst, with a mean accuracy of 0.682. Overall, these results show that models designed to capture long-range temporal dependencies, such as PatchTST, outperform short-term state transition Markov models that rely mainly on recent history when training on their own data. This finding highlights the limits of short-term sequence models in data-limited, privacy-preserving settings and underscores the strength of long-term sequence methods in this scenario.

Although the average accuracies differ only slightly, all three models exhibit large variability across users. The minimum accuracies range from 0.306 to 0.409, while some users reach perfect accuracy (1.000). This spread highlights substantial heterogeneity in user-level forecasting performance: a model can perform well for some users but poorly for others.

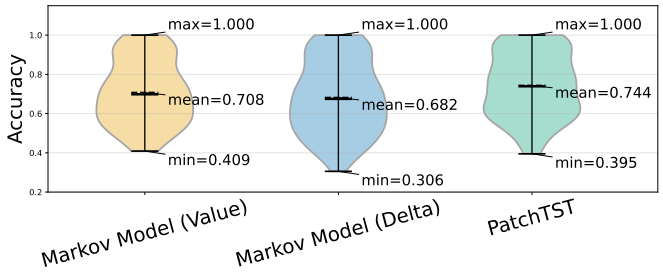

Fig. 3: Forecasting accuracy distribution of PHQ-4 scores among three single-user models, with the minimum, mean, and maximum values marked (RQ_1, §III-A).

TABLE I: Pearson correlations and p-values between forecast accuracy and PHQ-4 standard deviation among three models in single user scenario (RQ_1, §III-A).

Model	**Pearson Correlation**	p**-value**
Markov Model (Value)	-0.886	1.881×10^{-41}
Markov Model (Delta)	-0.854	1.499×10^{-35}
PatchTST	-0.889	2.937×10^{-42}

To further investigate this effect, we calculate the Pearson correlations and corresponding p-values between forecast accuracy and the standard deviation of PHQ-4 scores across three models in the single-user setting in Table I. We observe strong negative correlations between forecast accuracy and PHQ-4 variability, all of which are statistically significant ($p < 0.01$). This indicates that users with more stable PHQ-4 trajectories tend to achieve higher forecasting accuracy than users with highly variable patterns. We further validate this finding by inspecting the raw data: users with the highest accuracy consistently exhibit stable PHQ-4 scores over time, whereas users with larger fluctuations in behavior and PHQ-4 experience less stable and less accurate forecasts. Finally, the high average accuracy from the three models suggests the feasibility of deploying personalized models even under strict privacy constraints. These results offer promising evidence for personalized modeling and motivate future work to identify which users benefit most from such approaches.

B. RQ_2 – Model Granularity

As described in §II-A2, RQ_2 explores three levels of model granularity: ❶ Generic Model (population-level) trained on all training users' data; ❷ Similarity-based Model trained on the data from a subset of training users who are most similar to the target user; and ❸ Personalized Model (individual-level) fine-tuned based on the first two classes of models.

*1) **Results on Model Granularity**:* Table II summarizes the results for RQ_2. The first column (No finetune on target user's data) reports the performance of the first two model classes: Generic Models and Similarity-based Models. The second column (Finetuned on target user's data) reports the performance for Personalized Models. Since this is a multi-class classification task, we report accuracy, precision, recall, and F1 score to provide a complete view of performance across all four classes. Because RandomForest and XGBoost do not naturally support incremental updates without retraining [61], [62], their fine-tuning results are not reported.

Overall, the results show that generalization outperforms personalization for PHQ-4 forecasting in most cases. In particular, Generic Models trained on the full training population achieve the best performance, where the Generic XGBoost Regressor reaches the highest PHQ-4 accuracy of **0.777**. Looking closer at the first column (No finetune on target user's data), Generic Models consistently outperform Similarity-based Models across all six classical machine learning methods. For the transformer approach PatchTST, the two model classes achieve similar performance, suggesting that PatchTST is less sensitive to whether the training data come from all users or only similar users.

In the second column (Finetuned on target user's data), fine-tuning does not help Generic Models where PHQ-4 accuracy drops for both classical machine learning and transformer models. In contrast, fine-tuning improves the accuracy of all Similarity-based models except ElasticNet, for which performance remains unchanged. A notable case is the Neural Network, which benefits substantially more from fine-tuning than any other model. This suggests that neural networks are especially effective at leveraging per-user data to adapt, capturing user-specific nonlinear patterns and temporal dependencies that are not well modeled by globally trained or linear approaches. In contrast, other models exhibit less benefit from personalization, as their representations are less flexible to individual differences.

We also see a consistent trend: all models perform best on the Normal class, and performance steadily drops as mental health severity increases. This pattern is expected given the significant class imbalance in the dataset. Across the 121 users in our study, 62.87% of PHQ-4 reports fall into Normal, followed by 26.03% Mild, 7.24% Moderate, and only 3.87% Severe. With far fewer training data for higher-severity classes, the models struggle to learn these patterns and often miss Moderate and Severe cases, leading to lower recall and F1. In addition, the Severe class often achieves higher precision and F1 than Moderate, even though it has fewer samples. This happens because the middle categories are easier to mix up: a Moderate sample is likely to overlap with both Mild and Severe. In contrast, Severe cases usually show stronger behavior changes, so models can separate them more clearly.

To alleviate class imbalance in the dataset, we applied data augmentation by oversampling the minority Severe class [63] and reran all experiments. However, this approach did not improve overall accuracy. As this work focuses on establishing a large-scale empirical foundation for mental health forecasting, the design of new methods to further improve performance in challenging settings is beyond the scope. We therefore leave the development of more effective techniques to improve accuracy and recall for minority classes to future work.

