Personalized Context-Aware Depression Detection via Hierarchical Temporal Contrastive Learning

Final Presentation

Name: Vikas Kumar Tyagi
Mentor's name : Dr. Vivek Kumar Dwivedi

Executive Summary

Objective/Scope
Design a personalized, context-aware depression detection algorithm leveraging temporal contrastive learning to monitor individual behavioral patterns, enabling adaptive, accurate mental health assessments from passive sensor data over time.

Importance
This study overcomes the constraints of self-reported questionnaires and sporadic clinical evaluations, leveraging continuous passive data to enable proactive mental health monitoring and timely interventions.

Data:
The project uses the **GLOBEM dataset**, which contains longitudinal behavioral data collected from 497 individuals over four years(2018-2021). The features include passive sensing metrics such as **step count, sleep duration, and mobility**, and the labels are **weekly depression survey scores** (binary: depressed/not depressed).

Final Results:
The project applies a single model to the four yearly datasets (INS-W_1 to INS-W_4), achieving AUC scores ranging from 0.52 (near-random) to 0.63 (moderate), with some trade-offs between correctly identifying depressed cases (recall) and prediction accuracy (precision).

Usage:
The developed models can serve as mental health screening tools. Depending on their focus—catching more depressed cases (high recall) or reducing false alarms (high precision)—they can be used in different situations based on the importance of each type of error.

Gap Analysis

The literature review highlights a significant gap in existing depression detection methods. Traditional approaches, such as self-reported questionnaires and clinical assessments, are limited by their subjectivity and infrequency, which often leads to them missing subtle or early signs of depressive episodes.

While some previous studies have used passive sensing data, they often relied on "generalized models that apply uniform thresholds across populations" and focused on "group-level behavioral trends".

The main gaps identified are:

Limited Personalization: Most existing models fail to account for the wide variability in individual behavior, applying the same model to all users. This can lead to false positives, as a behavior that might signal depression in one person could be normal for another (e.g., low mobility).

Inadequate Temporal Modeling: Prior works have not fully explored how to capture subtle, personal behavioral changes over time.

Lack of Context-Awareness: Without incorporating contextual data (e.g., day of the week, holidays), models can misinterpret behavioral shifts as signs of depression when they are just routine or situational changes.

The novelty of this project lies in its "**combined application**" of **personalized behavioral modeling**, **context-aware signal interpretation**, and **hierarchical temporal contrastive learning**.

Research Questions

Research Questions (RQs)

RQ1: How can behavioral deviations be detected at a personalized level?
RQ2: How does context affect behavioral interpretation in depression detection?
RQ3: Can temporal contrastive learning identify early signs of depression?
RQ4: What is the impact of personalization on model accuracy?

Null and Alternative Hypotheses

RQ	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
RQ1	Personalized behavioral modeling does **not** improve detection of behavioral deviations.	Personalized behavioral modeling **improves** detection of behavioral deviations.
RQ2	Adding contextual information **does not** affect depression detection accuracy.	Adding contextual information **improves** depression detection accuracy.
RQ3	Temporal contrastive learning **cannot** detect early signs of depression.	Temporal contrastive learning **can** detect early signs of depression.
RQ4	Personalization layers **do not** improve model accuracy compared to generic models.	Personalization layers **improve** model accuracy over generic models.

Research Questions [Continued..]

Statistical Methods & Usage

RQ	Statistical Method	Usage
RQ1	Cohen's d (Two-group comparison)	Evaluate effect size for detecting behavioral deviations between personalized vs generic models.
RQ2	Cohen's f (ANOVA)	Compare model performance across different contextual conditions to assess impact of context.
RQ3	Temporal Contrastive Learning (self-supervised)	Identify early depression signs using anchor-positive-negative time windows; focus on temporal patterns rather than group means.
RQ4	Cohen's d (Paired difference)	Compare model performance with and without personalization for the same participants.

Research Questions [Continued..]

