# Multimodal TinyML for Suicidal Sentiment Detection on Wearable Devices

Samyak Jain, Anubhav Yadav, Aaryan Yadav

*Netaji Subhas University of Technology, Dwarka, Delhi, 110078, India*

**Abstract**

Suicide has a large public health impact.Suicidal sentiments are a serious mental health issue and requires timely intervention. Wearable devices allows for continuous monitoring of suicidal sentiments. This research proposes a TinyML based solution that combines Physiological parameters(heart rate variability, blood glucose, blood pressure, physical activity) along with textual data and tone of speech to detect suicidal sentiments in real time. We suggest a multi modal machine learning model for physiological signal processing, text and speech tone analysis and multi modal fusion. This system also incorporates TinyML models to preserve user privacy and run in a resource constrained environment.

*Keywords:* TinyML, Sentimental analysis of Text, Physiological Parameters, Sentiment using tone of Speech, Real Time Detection

## 1. Introduction

Introduction

Suicide and Major Depressive Disorder (MDD) are pressing global health issues that have far-reaching consequences for individuals, families, and societies[1]. According to the World Health Organization, approximately 700,000 people die by suicide every year, making it the fourth leading cause of death among 15-29-year-olds globally[2]. Moreover, MDD affects more than 264 million people worldwide and is a significant contributor to the overall global burden of disease[3].

Numerous studies have investigated the role of physiological markers in the development and progression of suicidal ideation and MDD7-10. For instance, dysregulation of the hypothalamic-pituitary-adrenal (HPA) axis, which is involved in the stress response, has been linked to an increased risk of suicidal behavior[4]. Additionally, abnormalities in heart rate variability (HRV), a measure of autonomic nervous system function, have been associated with depression severity and suicidal ideation[5]. Sleep disturbances, such as insomnia and nightmares, have also been identified as potential risk factors for suicidal thoughts and behaviors[6].

Despite the growing body of evidence highlighting the importance of physiological factors in suicidal ideation and MDD, the predictive potential of these markers, particularly when combined with textual data analysis, remains largely unexplored. The integration of physiological

*Email addresses:* samyak.jain.ug21@nsut.ac.in (Samyak Jain), anubhav.yadav.ug21@nsut.ac.in (Anubhav Yadav), aaryan.yadav.ug21@nsut.ac.in (Aaryan Yadav)

data with natural language processing techniques could provide a more comprehensive understanding of an individual's mental health status and enable the development of more accurate predictive models[7].

In this study, we propose a novel approach to predict PHQ-9 scores, a widely used measure of depression severity[8], by leveraging a multimodal machine learning model that integrates physiological data and textual analysis. Our model aims to harness the predictive power of various physiological parameters, including HRV and sleep quality metrics, collected over 7-day and 14-day periods. By incorporating textual data alongside these physiological features, we seek to further refine the prediction of suicidal ideation and PHQ-9 scores, capturing both objective and subjective aspects of an individual's mental health.

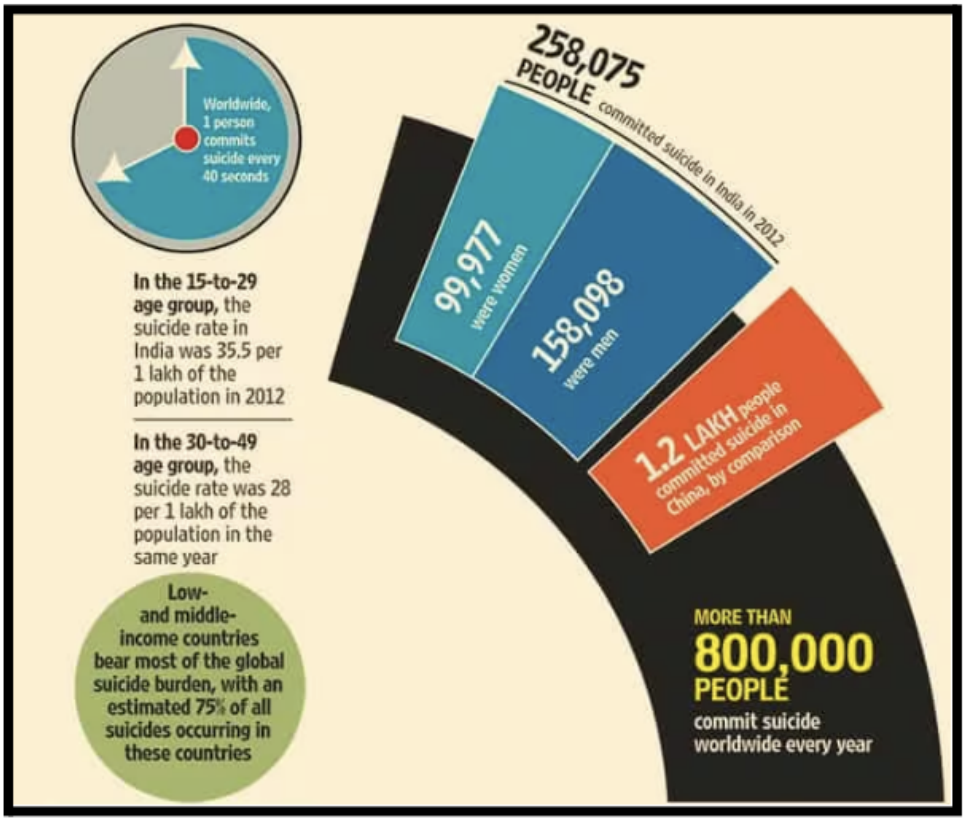



Figure 1: Suicide statistics highlighting the global burden, with over 800,000 deaths annually worldwide, a high rate among youth in India, disparities across age groups, and a disproportionate impact on low- and middle-income countries accounting for 75% of cases.

To ensure the robustness and generalizability of our model, we train it on a diverse dataset encompassing participants from various backgrounds and age groups. This approach seeks to mitigate potential biases and enhance the model's applicability across different populations. By

including a wide range of participants, we aim to develop a model that can accurately predict suicidal ideation and MDD risk in a real-world setting, accounting for the heterogeneity of individuals affected by these mental health conditions.

A key innovation of our study lies in the utilization of TinyML20, a cutting-edge technology that enables the deployment of machine learning models on resource-constrained devices. By employing TinyML, we aim to develop a smartband capable of detecting suicidal sentiment in real-time, without the need for cloud data upload. This approach prioritizes user privacy and allows for the prediction of mental health outcomes in a secure and efficient manner. The use of TinyML also makes our solution more accessible and cost-effective, as it eliminates the need for continuous data transmission and storage on external servers.

The implications of our research are far-reaching, particularly in the context of the mental health sector in India. India accounts for nearly 18% of the global population and has a high prevalence of mental health disorders, including MDD and suicidal behavior[9].By integrating our predictive model into future smartwatches, we envision a financially viable and accessible solution for early detection and intervention in cases of suicidal ideation and MDD. This could potentially revolutionize mental healthcare delivery in India, enabling timely support and treatment for individuals at risk, even in resource-limited settings.

In the following sections, we present a comprehensive literature review highlighting the key physiological factors associated with suicidal ideation and MDD. We then describe our methodology, including the dataset, feature engineering, and machine learning techniques employed. Finally, we discuss the results, limitations, and future directions of our study, emphasizing the potential impact on mental health management and suicide prevention strategies, both in India and globally.

## 2. Related Work

Previous research has demonstrated that monitoring digital biomarkers using machine learning models to assess passive smartphone data can aid in the screening, treatment, and remote monitoring of mental health disorders. Several studies have explored this concept, each contributing to the growing body of knowledge in this field.

Haines et al. [10] developed a mobile app that integrated data from a Fitbit device and a Facebook account to monitor sleep, mood, physical activity, and social engagement of individuals at risk of suicide. They applied machine learning techniques to the collected data and found that the k-nearest neighbors algorithm achieved the highest accuracy (68%) in predicting suicidal ideation. The authors suggested that adding more features, such as heart rate variability and stress levels, and increasing the sample size could improve the performance of their method. This study highlights the potential of combining data from multiple sources, such as wearable devices and social media, to predict mental health outcomes.