*2) **Results on Impact of Fine-Tuning Across Similarity Metrics**:* Figure 4 compares PHQ-4 forecasting accuracy

TABLE II: Forecasting Results at Different Model Granularity. Overall accuracy (Acc.) is reported per model, while precision (P), recall (R), and F1 are reported per PHQ-4 category (RQ_2, §III-B).

Initialization	Model	No Finetune on Target User's Data												Finetuned on Target User's Data (Personalized Model)											
			Normal			Mild			Moderate			Severe				Normal			Mild			Moderate			Severe	
		Acc.	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc.	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Generic Model	Linear Reg.	0.772	0.888	0.872	0.880	0.602	0.678	0.638	0.440	0.421	0.430	0.700	0.344	0.457	0.732	0.722	0.813	0.758	0.458	0.384	0.401	0.183	0.112	0.129	0.113	0.083	0.089
	Ridge Reg.	0.772	0.888	0.872	0.880	0.602	0.678	0.638	0.440	0.421	0.430	0.700	0.342	0.456	0.732	0.722	0.813	0.758	0.458	0.384	0.401	0.182	0.112	0.128	0.113	0.083	0.089
	ElasticNet	0.772	0.890	0.870	0.880	0.600	0.688	0.641	0.441	0.417	0.429	0.709	0.334	0.450	0.734	0.711	0.848	0.768	0.427	0.332	0.356	0.153	0.100	0.089	0.055	0.064	0.064
	Neural Network	0.773	0.876	0.893	0.885	0.621	0.624	0.622	0.418	0.379	0.397	0.580	0.461	0.508	0.728	0.809	0.931	0.865	0.593	0.465	0.519	0.417	0.287	0.335	0.552	0.303	0.369
	RandomForest	0.757	0.881	0.871	0.876	0.590	0.660	0.623	0.385	0.398	0.390	0.639	0.182	0.275	-	-	-	-	-	-	-	-	-	-	-		
	XGBoost	**0.777**	0.879	0.889	0.884	0.616	0.662	0.638	0.449	0.405	0.425	0.720	0.322	0.434	-	-	-	-	-	-	-	-	-	-	-		
	PatchTST	0.753	0.737	0.814	0.772	0.455	0.403	0.416	0.144	0.125	0.124	0.068	0.034	0.042	0.752	0.748	0.787	0.765	0.461	0.434	0.439	0.168	0.143	0.145	0.084	0.050	0.060
Similarity-based Model	Linear Reg.	0.591	0.597	0.967	0.686	0.062	0.025	0.029	0.016	0.004	0.005	0.014	0.004	0.005	0.596	0.597	0.975	0.691	0.061	0.022	0.026	0.023	0.011	0.009	0.026	0.004	0.005
	Ridge Reg.	0.591	0.597	0.969	0.687	0.062	0.023	0.028	0.016	0.011	0.006	0.014	0.002	0.003	0.596	0.597	0.976	0.691	0.060	0.021	0.026	0.021	0.007	0.006	0.026	0.003	0.004
	ElasticNet	0.596	0.596	0.990	0.694	0.010	0.001	0.002	0.008	0.000	0.000	0.008	0.000	0.000	0.596	0.596	0.990	0.694	0.013	0.002	0.002	0.000	0.000	0.000	0.017	0.000	0.000
	Neural Network	0.511	0.667	0.620	0.546	0.309	0.326	0.252	0.089	0.098	0.048	0.025	0.024	0.014	0.594	0.690	0.609	0.617	0.333	0.372	0.332	0.092	0.152	0.103	0.063	0.047	0.046
	RandomForest	0.594	0.679	0.710	0.623	0.316	0.385	0.288	0.044	0.058	0.041	0.017	0.013	0.014	-	-	-	-	-	-	-	-	-	-	-		
	XGBoost	0.579	0.666	0.696	0.618	0.330	0.384	0.291	0.073	0.046	0.037	0.016	0.007	0.008	-	-	-	-	-	-	-	-	-	-	-		
	PatchTST	0.753	0.760	0.788	0.772	0.463	0.439	0.439	0.156	0.139	0.135	0.082	0.046	0.056	0.754	0.759	0.800	0.777	0.465	0.427	0.435	0.169	0.145	0.146	0.088	0.046	0.057

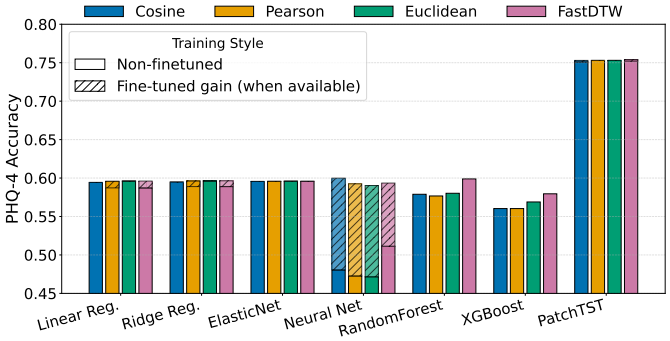

Fig. 4: Forecasting accuracy of Similarity-based models with and without fine-tuning across four similarity metrics (RQ_2, §III-B).