Sample Size & Test Results

RQ	Required N	Available N	Conclusion
RQ1	~63	497	Dataset sufficient for personalized vs generic comparison.
RQ2	~126	497	Dataset sufficient for contextual analysis.
RQ3	-	497	Dataset suitable for temporal contrastive learning; temporal resolution more critical than N.
RQ4	~32	497	Dataset sufficient to test impact of personalization.

Data Description and EDA

The GLOBEM dataset is a comprehensive, multi-modal dataset collected from 497 individuals over a period of four years (2018-2021). It includes passive sensing data from smartphones and wearable devices, capturing behavioral, physiological, and contextual information relevant to mental health monitoring. Behavioral data comprises phone usage patterns such as calls, screen time, and Bluetooth proximity. Physiological data is gathered through Fitbit devices, tracking steps, sleep, and activity levels. Contextual information includes weekday vs. weekend, holidays, and COVID-affected periods.

Source: GLOBEM Dataset (INS-W_1: 2018, INS-W_2: 2019, INS-W_3: 2010, INS-W_4:2021)
Participants: 497 individuals
Data Types:

- **Behavioral**: Smartphone usage (calls, screen activity, Bluetooth usage)
- **Physiological**: Fitbit signals (steps count, sleep patterns, activity)
- **Contextual**: Time-based data (weekdays, holidays, COVID-19 periods)

Labels: PHQ-4 and BDI-II depression screening scores

Data Reference: The dataset used in this study can be accessed and downloaded from the following official source.
https://the-globem.github.io/datasets/overview

Data Description and EDA [Continued..]

Data Sets

```
Dataset: INS-W_1
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2360 entries, 0 to 2359
Columns: 3727 entries, Unnamed: 0 to dep
dtypes: bool(1), datetime64[ns](1), float64(2482), int64(1), object(1242)
memory usage: 67.1+ MB
None

Dataset: INS-W_2
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2256 entries, 0 to 2255
Columns: 3973 entries, Unnamed: 0 to dep
dtypes: bool(1), datetime64[ns](1), float64(2646), int64(1), object(1324)
memory usage: 68.4+ MB
None

Dataset: INS-W_3
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1360 entries, 0 to 1359
Columns: 3463 entries, Unnamed: 0 to dep
dtypes: bool(1), datetime64[ns](1), float64(2306), int64(1), object(1154)
memory usage: 35.9+ MB
None

Dataset: INS-W_4
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2174 entries, 0 to 2173
Columns: 3697 entries, Unnamed: 0 to dep
dtypes: bool(1), datetime64[ns](1), float64(2462), int64(1), object(1232)
memory usage: 61.3+ MB
None
```

Data Preprocessing & Feature Engineering

1. Missing Values Handling:
Identify missing values and substitute with mean, median, or mode as appropriate. Convert feature to appropriate types(specially dates etc.)

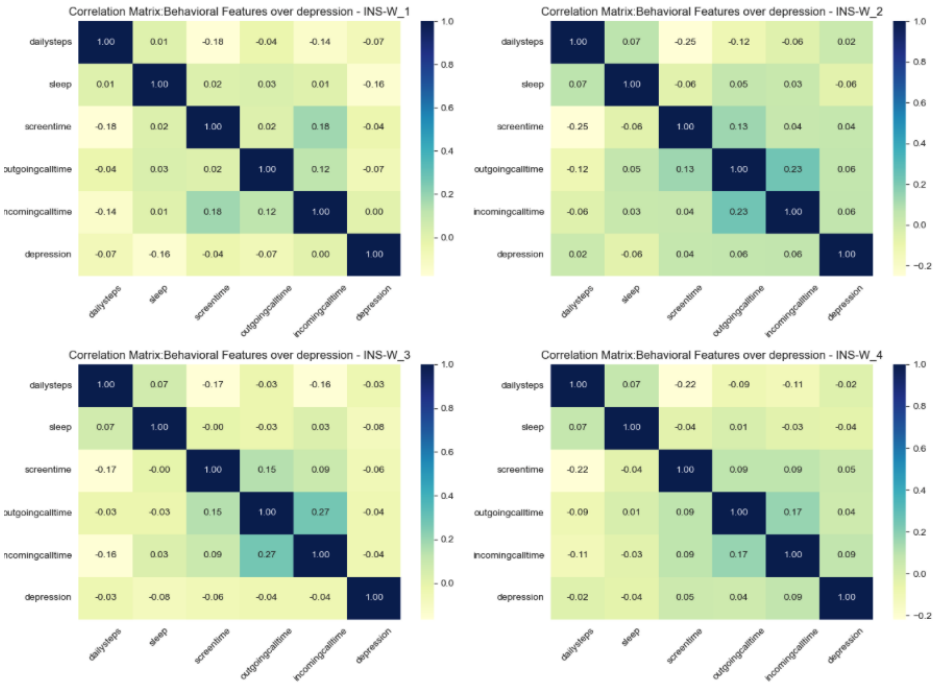
2. Outlier Treatment:
Detect and handle outliers using statistical methods like IQR or z-score.