Choudhary et al. [11] proposed a machine learning approach for detecting digital behavioral patterns of depression using nonintrusive smartphone data. They introduced a novel metric called the Mental Health Similarity Score, which was derived from analyzing 37 passive smartphone features, such as screen time, app usage, and communication patterns. The researchers collected smartphone data and PHQ-9 scores from 558 Android users and trained four supervised machine learning classifiers to predict depression and its severity. They also explored the effect of adding three gyroscope features to the models. The study reported high accuracy and correlation between the smartphone features and the PHQ-9 scores, demonstrating the feasibility of using

smartphone data for depression screening and monitoring. This research emphasizes the importance of identifying relevant smartphone features and developing novel metrics to assess mental health.

In a study by Wang et al. [12], the app Student Life was used to show the correlation between depression and accelerometer- and screen usebased biomarkers. This research provides evidence for the relationship between smartphone sensor data and mental health outcomes, specifically depression. The findings suggest that passive data collected from smartphones can be used as indicators of an individual's mental well-being.

Saeb et al. [13] found significant correlations between depression and passive data such as phone use and GPS in a sample of 40 participants. This study further supports the idea that smartphone usage patterns and location data can be used to infer an individual's mental health status. The results underscore the potential of leveraging passive smartphone data for mental health monitoring.

Asare et al. [14] found that age group and gender as predictors led to improved machine learning performance. Their study concluded that behavioral markers indicative of depression can be unobtrusively identified using smartphone sensor data [14]. This research highlights the importance of considering demographic factors when developing machine learning models for mental health prediction and the effectiveness of using smartphone sensor data to identify depression-related behavioral markers.

Taking a machine learning approach, a study found that the predictive power of mobile device use patterns was significant to continuously screen for depressive symptoms or monitor ongoing treatments [15]. This research demonstrates the practical application of machine learning techniques in mental health care, specifically for depression screening and treatment monitoring using smartphone data.

In line with this study, another study used the Remote Monitoring Application in Psychiatry to explore the validity of smartphone-based assessments for self-reporting mood symptoms and found high compatibility with nonsmartphone-based assessments [16], thus proving such tools to be helpful for clinicians and research. This research validates the use of smartphone-based assessments for mood symptom reporting and highlights their potential to support mental health professionals and researchers in their work.

## 3. Factors causing Suicide Ideation

### *3.1. Heart Rate Variability*

Heart rate variability (HRV) is a measure of the variation in the time intervals between consecutive heartbeats.In particular, HRV has been linked to suicidal behavior, a major public health concern and a leading cause of death worldwide. HRV is a measure of beat-to-beat temporal changes in heart rate and these changes, rather than being random noise, reflect the output of the central autonomic network.[17]

Several studies have shown that suicidal individuals have lower HRV than nonsuicidal individuals, suggesting a reduced capacity to adapt to environmental challenges and cope with negative emotions. In a study with 1461 patients, the MINI Suicide Score and its relation were studied. Both the resting HR and the root mean square of successive difference (RMSSD), a measure of HRV, had significant associations with the MINI suicidal score. It was found that for every 1 bpm increase in HR, the MINI suicidal score increased by 0.011 points and for every 1 ms increase in RMSSD, the MINI suicidal score decreased by 0.014 points.[18]

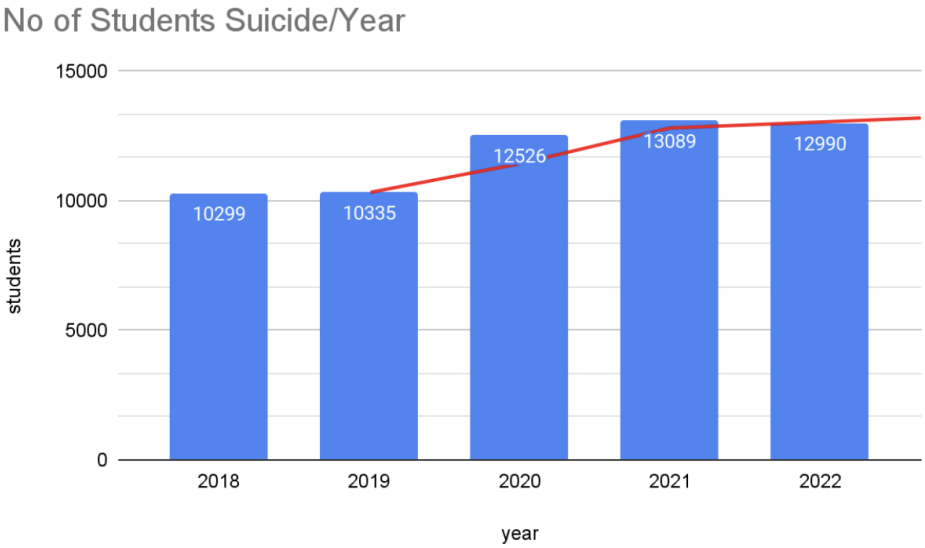



Figure 2: Number of Students committing suicide in a year. Vertical axis represents number of students in a scale of 5000 and the horizontal axis represents year.

Resting Heart Rate and HRV also have a strong correlation with Major Depressive Disorder. In a study with 72 patients MDD patients with melancholia displayed significantly increased heart rate and lower resting-state HRV relative to controls, findings associated with a moderate effect size (Cohens ds = 0.560.58). Patients with melancholia also displayed an increased heart rate relative to those with non-melancholia (Cohens d = 0.20). [19]

Now talking about the technologies and sensors used to measure HRV in wearable devices. In recent years, the utilization of Photoplethysmography (PPG) has emerged as a cost-effective and accessible method for obtaining Heart Rate Variability (HRV) indices. PPG-based devices are known for their simplicity and widespread availability. They offer an alternative approach to traditional HRV measurement tools. These devices are equipped with sensors employing an infrared emitter and a detector, allowing for the non-invasive monitoring of blood volume changes, and subsequently, heart rate variability. So far wearable devices can only be used as a surrogate for HRV at resting or mild exercise conditions, as their accuracy fades out with increasing exercise load. [20] During rest, the MAE (±standard deviation (SD)) of consumer wearables was an average of 7.2 ± 5.4 bpm and the MAE of research-grade wearables was 13.9 ± 7.8 bpm ($p < 0.0125$). During physical activity, the MAE ± SD of consumer wearables was 10.2 ± 7.5 bpm and the MAE of research-grade wearables was 15.9 ± 8.1 bpm ($p < 0.0125$). [21] On searching we a found a Machine Learning-Empowered System that monitors HRV and blood pressure with ear PPG. Applying the proposed machine learning-based signal processing framework to an acquired ear signal dataset, the MESTD, MAE and RMSE for HR estimation are 0.82.7, 1.8 and 2.8 BPM, respectively. [22]