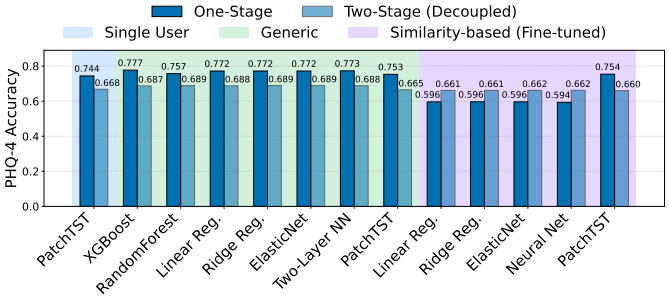

Fig. 5: Comparison of forecasting accuracy between one-stage and two-stage designs. Included are the single-user PatchTST (The only Single User model to produce Activity Scores), all Generic models, and then the Similarity-based models with fine-tuning (RQ_3, §III-C).

across seven Similarity-based Models under four similarity metrics. As indicated in §III-B1, RandomForest and XGBoost do not naturally support incremental updates without retraining, their fine-tuning results are not reported [61], [62]. For each model, the plot shows performance under non-finetuned training (baseline) and overlays the finetuned gain when available. Overall, the accuracy values range from 0.4 to 0.8, suggesting that the choice of model and similarity definition both affect performance but do not lead to extreme swings. Fine-tuning provides inconsistent but sometimes noticeable improvements, with some models showing clear upward shifts while others gain little, implying that the benefit of fine-tuning depends on whether the base model can effectively exploit personalized patterns under the selected similarity metric. Taken together, this result suggests that metric choice interacts with model capacity and training strategy, and the best results come from pairing a robust learner with a similarity metric that aligns with the underlying behavioral signal.

C. RQ_3 – Model Architecture Complexity

As described in §II-A3, RQ_3 compares two architectures: one-stage (end-to-end) design and two-stage design that separates prediction into two steps. Figure 2 shows their design differences. For both architectures, we evaluate 13 models in total, including the Single user PatchTST (the only single user model from RQ_1 that generates activity scores) along with the Generic Models and the Similarity-based Models from RQ_2.

*1) **Results on Model Complexity**:* Figure 5 compares the forecasting accuracy of one-stage design and two-stage (decoupled) design using final PHQ-4 forecasting accuracy as the metric. Overall, 9 of the 13 models (including the PatchTST in the Single User scenario, all Generic models, and PatchTST in the Similarity-based group) show lower accuracy when moving from the one-stage to the two-stage design. Only 4 Similarity-based models improve in the two-stage setting. A likely reason is that the two-stage pipeline introduces an extra interface between stages, which can amplify errors: mistakes in the first stage propagate to the second, and the second stage cannot recover if the intermediate outputs lose important information. In contrast, one-stage models optimize the full mapping jointly to maximize final accuracy.

These results contrast with the two-stage detection framework introduced in I-HOPE [11], highlighting that forecasting and detection are fundamentally different problems and that insights from detection do not necessarily carry over to forecasting. This distinction further underscores the value of our work, as the majority of existing studies focus on detection rather than forecasting.

*2) **Results on Activity Score Forecasting.**:* We also conduct a breakdown analysis to examine the two-stage pipeline by

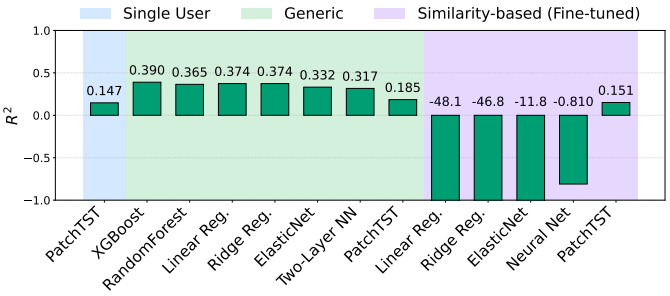

Fig. 6: Forecasting accuracy of activity scores in R^2 for Stage 1 in two-stage design (RQ_3, §III-C).

evaluating the forecasting accuracy of the intermediate activity score predicted in Stage 1 and then used as input to Stage 2. Figure 6 reports these results, using R^2 since activity score is continuous. Interestingly, the patterns are counterintuitive: the 9 models whose final PHQ-4 accuracy drops in the two-stage design actually achieve positive R^2 (better activity score forecasting), while the 4 models whose final PHQ-4 accuracy improves in the two-stage design show negative R^2 (worse activity score forecasting).

This suggests that better intermediate forecasts do not necessarily translate into better end-to-end performance. One possible reason is that Stage 2 may rely more on how the intermediate representation encodes information (e.g., its alignment with mental health signals) than on raw point-wise accuracy. In other words, Stage 1 can produce activity forecasts that look accurate numerically but remove or smooth out subtle patterns that are important for PHQ-4 prediction. In contrast, noisier activity predictions may retain useful variations or structure that Stage 2 can exploit, leading to higher final accuracy. This mismatch highlights a limitation of decoupled designs: optimizing each stage independently can produce intermediate outputs that are good under their own metric, but not optimal for the downstream mental-health forecasting task.

Compared to the prior work I-HOPE [11], which does not analyze the relationship between intermediate-stage outputs and final prediction accuracy, this work takes a further step by examining how forecasting accuracy trends differ from the Stage-1 output to the final predictions in a two-stage design. Therefore, our analysis provides new insights into how intermediate stage representations influence downstream forecasting performance, revealing design tradeoffs and limitations that are unique to forecasting and are not observable in detection tasks.

IV. Discussion

Our study is observational and relies on passive smartphone sensing plus weekly self-reported PHQ-4 surveys, so unobserved confounders may drive part of the measured relationship between behavior and mental health. For example, stressful life events, coursework load, therapy/medication changes, or social context shifts are not fully captured by sensing features, yet they can strongly affect PHQ-4. In addition, forecast outcomes may reflect measurement noise: PHQ-4 is a short questionnaire and self-reports can vary due to reporting bias, recall effects, and momentary mood. To reduce variability, we follow prior CES work [11] and group raw PHQ-4 scores into four categories, but this discretization may also hide clinically meaningful small changes. Finally, our preprocessing choices (e.g., imputing missing values and normalizing across users) and design decisions (e.g., sliding window formulation) could influence model behavior and should be interpreted as one reasonable pipeline rather than the only correct one.