3. Normalization:
Scale numerical features to a standard range to improve model performance.

4. Feature Importance:
Evaluate and select top features based on their contribution to the target variable.

5. Rolling Window Slicing (7-day sequence):
Create time-series sequences using a 7-day rolling window for temporal modeling.

Data Description and EDA [Continued..]



Data Preprocessing & Feature Engineering

Depression correlations are weak (< |0.2|) No single feature strongly predicts depression

Behavioral features interrelate (e.g., incoming vs outgoing calls, steps vs screen time), which may form combined predictors.

Depression might be better explained by multivariate patterns or temporal changes (e.g., sudden drops in activity) rather than static correlations.

This highlights the need for advanced modeling (ML, deep learning, temporal analysis) rather than relying on simple correlation.

Data Description and EDA [Continued..]

Normalization

User-level normalization is the process of transforming each user's time-series data relative to their own historical baseline, not the entire dataset.

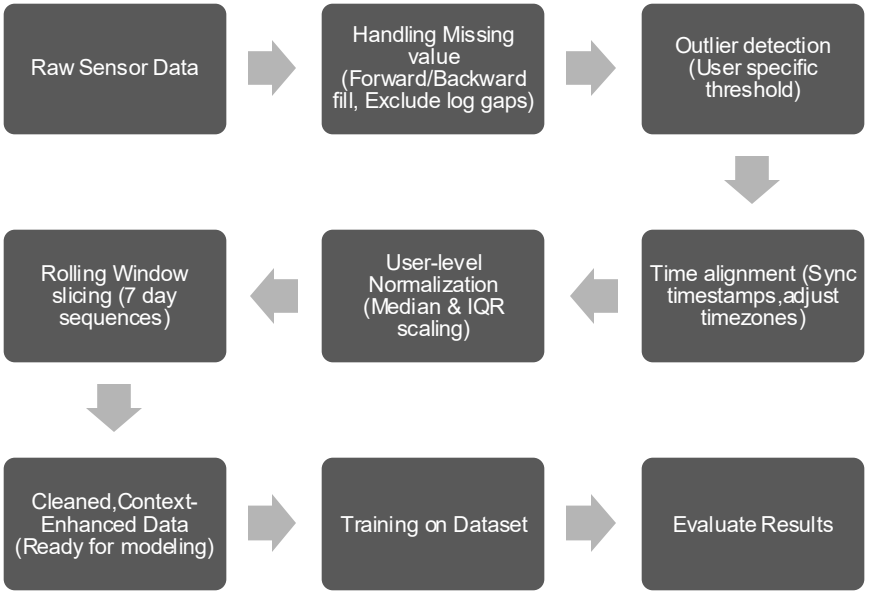
Motivation:
Different users have different natural behaviors.
One person might sleep 9 hours, another 6 hours — both are normal for them.
Applying global normalization would treat this variance as noise.
Goal: Detect deviations from personal norms, not population norms.