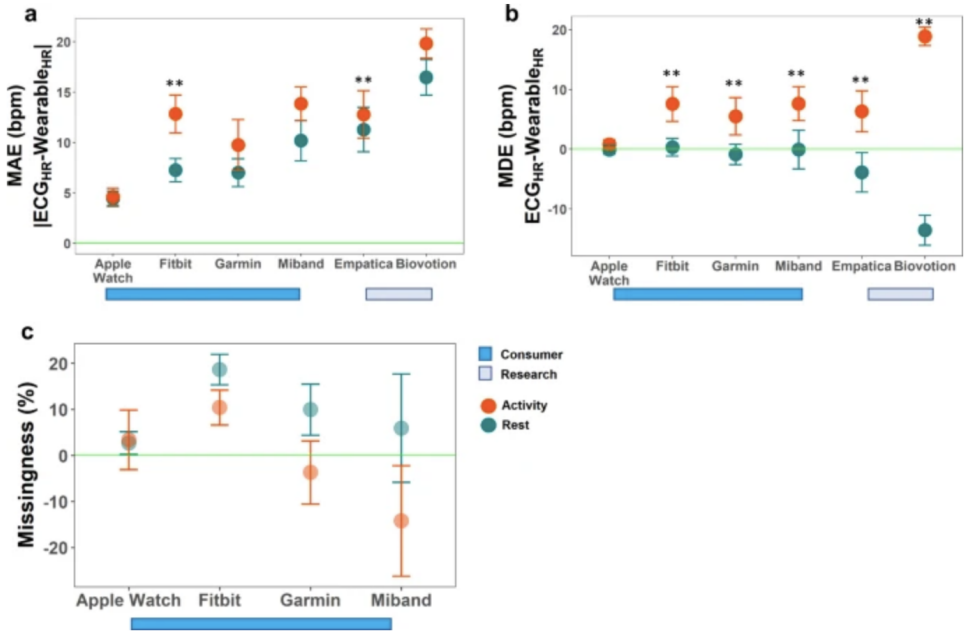



Figure 3: Comparative analysis of the accuracy and performance of popular wearable devices in estimating energy expenditure and classifying activity states. (a) Mean absolute error (MAE) in kcal/hour for estimating energy expenditure, with Apple Watch and Fitbit exhibiting the lowest errors across devices. (b) Distribution of energy expenditure estimation errors, revealing an underestimation bias for most devices except Apple Watch. (c) Percentage of incorrect classifications (missingress) in detecting activity and rest states across consumer, research, activity, and rest modes, with Apple Watch and Fitbit demonstrating superior performance in distinguishing activity levels. Lower values indicate higher accuracy in the respective metrics.

### *3.2. Sleep*

In a study investigating the link between sleep disturbances and suicidality in individuals who attempted suicide, an astonishing 89% of the participants reported encountering various forms of sleep disturbances. The most common complaint was difficulties initiating sleep (73%). Other complaints included difficulties maintaining sleep (69%), nightmares (66%) and early morning awakening (58%). [23] This relationship remained after adjustment for psychiatric diagnosis and psychiatric symptom intensity. On deeper examination of specific parameters of sleep, it revealed that there was a significant negative correlation between the BDI suicidality score and REM latency. REM percent was positively correlated with the BDI suicidality score. [24]

In summary, evidence suggests that suicidal ideation and behaviors are closely associated with sleep disturbances, and in some cases, this association appears to exist above and beyond depression. [25]

Now taking a look on the current methods and technologies being used to measure sleep stages we found a Long Short Term Memory(LSTM) neural network model using HRV data. The method presented in that study performs at a level that advances the state-of-the-art for HRV-based sleep stage classification.

Performance for each of the sleep stages was also measured (precision, recall, accuracy, Cohens k). For wake, precision was 0.73 ± 0.20, recall was 0.71 ± 0.20, accuracy was 0.90 ± 0.07 and Cohens k was 0.63 ± 0.19. For REM, precision was 0.71 ± 0.22, recall was 0.76 ± 0.24, accuracy was 0.92 ± 0.04 and Cohens k was 0.68 ± 0.22. For combined N1/N2, precision was 0.80 ± 0.11, recall was 0.82 ± 0.08, accuracy was 0.79 ± 0.08 and Cohens k was 0.56 ± 0.15. Finally, for N3, precision was 0.62 ± 0.33, recall was 0.61 ± 0.30, accuracy was 0.92 ± 0.04 and Cohens k was 0.53 ± 0.27. [26]

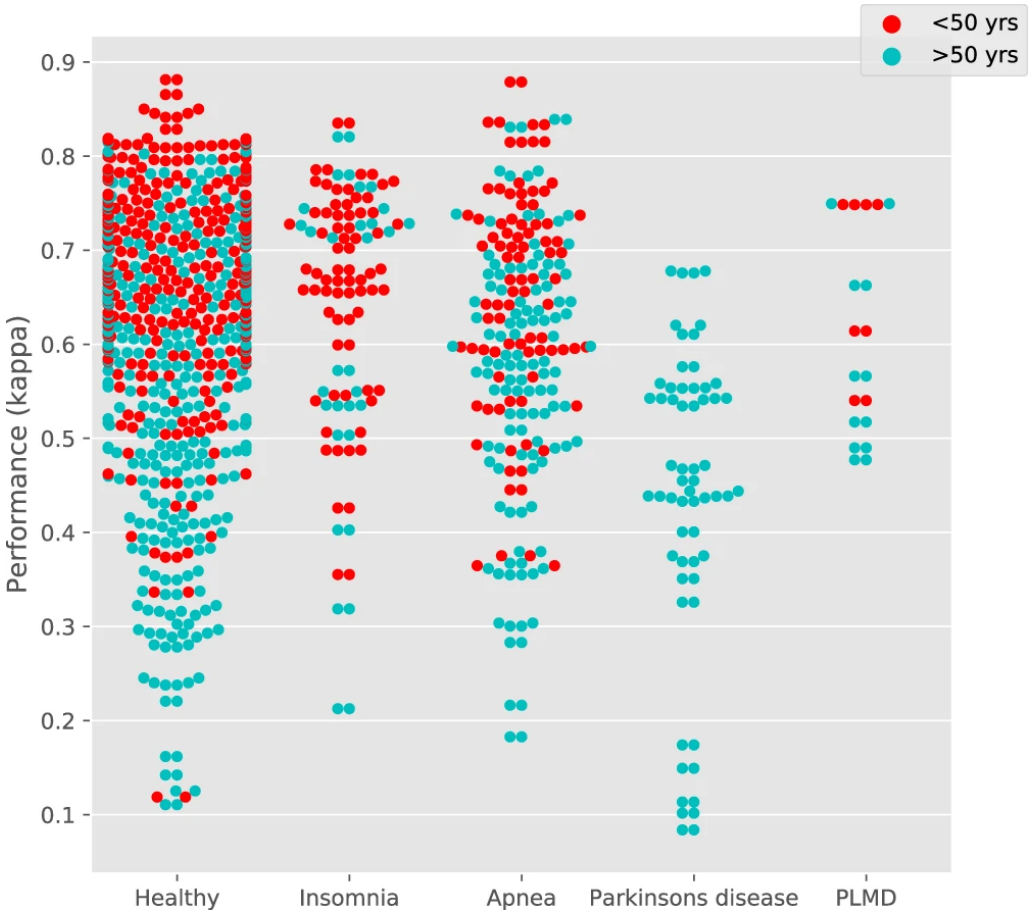
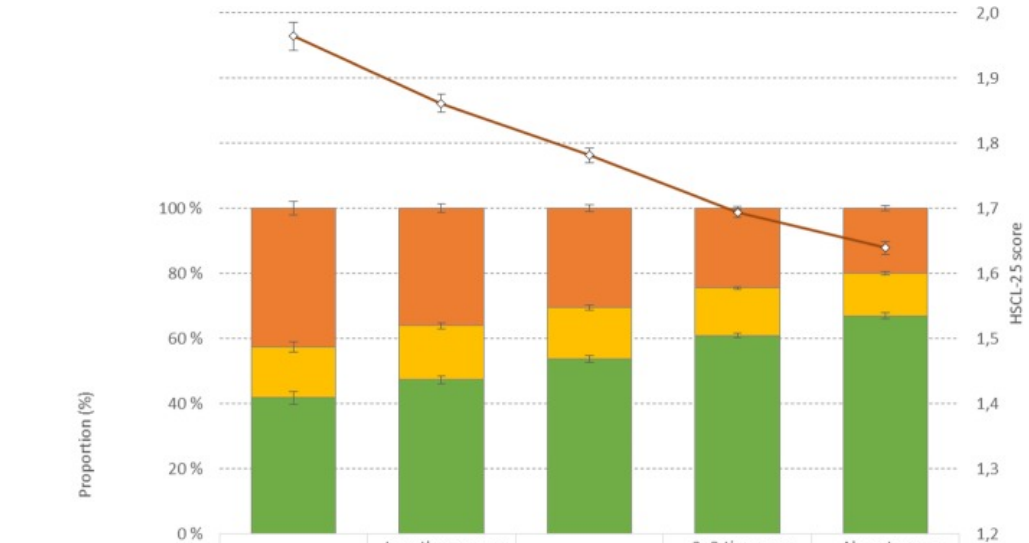



Figure 4: A scatter plot visualizes the performance comparison between two age groups (< 50 years and > 50 years) under various conditions. The y-axis represents a performance metric, *y*, ranging from 0.1 to 0.9, with higher values indicating better performance.Healthy controls serve as the baseline condition. In this case, the younger group (marked with red dots) generally exhibits higher scores with a wider distribution compared to the older group (marked with teal dots).