Our results are based on the specific CES dataset, which tracks Dartmouth students over multiple years, and may not generalize to other populations such as non-college young adults, students at different institutions, or groups with different demographics, stressors, and support resources. The sensing modality and feature construction also limit generalization: we use five derived activity scores (Leisure, Me Time, Phone Time, Sleep, Social Time) that summarize a subset of raw signals, and models trained on these abstractions may behave differently if future deployments use different sensors, sampling rates, wearable inputs, or feature definitions. Moreover, our evaluation focuses on forecasting accuracy for weekly PHQ-4 categories; real-world deployment would also require robustness to missing data, behavior drift, phone OS differences, and user disengagement. These factors may reduce performance outside the controlled dataset setting and motivate follow-up studies across sites and sensing configurations.

The class imbalance in the CES dataset poses challenges for training reliable forecasting models and raises ethical concerns for potential clinical deployment. Because any fair training or evaluation distribution is dominated by users with Normal users, model performance disproportionately degrades as symptom severity increases, as reflected in the declining recall for Severe class. In practice, deploying such a model could result in users with Severe symptoms being under-identified, preventing them from receiving timely interventions. These results highlight a fundamental tradeoff between sensitivity to high-risk individuals and the risk of over-intervention in low-risk populations. Addressing this tradeoff requires careful model design, rigorous evaluation across severity levels, and thorough validation prior to any real-world use. Further research is needed to develop principled data augmentation and testing methodologies before such forecasting approaches can be considered suitable for clinical settings.

V. Related Work

In this section, we first review mobile sensing and behavior modeling, and then describe prior work on mental health detection and forecasting using machine learning.

❶ **Mobile Sensing and Behavior Modeling:** Smartphones and wearables can capture rich signals about daily behavior, such as location changes and text or voice communication, as well as physiological measures like heart rate and sleep duration [64]–[66]. Over the past decade, researchers have collected longitudinal datasets and used them to model and predict user behavior from these activity traces. This progress

has enabled new behavioral modeling methods across many important behavior patterns [67], [68].

Mobile sensing has driven many machine learning and deep learning methods that use behavioral data to model human activities. [69] built a lightweight CNN that detects activities such as sleeping, walking, and running from smartphone signals. This type of activity detection provides a core building block for behavior modeling. Along the same lines, [70] developed compact deep learning models that fuse multiple sensing modalities and capture temporal patterns, while remaining small enough to run on mobile devices. These efforts highlight the need for models that stay lightweight but still track behavior changes over time. Recent work also shows that temporal activity recognition helps detect higher-level behavioral routines and long-term patterns [71], [72].

A major challenge is that behavior varies widely across individuals, which makes it hard for models to generalize. [73] addresses this issue in speech therapy and stress sensing by introducing personalization methods that improve performance across users. Similarly, [74] introduces a framework that adapts to new users using unlabeled data while keeping computation low, making on-device or in-the-field training more practical. These approaches also reduce the need to centralize personal data, which can improve privacy, an important concern in health-related sensing applications [75].

❷ **Mental Health Detection Using Machine Learning and Mobile Sensing:** Hundreds of millions of people worldwide struggle with mental health disorders, with significant rates of depression and anxiety. Alarmingly, depression rose in prevalence through the COVID-19 pandemic to impact over 216 million people globally and approximately 7.2% of U.S. adults annually [76], [77]. Traditional screening methods rely on self-reports and clinical assessments, which are subjective, time-consuming, and collected only occasionally. In contrast, the widespread use of smartphones and wearables now allows continuous, passive, and unobtrusive monitoring of everyday behavior in real-world settings [14], [78], [79].

To detect mental health issues accurately, we need reliable and continuous monitoring of a person's daily behavior. Early work attempting to connect mobile sensing data with mental health revolved around determining the relationship between certain measured behaviors and mental health risk [80]–[82]. By determining what actions might be correlated to rates of depression and anxiety, certain distributions of activities could be used as early warning signs. One such example is [83], which found that increased detection of speech, possibly related to increased social interaction, was associated with better mental health outcomes. [84], corroborated with that paper that certain behaviors such as staying in sedentary behavior, or irregular/extended periods of sleep were correlated with depressive symptoms. These features can be easily extracted from the use of a mobile device, through audio capture [85], global positioning [86], and physiological metrics [87], [88] .

Researchers have framed mental health detection as a machine learning classification task. This work has grown with the rise of large longitudinal sensing datasets StudentLife [89], GLOBEM [7], and College Experience Study [53] datasets. Recent advances also illustrate a move toward personalization in mental health prediction. For instance, [90] introduced automated feature extraction and collaborative-filtering methods that adapt to individual behavioral patterns, marking a step toward personalized detection systems. This line of work builds on earlier multimodal approaches, such as [91] and [10], who combined smartphone and wearable features to improve depression detection performance in college populations.

More recent work uses deep learning to capture longer and richer temporal patterns in mobile sensing data, including multi-layer perceptron [11], convolutional neural networks [92], [93], Long Short-Term Memory (LSTM) [94], [95], and transformer [22], [34]. By using diverse behavioral and physiological signals, these models often improve detection of emotions and stress [96]. However, generalization remains difficult: models trained on one group often fail on new users because human behavior varies widely. To reduce this gap, researchers use data merging, transfer learning, and cross-dataset methods to predict mood, personality, and stress more reliably [16], [97], [98]. For instance, [7] studies how to maintain reliability across institutions and collection periods. Surveys also summarize a standard workflow for mobile mental health sensing: collect and clean data, build features, train models, and evaluate results [99]–[101].