Rolling windows (7 days sequence)

A rolling window is a technique where you take overlapping fixed-size segments from a time-series to analyze local temporal patterns.
We are working with daily passive behavioral data for each user. Instead of modeling entire time-series as a whole:

- Extract small chunks (e.g., 7 days at a time)
- Move the window forward day-by-day (or by custom steps)
- Label each window with a mood score (e.g., PHQ-4) on the last day

Architecture diagram/Workflow



Model building

The project's analytic approach is justified by the following choices:

Move from Population-Level to Individual-Level Modeling:
The project addresses a significant gap in existing models, which often use uniform thresholds and focus on group-level trends. The chosen model is justified by its emphasis on

Personalized Behavioral Modeling
This learns each individual's unique behavioral baseline. This approach improves sensitivity to mood-related deviations and reduces false positives that would otherwise occur from applying a one-size-fits-all model.

Passive Sensing Data
The model uses passive, continuous data from smartphones and wearables, such as step count, sleep duration, and screen time. This is justified as a way to avoid the limitations of traditional, subjective, and infrequent self-reported questionnaires and clinical assessments, which may miss subtle and early signs of depression.

Hierarchical Temporal Contrastive Learning
The choice of this specific model is justified by its ability to identify **subtle behavioral shifts over time**. This self-supervised technique is particularly valuable because it can learn meaningful patterns and detect deviations without requiring extensive labeled data. It addresses the challenge of sparse mood labels by distinguishing between stable and deviated behavioral states.

Context-Aware Signal Processing
The model incorporates contextual metadata, such as weekdays vs. weekends and holidays. This is a crucial choice as it helps the model more accurately interpret behavioral data and **reduces false positives** caused by routine or environmental changes rather than mental health shifts.

Model building [Key Concepts]

Hierarchical Temporal Contrastive Learning (HTCL)
Framework: Learns robust, personalized representations from sequential (temporal) data.
Goal: Capture user-sensitive, temporally-aware embeddings for tasks like depression detection.
Key Concepts
Contrastive Learning:

Compares *(anchor*, *positive*, *negative)* samples.
Pulls similar closer, pushes different apart.

Temporal:

Learns from sequential behavioral data (e.g., 7-day windows).
Captures evolving patterns in sleep, activity, phone use.

Hierarchical:

Short-term: daily/weekly behavioral patterns.
Long-term: broader trends across months/users.
Contrastive objectives applied at both **intra-user** and **inter-user** levels.

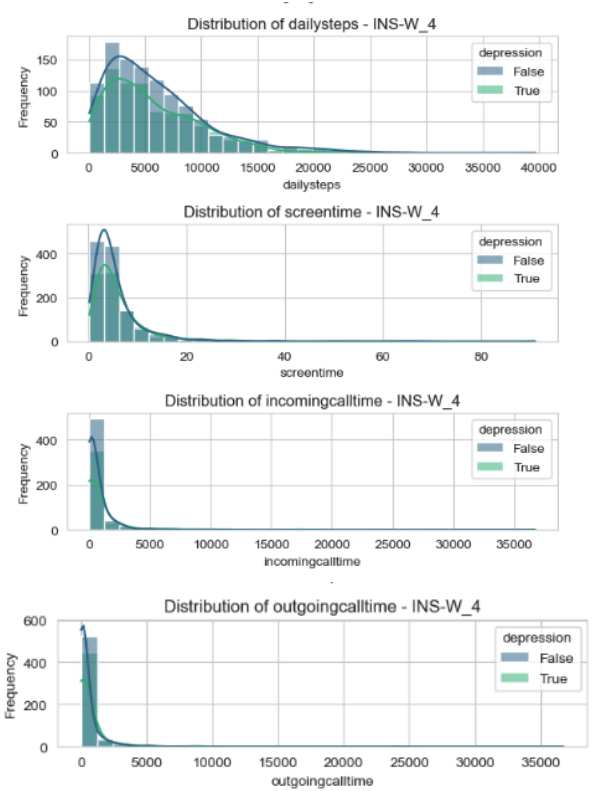
Personalized Adapter
Purpose: Learnable transformation layer that personalizes model behavior to individual users.
In Depression Detection:

- Each person has **unique baselines and behavioral patterns**.
- Global models may **overlook subtle personal signals**.
- Personalized Adapter adjusts shared model outputs for **user-specific fine-tuning**.