### 3.3. Stress

In a noteworthy study conducted by [A. H. Farabaugh et al., 2004], the impact of perceived stress on specific manifestations of depression was investigated, with a focus on anger attacks and atypical depression. Upon analysis, the study revealed a between higher level of stress and also found a presence of anger attacks. After adjusting for age, gender, and severity of depression at baseline, higher levels of perceived stress were significantly related to the presence of anger attacks (P<0.0001; t=-4.103) as well as to atypical depression (P=0.0013; t=3.26). [27]

### 3.4. Physical Activity




| | Never | Less than once a week | Once a week | 2–3 times per week | Almost every day |
|---|---|---|---|---|---|
| High (≥2.0) | 42,6 % | 36,1 % | 30,6 % | 24,6 % | 20,0 % |
| Moderate (≥1.75 <2.0) | 15,7 % | 16,7 % | 15,7 % | 14,5 % | 13,0 % |
| Low (<1.75) | 41,7 % | 47,3 % | 53,7 % | 60,9 % | 66,9 % |
| Average HSCL-25 score | 1,964 | 1,861 | 1,781 | 1,693 | 1,639 |

Figure 5: he image depicts a stacked bar chart showing the proportions of a population categorized into high, moderate, and low groups based on some criteria or score, across different frequencies of an activity or behavior. The x-axis represents the frequency, ranging from ”Never” to ”Almost every day”. The y-axis on the left shows the percentage breakdown, while the y-axis on the right displays an average numerical score, which decreases as the frequency increases. The chart suggests an inverse relationship between the frequency of the physical exercise and the HSCL-25 score.

Physical Activity has shown a strong correlation to suicide ideation and overall mental health of an individual. In a study monitoring 50.054 students aged between 18 - 35, physical exercise was found to be negatively associated with all measures of mental health problems and suicidality in a dose-response manner.The strongest effect-sizes were observed for frequency of physical exercise. Women with low levels of physical activity had a near three-fold increased odds of both

scoring high on the HSCL-25, and self-reported depression, compared to women exercising almost every day. Even stronger effect-sizes were observed for men (ORs ranging from 3.5 to 4.8). Also, physical exercise duration and intensity were significantly associated with mental health problems, but with generally smaller ORs. Similarly, graded associations were also observed when examining the link to self-harm and suicide attempts (ORs ranging from 1.9 to 2.5)[28]

Another study analysed data from 1021 participants concluded that suicide attempts were significantly reduced in participants randomized to exercise interventions as compared to inactive controls (OR = 0.23, CI 0.090.67, p = 0.04, k = 2). [29]

### *3.5. Blood Glucose*

Recent research has brought attention to a potential link between glucose disturbances and the prevalence of suicide attempts.A comprehensive investigation conducted by [Chen et al., 2022] shed light on this association, revealing a striking contrast in the rates of suicide attempts among MDD patients with and without glucose disturbances. The prevalence of suicide attempts was higher in MDD patients with glucose disturbances (n = 83.35.47%) than that in MDD patients without glucose disturbances (n = 263, 17.63%) ($\chi^2$ = 39.585, p < 0.001).

### *3.6. Text*

Recent advancements in technology, particularly in the realm of Natural Language Processing (NLP), offer promising avenues for the early detection of individuals at risk of suicide ideation. A notable research study highlighted that NLP could help in the early detection of individuals who have suicide ideation and allow timely implementation of preventive measures. It is also found that passive surveillance via mobile applications, online activity, and social media is feasible and may help in the early diagnosis and prevention of suicide in vulnerable groups.[30]

One of the papers proposed a novel method of detecting and preventing suicide in online text-based counseling services using natural language processing (NLP) and domain knowledge. The authors developed a domain knowledge-aware risk assessment (KARA) model that incorporates a suicide-knowledge graph into a deep learning model to identify help-seekers who are in crisis.More importantly,the KARA model achieved a much higher recall for positive cases than the baseline (0.870 vs 0.791). This encouraging result indicated that KARA reflected 87% of the high-risk cases, 10% more than the baseline BiLSTM model.[31]

Some research authors used a bidirectional long short-term memory (BiLSTM) model, which can capture both past and future contexts, and five ML models, such as logistic regression and support vector machine. They collected and pre-processed a dataset of 49,178 tweets using various natural language processing (NLP) techniques. They evaluated their models on accuracy, precision, recall, and F1-score. They found that the BiLSTM model performed the best, followed by the random forest model. They concluded that ML and DL models can offer effective and low-cost ways of identifying and preventing suicide in online platforms. The BiLSTM model surpasses the other ML and DL models with an accuracy of 93.6%.[32]

## 4. Objectives

This study focuses on the following things:

1. Collecting data from various sensors such as PPG sensor, using gyroscopes and accelerometer to collect sleep data and physical activity data. We then collect the heart rate variability and text data of a person. We then perform correlation analysis between HRV data and other physiological factors.

2. After preprocessing the data set and combining all the physiological factors and text of a person in a single dataset, we perform regression analysis to predict the PHQ - 9 value of a patient.

### *4.1. Abbreviations and Acronyms*

HRV - Heart Rate Variability, ECG - Electrocardiogram, PPG - Photoplethysmography, MDD - Major Depressive Disorder, REM - Rapid Eye Movement, BDI - Beck's Deppresion Index, PHQ-9 - Patient Health Questionaire

### *4.2. Dataset*

In this study, we utilize data from a large-scale longitudinal research project focused on mental health and lifestyle factors. The project, conducted between 2018 and 2020, involved over 10,000 participants in the United States who provided data from consumer-grade wearable devices and completed regular surveys related to their mental well-being and lifestyle changes. The study design and baseline participant characteristics have been previously described in detail.

The data subset employed in our work includes the following components:

1. **Wearable person-generated health data (PGHD):** This dataset comprises step count and sleep information collected from the participants' Fitbit devices throughout the study period.
2. **Screener survey:** Prior to the commencement of the study, participants self-reported their socio-demographic information and any pre-existing comorbidities.
3. **Lifestyle and medication changes (LMC) survey:** On a monthly basis, participants were asked to complete a brief survey documenting any changes in their lifestyle and medication over the preceding month.
4. **Patient Health Questionnaire (PHQ-9) score:** Every quarter, participants were requested to complete the PHQ-9, a well-established and validated 9-item questionnaire used to assess the severity of depressive symptoms [8].

From these diverse data sources, we derive a comprehensive set of input features, categorized as either static or dynamic. Static features, such as demographic variables, are defined once and remain constant for all samples from a given participant throughout the study. In contrast, dynamic features, such as behavioral metrics derived from the wearable devices, vary over time for each participant. The detailed process of feature extraction and engineering is elaborated upon in the Supplementary Materials.

#### *4.2.1. Survey Data*

Survey data was collected at multiple stages of the study, with extensive surveys delivered before the start and at the end of the study (pre/post surveys), along with short weekly Ecological Momentary Assessment (EMA) surveys during the study to collect in-the-moment self-report data.