❸ **Mental Health Forecasting Using Machine Learning and Mobile Sensing:** Forecasting mental health is more challenging than detection because it must predict how behavior and mental state change over time, instead of just identifying problems that already exist. [18] showed that wrist-worn physiological signals can predict next-day mental and physical health in office workers, supporting short-term forecasting from continuous sensing. [19] used social network activity and digital interactions to forecast mental health, but found that models often struggle to generalize across individuals. Together, these studies motivate longer-term forecasting using richer, multimodal behavioral and device-based data.

Prior work has explored different models depending on the forecasting horizon. [20] used multi-task recurrent neural networks to make near-term forecasts from sparse self-reports and highlighted the value of temporal context. [22] later applied transformer attention to track fine-grained emotional changes. For longer-term prediction, [21] combined mobile sensing with EMAs and showed that personalized models can outperform population-level models. More recently, [23] used exponentially weighted moving averages tailored to each user and found that even limited baseline data can support reliable, individualized depression forecasting.

VI. Conclusion

This paper presents the first large-scale empirical study of machine learning–based forecasting for college student mental health using the College Experience Study (CES) dataset. Instead of focusing on detection, we tackle the more challenging task of forecasting by predicting future mental health states based on the passive sensing behavioral data.

We systematically analyze key design dimensions in building forecasting models, including single-user privacy-restricted scenarios, model granularity (generic vs. similarity-based vs. personalized), and architectural design (one-stage vs. two-stage). Our study reveals new insights that differ from those found in detection-focused work. In particular, we show that decoupled two-stage pipelines, which perform well for detection tasks such as I-HOPE, do not consistently improve forecasting accuracy and can suffer from mismatches between intermediate objectives and downstream predictions. We also demonstrate that forecasting remains feasible in single-user, privacy-constrained settings, using models trained only on an individual user's own data. Together, these findings provide practical guidance for designing future mental health forecasting systems using smartphone sensing data.

REFERENCES

[1] D. Hammoudi Halat, A. Soltani, R. Dalli, L. Alsarraj, and A. Malki, "Understanding and fostering mental health and well-being among university faculty: A narrative review," *Journal of clinical medicine*, vol. 12, no. 13, p. 4425, 2023.

[2] T. Chu, X. Liu, S. Takayanagi, T. Matsushita, and H. Kishimoto, "Association between mental health and academic performance among university undergraduates: The interacting role of lifestyle behaviors," *International Journal of Methods in Psychiatric Research*, vol. 32, no. 1, p. e1938, 2023.

[3] G. Barbayannis, M. Bandari, X. Zheng, H. Baquerizo, K. W. Pecor, and X. Ming, "Academic stress and mental well-being in college students: correlations, affected groups, and covid-19," *Frontiers in psychology*, vol. 13, p. 886344, 2022.

[4] J. Colarossi, "Mental health of college students is getting worse," *The Brink*, 2022.

[5] BestColleges, "College student mental health statistics," https://www.bestcolleges.com/research/college-student-mental-health-statistics/, 2024.

[6] X. Xu, H. Zhang, Y. Sefidgar, Y. Ren, X. Liu, W. Seo, J. Brown, K. Kuehn, M. Merrill, P. Nurius *et al.*, "Globem dataset: multi-year datasets for longitudinal human behavior modeling generalization," *Advances in neural information processing systems*, vol. 35, pp. 24 655–24 692, 2022.

[7] X. Xu, X. Liu, H. Zhang, W. Wang, S. Nepal, Y. Sefidgar, W. Seo, K. S. Kuehn, J. F. Huckins, M. E. Morris *et al.*, "Globem: Cross-dataset generalization of longitudinal human behavior modeling," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–34, 2023.

[8] J. E. Bardram and A. Matic, "A decade of ubiquitous computing research in mental health," *IEEE Pervasive Computing*, vol. 19, no. 1, pp. 62–72, 2020.

[9] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr, "Harnessing context sensing to develop a mobile intervention for depression," *Journal of medical Internet research*, vol. 13, no. 3, p. e1838, 2011.

[10] R. Wang, W. Wang, A. DaSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking depression dynamics in college students using mobile phone and wearable sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–26, 2018.

[11] M. R. Chowdhury, W. Xuan, S. Sen, Y. Zhao, and Y. Ding, "Predicting and understanding college student mental health with interpretable machine learning," in *2025 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2025.

[12] D. A. Adler, V. W.-S. Tseng, G. Qi, J. Scarpa, S. Sen, and T. Choudhury, "Identifying mobile sensing indicators of stress-resilience," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 5, no. 2, pp. 1–32, 2021.

[13] I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, and J.-P. Onnela, "Relapse prediction in schizophrenia through digital phenotyping: a pilot study," *Neuropsychopharmacology*, vol. 43, no. 8, pp. 1660–1666, 2018.

[14] R. Wang, M. S. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer, and D. Tseng, Vincent W. S.and Ben-Zeev, "Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 2016, pp. 886–897.

[15] D. A. Adler, C. A. Stamatis, J. Meyerhoff, D. C. Mohr, F. Wang, G. J. Aranovich, S. Sen, and T. Choudhury, "Measuring algorithmic bias to analyze the reliability of ai tools that predict depression risk using smartphone sensed-behavioral data," *npj Mental Health Research*, vol. 3, no. 1, p. 17, 2024.