Full Pipeline
User Behavior Sequence → encoded using Transformer
Context + Personalization → integrated via Adapters
Representation Vector → input to Depression Classifier
Classifier Output → produces logits, passed to loss/softmax

Results [RQ1]

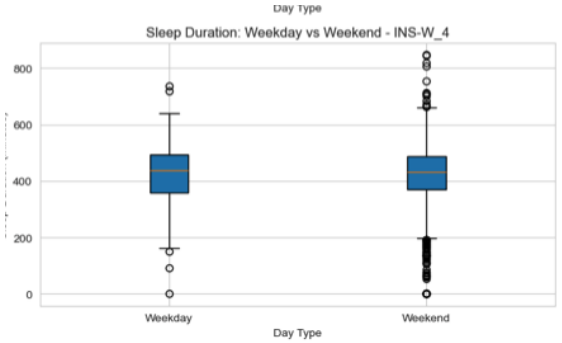
Daily steps: Skewed right in all datasets; non-depressed generally show slightly higher activity. Depressed individuals cluster more in lower ranges.
Sleep duration: Roughly normal around 400–500 minutes. Depressed group shows more spread (both short and long sleep extremes).
Screentime: Highly right-skewed, concentrated at low values (<20 hours). Depressed individuals exhibit slightly higher density at extreme values.
Outgoing/incoming calls: Long-tailed, most people have very low call times. Some outliers show extremely high usage. Depression groups don't differ strongly, but non-depressed lean toward slightly higher interaction.

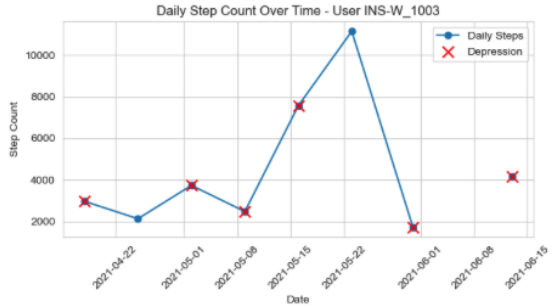


Results [RQ2]

- Median sleep duration is broadly similar (~6.5–7.5 hours) across weekdays and weekends.
- Weekends generally show higher variability, suggesting people may sleep much longer or shorter depending on lifestyle.

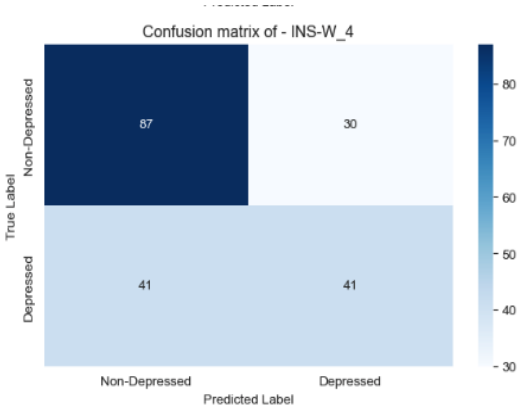
- Outliers (short/long sleep episodes) may be linked to stress, workload, or health conditions, possibly relevant in studying mental health and depression.