The depression detection task was used as a starting point for behavior modeling, with BDI-II (post) and PHQ-4 (EMA) employed as the ground truth. Both were screening tools for further inquiry of clinical depression diagnosis. A binary classification problem was focused on to distinguish whether participants scores indicated at least mild depressive symptoms through the scales (i.e., PHQ-4 > 2, BDI-II > 13). The average number of depression labels was 11.6  2.6 per person. The percentage of participants with at least mild depression was 39.8  2.7% for BDI-II and 46.2  2.5% for PHQ-4.

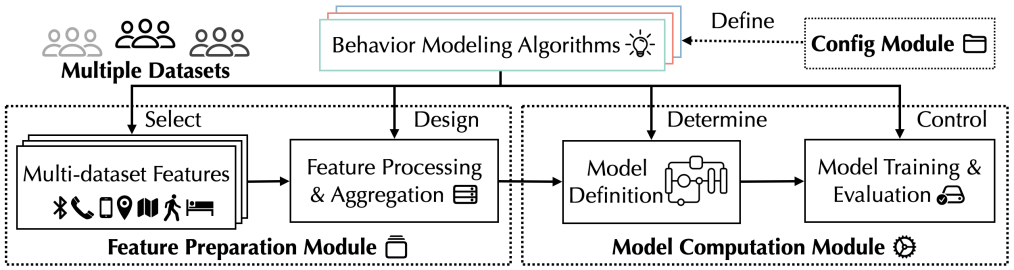



Figure 6: Participants are subjected to periodic surveys and are tracked upon various physiological parameters such as heart rate, sleep, phone usage, and other factors.

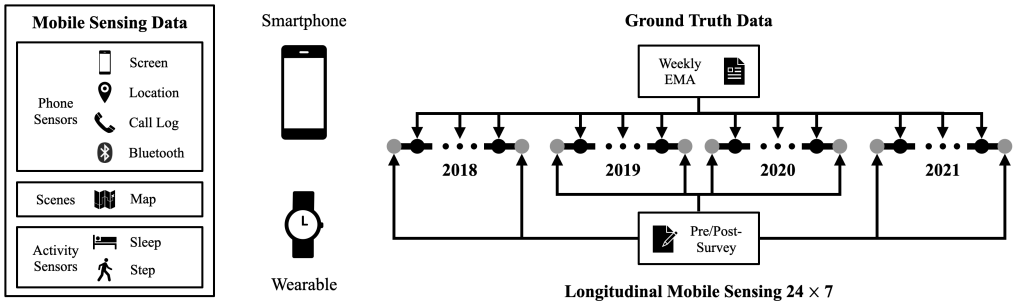



Figure 7: Dataset

#### *4.2.2. Sensor Data*

A mobile app was developed using the AWARE Framework to continuously collect location, phone usage (screen status), Bluetooth scans, and call logs. The app was made compatible with both the iOS and Android platforms. Participants were instructed to install the app on smartphones and leave it running in the background. Additionally, wearable Fitbits were provided to collect their physical activities and sleep behaviors. Sensor data was passively collected 247 during the study through the mobile app and wearable. The average number of days per person per year was 77.5  8.9 among the four datasets.

## 5. Methodology

### *5.1. Phase-1*

The objective of the phase 1c model is to anticipate the PHQ-9 score categories of participants based on their sociodemographic, medical, and wearable PGHD.

In a previous study by Makhmutova et al [33], an initial version of Phase 1c was introduced. However, we present an enhanced variant that has been modified to reduce overfitting. To achieve this, we employ the CatBoost algorithm with Dropouts Meet Multiple Additive Regression Trees (DART) boosting [34], an ensemble model of boosted regression trees with dropout.

This particular algorithm is chosen for its capability to handle sparse data and its ability to adjust a dropout parameter to mitigate overfitting. The process of feature selection eliminates highly correlated features, utilizing recursive feature elimination [35] to exclude features with

lesser contributions to the model's performance. We primarily assess the model's performance using quadratic weighted Cohen's Kappa [36], while adjacent accuracy, balanced accuracy, and weighted F1-scores serve as secondary performance metrics. To optimize the hyperparameters of our LightGBM model, we conduct randomized search 5-fold cross validation.

The reason for employing a 5-fold cross validation is to minimize the influence of overfitting. The performance metrics of the best tuned models, along with 95% confidence intervals, are reported across 5 training runs (5 outer shuffle splits). Additional information on hyperparameters can be found in the Supplementary Materials and elsewhere [37]. Given the expansive feature space encompassing various static and dynamic input features, we construct the model in three steps.

First, we extensively explore and identify the best feature subsets for each type of input. Next, we perform an initial optimization on input sets that combine different types of input, taking into account an initial estimation of model error. Lastly, we conduct a final tuning process to obtain the best performing model. The outcome of phase 1 produces intermediate monthly PHQ-9 score categories for SM1, representing sample month 1, and SM2.

### *5.2. Phase-2*

In phase 2c, we predicted an increase in PHQ-9 category using various inputs, including the participant's PHQ-9 score from SM0, intermediate generated PHQ-9 categories at SM1 and SM2, the generated probabilities of each PHQ-9 category for SM1 and SM2, LMC survey responses, and wearable PGHD collected over the 2 weeks prior to the final PHQ-9 completion at SM3. We also used screener survey responses as input features to account for sociodemographic factors. To calculate the target variable in each sample in the phase 2c model, we observed whether there was an increase in PHQ-9 category between SM0 and SM3.

The model construction procedure for phase 2c was similar to that of phase 1c. The feature selection process involved reducing the initial number of input features by removing highly correlated features and selecting the most important features using recursive feature elimination with cross-validation for the largest sets of input features, grouped by source. We then used forward sequential feature selection, a greedy method that has been successfully used to develop digital measures in mental health studies, to identify the optimal features. LightGBM DART was chosen as the algorithm due to its high accuracy in similar classification tasks, ability to handle sparse data, and generation of interpretable models.

We prioritized Specificity and Area Under the Precision-Recall Curve (AUPRC) as performance metrics. Feature importance was evaluated using a combination of 'Gain' importance and 'split' importance. Gain importance measures the improvement in accuracy provided by a feature, while split importance considers the number of times the feature is used in a model. Together, these metrics help us understand which features contribute the most to the model's decision-making process.

Figure 3 summarizes the construction of the PSYCHE-D combined pipeline, consisting of phase 1c followed by phase 2c. The diagram also illustrates the participant-based splitting approach used to ensure that predictions are generated on previously unseen participants, allowing us to evaluate the approach's generalization capabilities.

### *5.3. Combining Text Data with Phsyiological Data*

We leverage a pretrained BERT (Bidirectional Encoder Representations from Transformers) model, which has been fine-tuned on a Reddit dataset, to detect suicidal sentiment in text data.

The BERT model, a state-of-the-art transformer-based architecture, has demonstrated exceptional performance in various natural language processing tasks, including sentiment analysis and text classification. By fine-tuning the pretrained BERT model on the Reddit dataset, which contains a significant amount of mental health-related content, we aim to enhance its ability to capture the subtle nuances and contextual information specific to suicidal sentiment.

To further improve the accuracy and robustness of our suicide risk assessment, we propose a novel approach that combines the results obtained from the fine-tuned BERT model with physiological parameters acquired from a wearable device. The integration of textual and physiological data is achieved through a multi-modal fusion technique, which allows for the effective combination of heterogeneous data sources.

The textual data, processed by the fine-tuned BERT model, undergoes a series of preprocessing steps, including tokenization, lowercasing, and removal of stop words and special characters. The preprocessed text is then passed through the BERT model, which generates high-dimensional embeddings that capture the semantic and contextual information present in the text. These embeddings are subsequently fed into a dense neural network layer, which learns to map the embeddings to a lower-dimensional representation suitable for fusion with the physiological data.

Concurrently, the physiological parameters collected from the wearable device, such as heart rate variability, sleep patterns, and physical activity levels, are preprocessed and normalized to ensure compatibility with the textual data. These physiological features are then concatenated with the lower-dimensional representation of the textual embeddings, forming a unified multi-modal representation.