[16] L. Meegahapola, W. Droz, P. Kun, A. de Götzen, C. Nutakki, S. Diwakar, S. R. Correa, D. Song, H. Xu, M. Bidoglia, G. Gaskell, A. Chagnaa, A. Ganbold, T. Zundui, C. Caprini, D. Miorandi, A. Hume, J. L. Zarza, L. Cernuzzi, I. Bison, M. R. Britez, M. Busso, R. Chenu-Abente, C. Günel, F. Giunchiglia, L. Schelenz, and D. Gatica-Perez, "Generalization and personalization of mobile sensing-based mood inference models: an analysis of college students in eight countries," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 6, no. 4, pp. 1–32, 2023.

[17] L. Meegahapola, H. Hassoune, and D. Gatica-Perez, "M3bat: Unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–30, 2024.

[18] T. Umematsu, A. Sano, S. Taylor, M. Tsujikawa, and R. W. Picard, "Forecasting stress, mood, and health from daytime physiology in office workers and students," in *2020 42nd annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5953–5957, not mobile sensing data.

[19] A. M. Langener, L. F. Bringmann, M. J. Kas, and G. Stulp, "Predicting mood based on the social context measured through the experience sampling method, digital phenotyping, and social networks," *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 51, no. 4, pp. 455–475, 2024.

[20] D. Spathis, S. Servia-Rodriguez, K. Farrahi, C. Mascolo, and J. Rentfrow, "Sequence multi-task learning to forecast mental wellbeing from sparse self-reported data," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2886–2894.

[21] A. Kathan, M. Harrer, L. Küster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, , I. Grossmann, D. D. Ebert, and B. W. Schuller, "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," *Frontiers in digital health*, vol. 4, p. 964582, 2022.

[22] L. Paz-Arbaizar, J. Lopez-Castroman, A. Artés-Rodríguez, P. M. Olmos, and D. Ramírez, "Emotion forecasting: A transformer-based approach," *Journal of Medical Internet Research*, vol. 27, p. e63962, 2025.

[23] E. Schat, F. Tuerlinckx, M. J. Schreuder, B. De Ketelaere, and E. Ceulemans, "Forecasting the onset of depression with limited baseline data only: A comparison of a person-specific and a multilevel modeling based exponentially weighted moving average approach." *Psychological Assessment*, vol. 36, no. 6-7, p. 379, 2024.

[24] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, "Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support," *Annals of behavioral medicine*, 2017.

[25] P. Klasnja, S. Smith, N. J. Seewald, A. Lee, K. Hall, B. Luers, E. B. Hekler, and S. A. Murphy, "Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of heartsteps," *Annals of Behavioral Medicine*, vol. 53, no. 6, pp. 573–582, 2019.

[26] K. Kroenke, R. L. Spitzer, J. B. Williams, and B. Löwe, "An ultra-brief screening scale for anxiety and depression: the phq–4," *Psychosomatics*, vol. 50, no. 6, pp. 613–621, 2009.

[27] D. Claudio, S. Moyce, T. Albano, E. Ibe, N. Miller, and M. O'Leary, "A markov chain model for mental health interventions," *International*

Journal of Environmental Research and Public Health, vol. 20, no. 4, 2023.

[28] W. J. Hulme, G. P. Martin, M. Sperrin, A. J. Casson, S. Bucci, S. Lewis, and N. Peek, "Adaptive symptom monitoring using hidden markov models – an application in ecological momentary assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1770–1780, 2021.

[29] D. Kulić and Y. Nakamura, "Incremental learning of human behaviors using hierarchical hidden markov models," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4649–4655.

[30] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *The Eleventh International Conference on Learning Representations*, 2023.

[31] F. Rizza, G. Bellitto, S. Calcagno, and S. Palazzo, "Exploring wearable emotion recognition with transformer-based continual learning," in *International Workshop on Personalized Incremental Learning in Medicine*. Springer, 2024, pp. 86–101.

[32] S. Saini and P. Sen, "Early detection of anorexia from reddit posts using time series based transformer model," *Discover Computing*, vol. 29, no. 1, p. 19, Jan. 2026.

[33] M. Z. Hossain, S. Sahoo, C. Shende, P. Patel, R. Morillo, Z. Pan, X. Wang, J. Bi, J. Kamath, A. Russell, D. Song, and B. Wang, "Predicting depression treatment outcome using daily step count sensory data," *ACM Trans. Comput. Healthcare*, Nov. 2025, just Accepted.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] Y. Zhao, S. Yin, A. Sejfia, M. S. Laser, H. Wang, and N. Medvidovic, "Assessing the feasibility of web-request prediction models on mobile platforms," in *Proceedings of the 8th IEEE/ACM International Conference on Mobile Software Engineering and Systems*, ser. MOBILESoft '21, 2021.

[36] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "Deep learning and model personalization in sensor-based human activity recognition," *Journal of Reliable Intelligent Environments*, vol. 9, no. 1, pp. 27–39, Mar. 2023.

[37] C. Ding, T. Yao, C. Wu, and J. Ni, "Advances in deep learning for personalized ECG diagnostics: A systematic review addressing inter-patient variability and generalization constraints," *Biosensors and Bioelectronics*, vol. 271, p. 117073, Mar. 2025.

[38] A. Mishra, A. Singh, A. Singh, M. Chauhan, and D. Kamboj, "Analysis of mental health disorders from survey reports using time series based linear regression," in *2024 2nd International Conference on Networking, Embedded and Wireless Systems (ICNEWS)*, 2024, pp. 1–6.

[39] N. Pritam, K. S. Gill, M. Kumar, R. Rawat, and D. Banerjee, "Classification of student mental health analysis using logistic regression and other classification techniques through machine learning methods," in *2024 3rd International Conference for Innovation in Technology (INOCON)*, 2024, pp. 1–5.