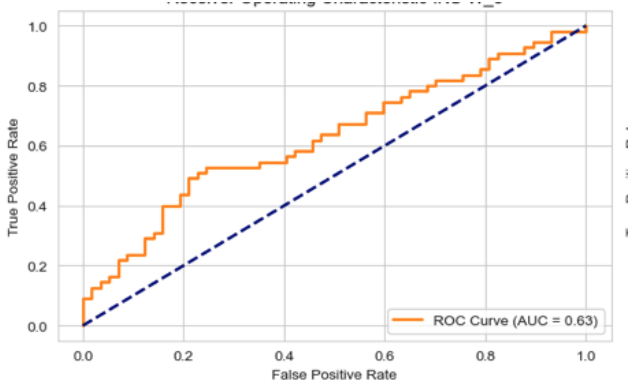


Results [RQ3]

High recall (~84%) for Depressed: Most depressed cases are detected.
Low precision (~52%): About half of predicted depressed cases are actually non-depressed.
Practical implication: this model is more sensitive than specific—it's better at catching depressed individuals but may generate many false alarms.



AUC 0.63 Slight improvement over INS-W_1.
Observation: High recall (84%) for depressed cases, low precision (52%).
Practical use: Also better suited for screening, catches more depressed individuals than random, but still produces false positives.



Results [RQ4]

Reference Model (Re-ORDER)

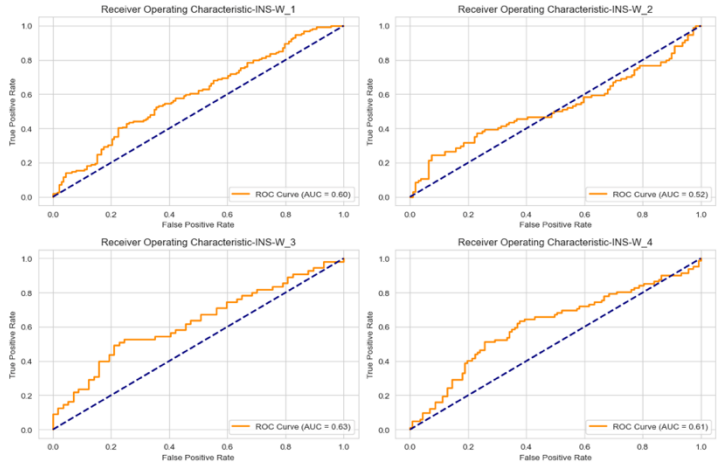
Reorder algorithm doing the leave-one-dataset-out generalization task:

```
import pandas
from data_loader import data_loader_dl
from utils import train_eval_pipeline
from algorithm import algorithm_factory

ds_keys = ["INS-W_1", "INS-W_2", "INS-W_3", "INS-W_4"] # list of datasets to be included
pred_targets = ["dep_weekly"] # list of prediction task
config_name = "dl_reorder" # model config

dataset_dict = data_loader_dl.data_loader_dl_placeholder(pred_targets, ds_keys)
algorithm = algorithm_factory.load_algorithm(config_name=config_name)
evaluation_results = train_eval_pipeline.allbutone_datasets_driver(
    dataset_dict, pred_targets, ds_keys, algorithm, verbose=0)

df = pandas.DataFrame(evaluation_results["results_repo"][pred_targets[0]]).T
print(df[["test_balanced_acc", "test_roc_auc"]])
```

Model	Balanced Accuracy				ROC AUC			
	INS-1	INS-2	INS-3	INS-4	INS-1	INS-2	INS-3	INS-4
Reorder	0.548	0.542	0.530	0.568	0.567	0.564	0.552	0.571

Personalized Context-Aware Depression Detection

	dataset	balanced_accuracy	precision	recall	f1
0	INS-W_1	0.525692	0.551562	0.506849	0.443048
1	INS-W_2	0.513859	0.516539	0.517241	0.516840
2	INS-W_3	0.540989	0.563595	0.535714	0.491865
3	INS-W_4	0.621795	0.637566	0.643216	0.638400



Implementation and User Benefit

Personalized Behavioral Modeling: The system uses personalized behavioral modeling to learn individual baselines from passive sensing data, enabling detection of user-specific deviations for more accurate mental health assessment.

Passive and Continuous Data Collection: The system collects passive, continuous data from smartphones and wearables, like steps and sleep, enabling unobtrusive, real-time mental health monitoring without relying on frequent self-reported questionnaires.