The fused multi-modal representation is then passed through a series of fully connected layers, which learn to capture the complex interactions and correlations between the textual and physiological data. The output of the final fully connected layer is connected to a regression head, which is trained to predict the PHQ-9 (Patient Health Questionnaire-9) score of the individual.

To train the model, we employ a mean squared error (MSE) loss function, which minimizes the difference between the predicted PHQ-9 scores and the ground truth scores obtained from clinical assessments. The model is trained using the Adam optimizer, with a learning rate of 0.001 and a batch size of 32. We employ techniques such as dropout and L2 regularization to prevent overfitting and improve the generalization ability of the model.

The proposed multi-modal approach, combining the fine-tuned BERT model for textual analysis with physiological data from wearable devices, aims to provide a more comprehensive and accurate assessment of an individual's mental health status, particularly in the context of suicide risk.

## 6. Observations

### *6.1. Intermediate Classification of PHQ-9 Score*

Acquiring PGHD on a large scale requires a low-burden data collection approach, thus participants were only asked to complete the PHQ-9 at sparse intervals, once every 3 months. Consequently, we are limited to a relatively small set of reference labels: on average 2.07 labels per enrolled participant over the course of one year. The first phase of our approach thus generates more frequent intermediate depression severity labels, which are used in combination with self-reported reference labels to reduce the sparsity of the dataset by up to 3x. We were able to

construct a multi-class classification model that determines a participants depression severity for a given month, by assigning an individual to one of 5 ordinal PHQ-9 classes describing severity from minimal to severe [8].

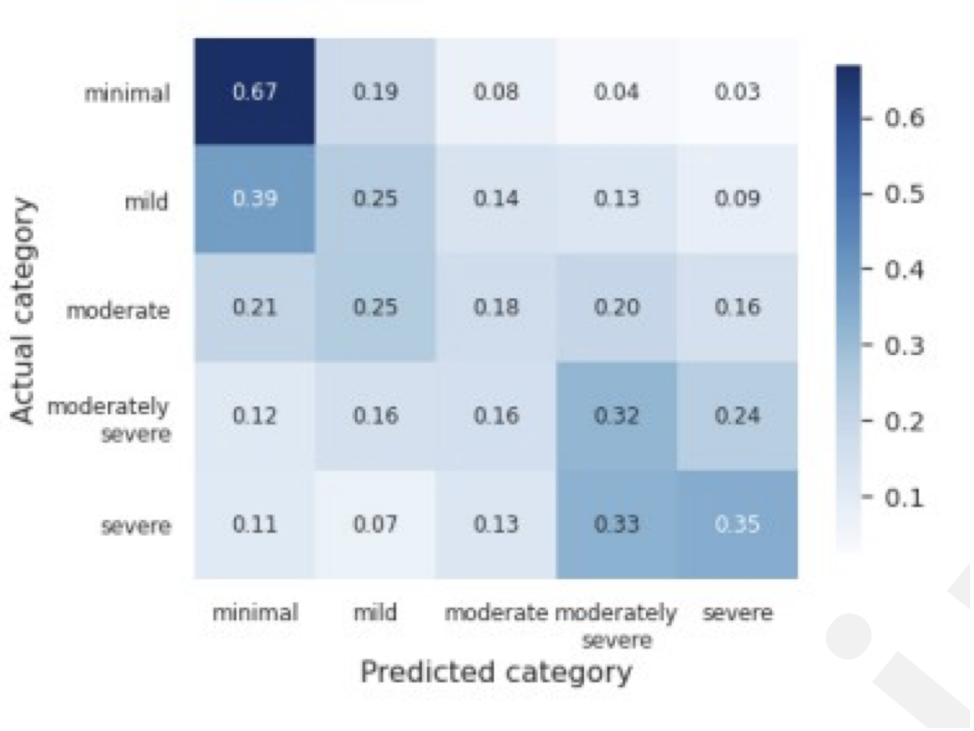



Figure 8: Confusion Matrix for Category Depression

The model includes features selected across all input sources, including demographics (gender, birth year, education, body mass index), life events and conditions at baseline (whether they had received financial assistance, experienced trauma or given birth, diagnosis with a range of chronic conditions), LMC (changes to medications or lifestyle) and sleep-related wearable PGHD (the number of hypersomnia nights, range of bedtime, and average ratio of the time spent asleep to the time spent in bed). A full list of the final features and their relative importance is included in the Supplementary Materials. This model was then used to generate intermediate PHQ-9 category labels for each individual for SM1 and SM2.

### *6.2. Prediction of longitudinal change*

The intermediate generation of depression severity labels means that each sample now consists of the PHQ-9 depression severity at SM0, the LMC surveys, wearable PGHD, and up to 2 generated labels which provide a weak estimate of depression severity (PHQ-9 category) at SM1 and SM2. We then posed our original aim as a binary problem: can we predict increased depression severity?

We defined increased depression severity as when a participant changed PHQ-9 category between SM0 and SM3. From our 10,866 samples, 2,252 (20.7%) were thus labeled as positive

cases.

The second phase model was optimized using various input feature sets and LightGBM model hyperparameters. As shown in Figure 3, the optimization process also relied on the outputs generated by the first phase.

We used several metrics to evaluate performance, but prioritized sensitivity as the primary metric, as our main objective was to accurately identify the highest proportion of individuals reporting increased depression severity. Since the dataset was highly imbalanced, with only 21% of individuals reporting increased depression severity, we optimized performance for both the majority and minority classes. Therefore, we considered specificity and AUPRC as secondary performance metrics to observe the tradeoff in performance for each class.

The best-performing model selected 13 input features, achieving a sensitivity of 55.4% (95% CI 0.8%), specificity of 65.3% (95% CI 4.2%), and AUPRC of 0.31 (95% CI 0.024). In comparison, a baseline model that randomly assigned 20.7% positive labels reported an AUPRC of 0.21, a sensitivity of 19.8%, and a specificity of 80.0% (averaged across 10 runs of 1,000 samples).

We analyzed the most important features in the second phase of the combined pipeline and found that the selected features for predicting relative changes in depression were similar to those selected for predicting absolute depression in the first phase. Figure 9 presents the most

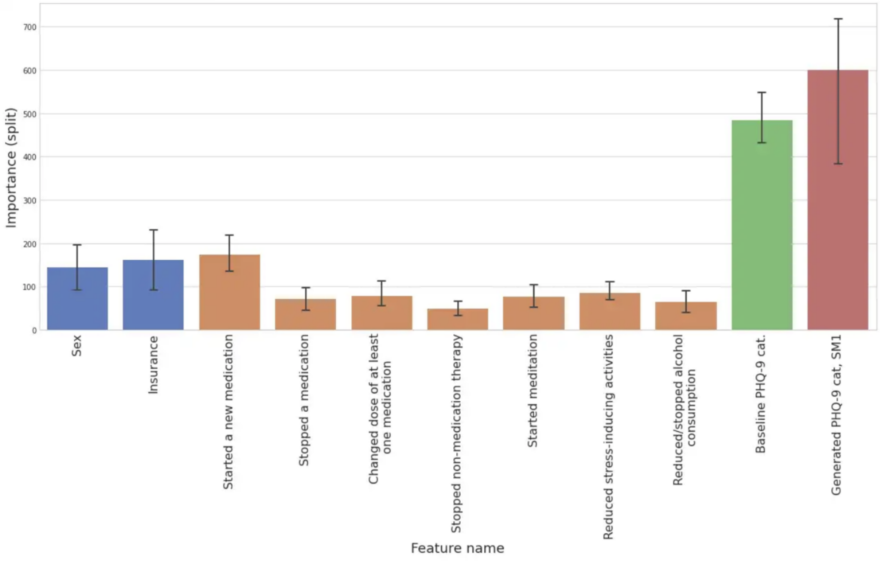



Figure 9: The split feature importance of the features that were selected in at least three out of five train-test splits in the best-performing phase 2c model is presented. The colors in the visualization represent different types of features: static screener features are depicted in blue, lifestyle and medication changes (LMC) features are shown in orange, baseline Patient Health Questionnaire-9 (PHQ-9) features are displayed in green, and generated PHQ-9 features are represented in red. The figure also includes the 95% confidence intervals (CIs) for the split feature importance, providing a measure of the uncertainty associated with each feature's importance. The abbreviation "SM" in the description likely refers to "sample month," indicating that the PHQ-9 features were collected at different time points throughout the study.

important features, with additional details in Multimedia Appendix 2. Regardless of the cohort, PHQ-9related features were consistently selected as strong predictors of an increase in depression severity. Specifically, the self-reported starting PHQ-9 category and the generated intermediate PHQ-9 category for SM1 were the most important features, as shown in Figure 5. Among the static demographic and socioeconomic features, sex and having health insurance were the most important. Various self-reported LMC features were frequently selected, including medication changes (starting, stopping, and changing doses) and stress-related lifestyle changes (starting meditation and reducing stress-inducing activities), as well as reducing or stopping alcohol consumption. We observed that objective sleep features were again selected, but no specific individual wearable PGHD feature (sleep or otherwise) was consistently selected to be included in the final model.