[40] M. A. Abid, Z. Dehghan, T. Shinde, and G. Narang, "Machine learning based approaches for identification and prediction of diverse mental health conditions," in *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, vol. 1. IEEE, 2023, pp. 1–5.

[41] F. Li, P. Xu, S. Zheng, W. Chen, Y. Yan, S. Lu, and Z. Liu, "Photoplethysmography based psychological stress detection with pulse rate variability feature differences and elastic net," *International Journal of Distributed Sensor Networks*, vol. 14, no. 9, p. 1550147718803298, 2018.

[42] N. Saravanan, G. Moheshkumar, M. S. VM *et al.*, "Accurate prediction and detection of suicidal risk using random forest algorithm," in *2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN)*. IEEE, 2024, pp. 287–292.

[43] Y. Su, L. Ge, and G. Wei, "Random forest model predicts stress level in a sample of 18,403 college students," in *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Big Data and Algorithms*, ser. CAIBDA '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 588–593.

[44] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794.

[45] S. Verma, C. Sharma, G. Aggarwal, and P. Upadhya, "Artificial intelligence-based approach for classification and prediction of mental health," in *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2024, pp. 708–713.

[46] A. Sharma and W. J. M. I. Verbeke, "Improving diagnosis of depression with xgboost machine learning model and a large biomarkers dutch dataset (n = 11,081)," *Frontiers in Big Data*, vol. Volume 3 - 2020, 2020.

[47] A. Shrestha, S. Bergquist, E. Montz, and S. Rose, "Mental health risk adjustment with clinical categories and machine learning," *Health Services Research*, vol. 53, no. S1, pp. 3189–3206, 2018.

[48] Y. Wang, X. Wang, L. Zhao, and K. Jones, "A case for the use of deep learning algorithms for individual and population level assessments of mental health disorders: Predicting depression among china's elderly," *Journal of Affective Disorders*, vol. 369, pp. 329–337, 2025.

[49] H. Lu, S. Uddin, F. Hajati, M. Khushi, and M. A. Moni, "Predictive risk modelling in mental health issues using machine learning on graphs," in *Proceedings of the 2022 Australasian Computer Science Week*, ser. ACSW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 168–175.

[50] T. Sharma, R. Panchendrarajan, and A. Saxena, "Characterisation of mental health conditions in social media using deep learning techniques," *Deep Learning for Social Media Data Analytics*, pp. 157–176, 2022.

[51] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[53] S. Nepal, W. Liu, A. Pillai, W. Wang, V. Vojdanovski, J. F. Huckins, C. Rogers, M. L. Meyer, and A. T. Campbell, "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*, vol. 8, no. 1, pp. 1–37, 2024.

[54] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, no. 1, pp. 1–32, 2008.

[55] B. Löwe, I. Wahl, M. Rose, C. Spitzer, H. Glaesmer, K. Wingenfeld, A. Schneider, and E. Brähler, "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*, vol. 122, no. 1-2, pp. 86–95, 2010.

[56] K. Kroenke, R. L. Spitzer, J. B. Williams, and B. Löwe, "An ultra-brief screening scale for anxiety and depression: The phq–4," *Psychosomatics*, vol. 50, no. 6, pp. 613–621, 2009.

[57] P. Xia, L. Zhang, and F. Li, "Learning similarity with cosine similarity ensemble," *Information sciences*, vol. 307, pp. 39–52, 2015.

[58] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[59] P.-E. Danielsson, "Euclidean distance mapping," *Computer Graphics and image processing*, vol. 14, no. 3, pp. 227–248, 1980.

[60] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent data analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[61] E. Ikonomovska, J. Gama, and S. Džeroski, "Learning model trees from evolving data streams," *Data Mining and Knowledge Discovery*, vol. 23, no. 1, pp. 128–168, Jul. 2011.

[62] C. Zhang, Y. Zhang, X. Shi, G. Almpanidis, G. Fan, and X. Shen, "On Incremental Learning for Gradient Boosting Decision Trees," *Neural Processing Letters*, vol. 50, no. 1, pp. 957–987, Aug. 2019.

[63] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[64] A. Seifert, M. Hofer, and M. Allemand, "Mobile data collection: Smart, but not (yet) smart enough," *Frontiers in Neuroscience*, vol. Volume 12 - 2018, 2018.

[65] F. De Arriba-Pérez, M. Caeiro-Rodríguez, and J. M. Santos-Gago, "Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios," *Sensors*, vol. 16, no. 9, 2016.

[66] M. Kheirkhahan, S. Nair, A. Davoudi, P. Rashidi, A. A. Wanigatunga, D. B. Corbett, T. Mendoza, T. M. Manini, and S. Ranka, "A smartwatch-based framework for real-time and online assessment and mobility monitoring," *Journal of Biomedical Informatics*, vol. 89, pp. 29–40, 2019.

[67] X. Gu, A. Taya, Y. Nishiyama, and K. Sezaki, "Toward detecting maternity neurosis by using passive mobile sensing: Preliminary investigation," in *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*. IEEE, 2024, pp. 1–1.

[68] B. Lamichhane, J. Zhou, and A. Sano, "Psychotic relapse prediction in schizophrenia patients using a personalized mobile sensing-based supervised deep learning model," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3246 – 3257, 2023.

[69] T. Liu, S. Wang, Y. Liu, W. Quan, and L. Zhang, "A lightweight neural network framework using linear grouped convolution for human activity recognition on mobile devices," *J. Supercomput.*, vol. 78, no. 5, p. 6696–6716, Apr. 2022.

[70] Y. Zhou, H. Zhao, Y. Huang, T. Riedel, M. Hefenbrock, and M. Beigl, "Tinyhar: A lightweight deep learning model designed for human activity recognition," in *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, ser. ISWC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 89–93.