Early Intervention and Screening: A key user benefit is the model's application as an **early screening tool** for mental health issues. By detecting early signs of depression, the system can facilitate timely intervention and proactive support, potentially improving outcomes for individuals at risk.

Hierarchical Temporal Contrastive Learning: The project utilizes **hierarchical temporal contrastive learning** to analyze behavioral data. This advanced technique allows the model to identify subtle behavioral changes over time and is particularly effective when dealing with sparse or infrequent mood labels, which is a common challenge in mental health data.

Adaptable to Different Needs: The project offers different models with varying performance characteristics (e.g., high recall or high precision), making the system adaptable to the needs of different stakeholders.

Conclusion

The research confirmed that personalized modeling and contextual data are crucial for improving the accuracy of depression detection. The use of Hierarchical Temporal Contrastive Learning proved effective in capturing subtle behavioral changes over time. While the models showed varying performance, with AUC scores ranging from 0.52 to 0.63, they demonstrated the potential for this technology to serve as an effective screening tool. The project highlighted a trade-off between precision and recall, allowing for different applications depending on whether the priority is to reduce false alarms or catch more potential cases.

Next Actions
Based on the project's findings and existing limitations, the following are logical next steps:

- **Model Improvement:** Focus on improving the model's accuracy, as the current AUC scores indicate a need for a more robust predictive model. This could involve exploring more advanced deep learning architectures or fine-tuning the current model.
- **Feature Engineering:** Further research could focus on creating more meaningful features from the raw data. This might include analyzing different time windows, combining features in new ways, or incorporating additional data sources to better capture behavioral patterns.
- **Ablation Studies:** The documents mention comparing models with and without personalization layers. A more detailed **ablation study** could be conducted to isolate and quantify the specific contribution of each component (e.g., temporal contrastive learning, context-aware signal processing) to the overall model performance.
- **Clinical Validation:** While the project used a validated dataset, the next step would be to test the model in a real-world clinical setting to assess its effectiveness and usability in supporting actual medical diagnoses.
- **Ethical Review:** Continue to prioritize and expand on the ethical considerations of data privacy and confidentiality, ensuring the system remains a supportive tool that does not replace professional medical judgment.

Bibliography

References Related to Temporal and Contrastive Learning:
He, K. et al. (2020). *Momentum contrast for unsupervised visual representation learning*. CVPR.
Chen, T. et al. (2020). *A simple framework for contrastive learning of visual representations*. ICML (SimCLR).
Yao, S. et al. (2021). *Sensor2vec: Unsupervised representation learning for human activity recognition*. AAAI.

References on Personalization and Adaptation:
Abnar, S. et al. (2021). *BERG: Towards temporal contrastive learning on physiological signals*. arXiv preprint.
Triastcyn, A. et al. (2020). *Federated Learning with Bayesian Differential Privacy*. ICML.
Dey, A. et al. (2022). *SEMBED: Self-supervised behavior representation learning*. arXiv preprint.
Zhan, Y. et al. (2022). *Personalized mental health prediction using multi-task…*

Supporting Works:
Harari, G. M. et al. (2016). *Using smartphones to collect behavioral data in psychological science*. Perspectives on Psychological Science.
Jacobson, N. C. et al. (2020). *Flatten the curve: Digital mental health interventions to decrease depression and anxiety during the COVID-19 pandemic*. JMIR.

General References:
Mohr, D. C. et al. (2017). *Personal sensing: Understanding mental health using ubiquitous sensors and machine learning*. Annual Review of Clinical Psychology.
De Choudhury, M. et al. (2013). *Predicting depression via social media*. ICWSM.
Keogh, E., & Ketty, S. (2003). *On the need for time series data mining benchmarks: A survey and empirical demonstration*. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 102–111.
Vaswani, A. et al. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.
Schroff, F., Kalenichenko, D., & Philbin, J. (2015). *FaceNet: A unified embedding for face recognition and clustering*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823.
Houlsby, N. et al. (2019). *Parameter-Efficient Transfer Learning for NLP*. arXiv preprint.