## 7. Conclusion and future work

Patient-Generated Health Data (PGHD) offer a low-burden, direct connection to the patient journey and have been shown to be a valuable component of models that predict health-relevant outcomes [[38],[39]]. In this study, we introduce a two-phase approach for predicting longitudinal deterioration in depression status. Phase 1c involves increasing label density by generating intermediate PHQ-9 category labels using wearable PGHD and Lifestyle and Medication Change (LMC) information. In the second phase, we combine self-reported and generated PHQ-9 category labels with additional recent wearable PGHD and LMC information to predict the deterioration of depression status three months after the initial self-report. This two-phase approach is low-burden and requires minimal participant interaction, as the input information consists of simple self-reports and data from consumer-grade wearables.

Although the overall performance in phase 1c was not exceptionally strong ($\kappa$=0.476, 95% CI ± 0.017), we were encouraged by the high adjacent accuracy (77.6%) and the correspondence between the features in the final tuned models and known risk factors for depression, such as gender, experience of trauma, and chronic comorbidities. These factors have been shown to influence depression in large-scale studies[40]. We also observed the selection of objective sleep features, which have been previously associated with depressive disorders using low-cost wearable devices [41], PGHD [42], and smartphones [43]. Additionally, we noted that performance varied across severity groups, with high performance for individuals with either relatively mild or severe depression.

In phase 2c, our best-performing model achieved a sensitivity of 55.4%, specificity of 65.3% (95% CI 4.2%), and an Area Under the Precision-Recall Curve (AUPRC) of 0.31 (95% CI 0.024). Compared to simulating random assignment of 20.7% positive labels across 10 iterations of 1,000 samples, which resulted in an AUPRC of 0.21, a sensitivity of 19.8%, and a specificity of 80.0%, our model nearly tripled sensitivity while only slightly reducing specificity. We prioritized sensitivity because the potential consequences of false negatives (i.e., not identifying a person with deteriorating depression) are much higher than the cost of false positives (i.e., incorrectly suspecting someone of deteriorating depression).

Features from all input sources were selected in the best-performing models, but with varying relative importance. Static features, defined at enrollment and remaining constant, were selected but had relatively low importance. These included features known to be relevant to the risk of developing depression, such as the presence of chronic comorbidities [44], ethnicity [45], financial difficulties [46], and pregnancy [47]. Features derived from wearable devices, including trends in sleep onset time, percentage of sleep time spent awake, and overall number of hypersomnia

days, were also selected. The most important features were those generated in phase 1c, i.e., the probability of an individual being in a given PHQ-9 class, summarizing features from across all input sources. The intermediate labels generated in phase 1c are inspired by the concept of "weak labeling," which can help reduce large-scale noisy data to a signal useful for supervised learning [48]. Due to data sparsity, intermediate labeling was not always available, resulting in some samples having one or no intermediate PHQ-9 category labels. However, LightGBM's ability to handle missing values ensured that the lack of intermediate labeling or missing PGHD values did not pose problems in the phase 2c model predictions, highlighting the low-burden and robust nature of the described approach.

From these findings, we deduced that the average sleep onset time is a good determinant of increasing depression severity, consistent with previous research [41], but that variability in sleep is participant-specific and not necessarily a good predictor for generalizing to other participants.

## References


[1] H. Cai, X.-M. Xie, Q. Zhang, X. Cui, J.-X. Lin, K. Sim, G. S. Ungvari, L. Zhang, Y.-T. Xiang, Prevalence of suicidality in major depressive disorder: a systematic review and meta-analysis of comparative studies, Frontiers in psychiatry 12 (2021) 690130.

[2] K. L. Lovero, P. F. Dos Santos, A. X. Come, M. L. Wainberg, M. A. Oquendo, Suicide in global mental health, Current psychiatry reports 25 (6) (2023) 255–262.

[3] W. H. Organization, Depression and other common mental disorders: global health estimates (2017) 24 p.

[4] I. Berardelli, G. Serafini, N. Cortese, F. Fiaschè, R. C. OConnor, M. Pompili, The involvement of hypothalamus–pituitary–adrenal (hpa) axis in suicide risk, Brain sciences 10 (9) (2020) 653.

[5] V. Vaccarino, L. Badimon, J. D. Bremner, E. Cenko, J. Cubedo, M. DorobanÅ£u, D. J. Duncker, . Koller, O. Manfrini, D. Milii, T. Padr, A. R. Pries, A. A. Quyyumi, D. Tousoulis, D. Trifunovi, Z. Vasiljevi, C. De Wit, R. Bugiardini, Depression and coronary heart disease: 2018 position paper of the ESC working group on coronary pathophysiology and microcirculation, European heart journal 41 (17) (2019) 1687–1696. doi:10.1093/eurheartj/ehy913.
URL https://doi.org/10.1093/eurheartj/ehy913

[6] R. A. Bernert, T. E. Joiner, Sleep disturbances and suicide risk: A review of the literature, Neuropsychiatric disease and treatment Volume 3 (2008) 735743. doi:https://doi.org/10.2147/ndt.s1248.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656315/

[7] A. L. Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, C. Lemey, Machine learning and natural language processing in mental health: Systematic review, JMIR. Journal of medical internet research 23 (5) (2021) e15708e15708. doi:https://doi.org/10.2196/15708.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8132982/

[8] K. Kroenke, R. L. Spitzer, J. B. Williams, The phq-9: validity of a brief depression severity measure, Journal of general internal medicine 16 (9) (2001) 606–613.

[9] V. R. Meghrajani, M. Marathe, R. Sharma, A. Potdukhe, M. B. Wanjari, A. B. Taksande, A comprehensive analysis of mental health problems in india and the role of mental asylums, Curusdoi:https://doi.org/10.7759/cureus.42559.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10460242/

[10] A. Haines-Delmont, G. Chahal, A. J. Bruen, A. Wall, C. T. Khan, R. Sadashiv, D. Fearnley, et al., Testing suicide risk prediction algorithms using phone measurements with patients in acute mental health settings: feasibility study, JMIR mHealth and uHealth 8 (6) (2020) e15901.

[11] S. Choudhary, N. Thomas, J. Ellenberger, G. Srinivasan, R. Cohen, et al., A machine learning approach for detecting digital behavioral patterns of depression using nonintrusive smartphone data (complementary path to patient health questionnaire-9 assessment): Prospective observational study, JMIR Formative Research 6 (5) (2022) e37736.

[12] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, 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.

[13] S. Saeb, M. Zhang, M. Kwasny, C. J. Karr, K. Kording, D. C. Mohr, The relationship between clinical, momentary, and sensor-based assessment of depression, in: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), IEEE, 2015, pp. 229–232.