[71] M. Bock, M. Moeller, and K. Van Laerhoven, "Temporal action localization for inertial-based human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, Nov. 2024.

[72] Y. Enokibori, "rtsfnet: A dnn model with multi-head 3d rotation and time series feature extraction for imu-based human activity recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 4, Nov. 2024.

[73] S. Kaur, A. Gump, Y. Xiao, J. Xin, H. Sharma, N. R. Benway, J. L. Preston, and A. Salekin, "Crop: Context-wise robust static human-sensing personalization," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 2, Jun. 2025.

[74] T. Gong, Y. Kim, A. Orzikulova, Y. Liu, S. J. Hwang, J. Shin, and S.-J. Lee, "Dapper: Label-free performance estimation after personalization for heterogeneous mobile sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 2, Jun. 2023.

[75] G. J. Fernandes, J. Zheng, M. Pedram, C. Romano, F. Shahabi, B. Rothrock, T. Cohen, H. Zhu, T. S. Butani, J. Hester, A. K. Katsaggelos, and N. Alshurafa, "Habitsense: A privacy-aware, ai-enhanced multimodal wearable platform for mhealth applications," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 3, Sep. 2024.

[76] J. F. Huckins, A. W. DaSilva, W. Wang, E. Hedlund, C. Rogers, S. K. Nepal, J. Wu, M. Obuchi, E. I. Murphy, M. L. Meyer, D. D. Wagner, P. E. Holtzheimer, and A. T. Campbell, "Mental health and behavior of college students during the early phases of the covid-19 pandemic: Longitudinal smartphone and ecological momentary assessment study," *Journal of medical Internet research*, vol. 22, no. 6, p. e20185, 2020.

[77] S. Nepal, W. Wang, V. Vojdanovski, J. F. Huckins, A. Dasilva, M. Meyer, and A. Campbell, "Covid student study: A year in the life of college students during the covid-19 pandemic through the lens of mobile phone sensing," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–19.

[78] M. S. H. Aung, F. Alquaddoomi, C.-K. Hsieh, M. Rabbi, L. Yang, J. P. Pollak, D. Estrin, and T. Choudhury, "Leveraging multi-modal sensing for mobile health: a case review in chronic pain," *IEEE journal of selected topics in signal processing*, vol. 10, no. 5, pp. 962–974, 2016.

[79] L. Canzian and M. Musolesi, "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1293–1304.

[80] A. Sano, S. Taylor, A. W. McHill, A. J. Phillips, L. K. Barger, E. Klerman, and R. Picard, "Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study," *J Med Internet Res*, vol. 20, no. 6, p. e210, Jun 2018.

[81] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," *Annual Review of Clinical Psychology*, vol. 13, no. Volume 13, 2017, pp. 23–47, 2017.

[82] M. Rabbi, S. Ali, T. Choudhury, and E. Berke, "Passive and in-situ assessment of mental and physical well-being using mobile sensors," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, ser. UbiComp '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 385–394.

[83] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell, "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health." *Psychiatric rehabilitation journal*, vol. 38, no. 3, p. 218, 2015.

[84] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study," *Journal of medical Internet research*, vol. 17, no. 7, p. e4273, 2015.

[85] M. Cheffena, "Fall detection using smartphone audio features," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1073–1080, 2016.

[86] D. Kelly, B. Smyth, and B. Caulfield, "Uncovering measurements of social and demographic behavior from smartphone location data," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 188–198, 2013.

[87] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss 'n' turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 477–486.

[88] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, "Smartwatch-based activity recognition: A machine learning approach," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE Press, 2016, pp. 426–429.

[89] 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 *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, 2014, pp. 3–14.

[90] X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell, J. Mankof, and A. K. Dey, "Leveraging routine behavior and contextually-filtered features for depression detection among college students," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–33, 2019.

[91] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data," in *2016 IEEE wireless health (WH)*. IEEE, 2016, pp. 1–8.

[92] T. Tran and R. Kavuluru, "Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks," *Journal of Biomedical Informatics*, vol. 75, pp. S138–S148, 2017, supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

[93] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[94] T. Umematsu, A. Sano, and R. W. Picard, "Daytime data and lstm can forecast tomorrow's stress, health, and happiness," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2186–2190.

[95] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[96] G. Dong, M. Tang, L. Cai, L. E. Barnes, and M. Boukhechba, "Semi-supervised graph instance transformer for mental health inference," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021, pp. 1221–1228.

[97] M. Khwaja, S. S. Vaid, S. Zannone, G. M. Harari, A. A. Faisal, and A. Matic, "Modeling personality vs. modeling personalidad: In-the-wild mobile data analysis in five countries suggests cultural impact on personality models," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–24, 2019.

[98] D. A. Adler, F. Wang, D. C. Mohr, and T. Choudhury, "Machine learning for passive mental health symptom prediction: Generalization

across different longitudinal mobile sensing studies," *Plos one*, vol. 17, no. 4, p. e0266516, 2022.

[99] Y. Fukazawa, N. Yamamoto, T. Hamatani, K. Ochiai, A. Uchiyama, and K. Ohta, "Smartphone-based mental state estimation: A survey from a machine learning perspective," *Journal of Information Processing*, vol. 28, pp. 16–30, 2020.

[100] K. Yang, B. Tag, C. Wang, Y. Gu, Z. Sarsenbayeva, T. Dingler, G. Wadley, and J. Goncalves, "Survey on emotion sensing using mobile devices," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2678–2696, 2022.

[101] G. Vos, K. Trinh, Z. Sarnyai, and M. R. Azghadi, "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review," *International Journal of Medical Informatics*, vol. 173, p. 105026, 2023.