[14] K. Opoku Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, D. Ferreira, Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study, JMIR mHealth and uHealth 9 (7) (2021) e26540.
[15] R. Razavi, A. Gharipour, M. Gharipour, Depression screening using mobile phone usage metadata: a machine learning approach, Journal of the American Medical Informatics Association 27 (4) (2020) 522–530.
[16] J. Goltermann, D. Emden, E. J. Leehr, K. Dohm, R. Redlich, U. Dannlowski, T. Hahn, N. Opel, et al., Smartphone-based self-reports of depressive symptoms using the remote monitoring application in psychiatry (remap): interformat validation study, JMIR Mental Health 8 (1) (2021) e24333.
[17] A. H. Kemp, D. S. Quintana, The relationship between mental and physical health: insights from the study of heart rate variability, International journal of Psychophysiology 89 (3) (2013) 288–296.
[18] D. Lee, J. H. Baek, Y. J. Cho, K. S. Hong, Association of resting heart rate and heart rate variability with proximal suicidal risk in patients with diverse psychiatric diagnoses, Frontiers in psychiatry 12 (2021) 652340.
[19] A. H. Kemp, D. S. Quintana, C. R. Quinn, P. Hopkinson, A. W. F. Harris, Major depressive disorder with melancholia displays robust alterations in resting state heart rate and its variability: implications for future morbidity and mortality, Frontiers in Psychology 5. doi:10.3389/fpsyg.2014.01387.
URL https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01387
[20] K. Georgiou, A. V. Larentzakis, N. N. Khamis, G. I. Alsuhaibani, Y. A. Alaska, E. J. Giallafos, Can wearable devices accurately measure heart rate variability? a systematic review, Folia medica 60 (1) (2018) 7–20.
[21] B. Bent, B. A. Goldstein, W. A. Kibbe, J. P. Dunn, Investigating sources of inaccuracy in wearable optical heart rate sensors, NPJ digital medicine 3 (1) (2020) 18.
[22] Q. Zhang, X. Zeng, W. Hu, D. Zhou, A machine learning-empowered system for long-term motion-tolerant wearable monitoring of blood pressure and heart rate with ear-ecg/ppg, IEEE Access 5 (2017) 10547–10561. doi:10.1109/ACCESS.2017.2707472.
[23] N. Sjöström, M. Waern, J. Hetta, Nightmares and sleep disturbances in relation to suicidality in suicide attempters, Sleep 30 (1) (2007) 91–95.
[24] M. Y. Agargun, R. Cartwright, Rem sleep, dream variables and suicidality in depressed patients, Psychiatry research 119 (1-2) (2003) 33–39.
[25] R. A. Bernert, T. E. Joiner, Sleep disturbances and suicide risk: a review of the literature, Neuropsychiatric disease and treatment 3 (6) (2007) 735–743.
[26] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, R. M. Aarts, Sleep stage classification from heart-rate variability using long short-term memory neural networks, Scientific reports 9 (1) (2019) 14149.
[27] A. H. Farabaugh, D. Mischoulon, M. Fava, C. Green, W. Guyker, J. Alpert, The potential relationship between levels of perceived stress and subtypes of major depressive disorder (mdd), Acta Psychiatrica Scandinavica 110 (6) (2004) 465–470.
[28] M. Grasdalsmoen, H. R. Eriksen, K. J. Lønning, B. Sivertsen, Physical exercise, mental health problems, and suicide attempts in university students, BMC psychiatry 20 (1) (2020) 1–11.
[29] N. Fabiano, A. Gupta, J. G. Fiedorowicz, J. Firth, B. Stubbs, D. Vancampfort, F. B. Schuch, L. J. Carr, M. Solmi, The effect of exercise on suicidal behaviors: A systematic review and meta-analysis of randomized controlled trials, Journal of affective disorders.
[30] A. Arowosegbe, T. Oyelade, Application of natural language processing (nlp) in detecting and preventing suicide ideation: A systematic review, International Journal of Environmental Research and Public Health 20 (2) (2023) 1514.
[31] Z. Xu, Y. Xu, F. Cheung, M. Cheng, D. Lung, Y. W. Law, B. Chiang, Q. Zhang, P. S. Yip, Detecting suicide risk using knowledge-aware natural language processing and counseling service data, Social science & medicine 283 (2021) 114176.
[32] R. Haque, N. Islam, M. Islam, M. M. Ahsan, A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning, Technologies 10 (3) (2022) 57.
[33] M. Makhmutova, R. Kainkaryam, M. Ferreira, J. Min, M. Jaggi, I. Clay, Prediction of self-reported depression scores using person-generated health data from a virtual 1-year mental health observational study (2021) 411doi:10.1145/3469266.3469878.
URL https://doi.org/10.1145/3469266.3469878
[34] R. K. Vinayak, R. Gilad-Bachrach, Dart: Dropouts meet multiple additive regression trees, in: Artificial Intelligence and Statistics, PMLR, 2015, pp. 489–497.
[35] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning 46 (2002) 389–422.
[36] J. L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and psychological measurement 33 (3) (1973) 613–619.
[37] D. W. Aha, R. L. Bankert, A comparative evaluation of sequential feature selection algorithms (1995) 1–7.
[38] M. Karas, N. Marinsek, J. Goldhahn, L. Foschini, E. Ramirez, I. Clay, Predicting subjective recovery from lower

limb surgery using consumer wearables, Digital biomarkers 4 (Suppl. 1) (2020) 73–86.
[39] J. Dunn, L. Kidzinski, R. Runge, D. Witt, J. L. Hicks, S. M. Schüssler-Fiorenza Rose, X. Li, A. Bahmani, S. L. Delp, T. Hastie, et al., Wearable sensors enable personalized predictions of clinical laboratory measurements, Nature medicine 27 (6) (2021) 1105–1112.
[40] A. Brailean, J. Curtis, K. Davis, A. Dregan, M. Hotopf, Characteristics, comorbidities, and correlates of atypical depression: evidence from the uk biobank mental health survey, Psychological medicine 50 (7) (2020) 1129–1138.
[41] Y. Zhang, A. A. Folarin, S. Sun, N. Cummins, R. Bendayan, Y. Ranjan, Z. Rashid, P. Conde, C. Stewart, P. Laiou, et al., Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study, JMIR mHealth and uHealth 9 (4) (2021) e24604.
[42] S. Kumar, J. L. Tran, E. Ramirez, W.-N. Lee, L. Foschini, J. L. Juusola, et al., Design, recruitment, and baseline characteristics of a virtual 1-year mental health study on behavioral data and health outcomes: observational study, JMIR Mental Health 7 (7) (2020) e17075.
[43] I. Moshe, Y. Terhorst, K. Opoku Asare, L. B. Sander, D. Ferreira, H. Baumeister, D. C. Mohr, L. Pulkki-Råback, Predicting symptoms of depression and anxiety using smartphone and wearable data, Frontiers in psychiatry 12 (2021) 625247.
[44] H. Li, S. Ge, B. Greene, J. Dunbar-Jacob, Depression in the context of chronic diseases in the united states and china, International journal of nursing sciences 6 (1) (2019) 117–122.
[45] R. K. Bailey, J. Mokonogho, A. Kumar, Racial and ethnic differences in depression: current perspectives, Neuropsychiatric disease and treatment (2019) 603–609.
[46] T. Richardson, P. Elliott, R. Roberts, M. Jansen, A longitudinal study of financial difficulties and mental health in a national sample of british undergraduate students, Community mental health journal 53 (2017) 344–352.
[47] E. OConnor, C. A. Senger, M. L. Henninger, E. Coppola, B. N. Gaynes, Interventions to prevent perinatal depression: evidence report and systematic review for the us preventive services task force, Jama 321 (6) (2019) 588–601.
[48] A. Zhan, S. Mohan, C. Tarolli, R. B. Schneider, J. L. Adams, S. Sharma, M. J. Elson, K. L. Spear, A. M. Glidden, M. A. Little, et al., Using smartphones and machine learning to quantify parkinson disease severity: the mobile parkinson disease score, JAMA neurology 75 (7) (2018) 876–880.