Original Paper

Empowering Mental Health Monitoring Using a Macro-Micro Personalization Framework for Multimodal-Multitask Learning: Descriptive Study

Meishu Song[1], MSc; Zijiang Yang[1], MSc; Andreas Triantafyllopoulos[2], MSc; Zixing Zhang[3], PhD; Zhe Nan[4], MSc; Muxuan Tang[1], MSc; Hiroki Takeuchi[1], PhD; Toru Nakamura[5], PhD; Akifumi Kishi[1], PhD; Tetsuro Ishizawa[1], MD, PhD; Kazuhiro Yoshiuchi[1], MD, PhD; Björn Schuller[6], PhD; Yoshiharu Yamamoto[1], PhD

[1]Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo, Japan

[2]Technical University of Munich, Munich, Germany

[3]Hunan University, Changsha, China

[4]Google, San Carlos, CA, United States

[5]Osaka University, Osaka, Japan

[6]Imperial College London, London, United Kingdom

Corresponding Author:
Yoshiharu Yamamoto, PhD
Educational Physiology Laboratory
Graduate School of Education
The University of Tokyo
7 Chome-3-1 Hongo
Bunkyo City
Tokyo, 1130033
Japan
Phone: 81 09035784363
Email: yamamoto@p.u-tokyo.ac.jp

Abstract

Background: The field of mental health technology presently has significant gaps that need addressing, particularly in the domain of daily monitoring and personalized assessments. Current noninvasive devices such as wristbands and smartphones are capable of collecting a wide range of data, which has not yet been fully used for mental health monitoring.

Objective: This study aims to introduce a novel dataset for personalized daily mental health monitoring and a new macro-micro framework. This framework is designed to use multimodal and multitask learning strategies for improved personalization and prediction of emotional states in individuals.

Methods: Data were collected from 298 individuals using wristbands and smartphones, capturing physiological signals, speech data, and self-annotated emotional states. The proposed framework combines macro-level emotion transformer embeddings with micro-level personalization layers specific to each user. It also introduces a Dynamic Restrained Uncertainty Weighting method to effectively integrate various data types for a balanced representation of emotional states. Several fusion techniques, personalization strategies, and multitask learning approaches were explored.

Results: The proposed framework was evaluated using the concordance correlation coefficient, resulting in a score of 0.503. This result demonstrates the framework's efficacy in predicting emotional states.

Conclusions: The study concludes that the proposed multimodal and multitask learning framework, which leverages transformer-based techniques and dynamic task weighting strategies, is superior for the personalized monitoring of mental health. The study indicates the potential of transforming daily mental health monitoring into a more personalized app, opening up new avenues for technology-based mental health interventions.

(JMIR Ment Health 2024;11:e59512) doi: 10.2196/59512

KEYWORDS

multimodal; multitask; daily mental health; mental health; monitoring; macro; micro; framework; personalization; strategies; prediction; emotional state; wristbands; smartphone; mobile phones; physiological; signals; speech data;

Introduction

Background

Mental health, recognized as a critical component of overall well-being, has garnered increasing attention and concern. The World Health Organization [1] defines mental health as a state of well-being where individuals realize their potential, cope with normal life stresses, work productively, and contribute to their community. However, mental health issues continue to present a significant burden globally, affecting individuals' quality of life and posing challenges. In response to these challenges, the concept of "daily mental health monitoring" has emerged as a critical area of research and application [2,3]. This concept refers to the regular, continuous observation and assessment of an individual's emotional states, using a variety of methods and tools to capture data in real time [4]. Such monitoring aims to provide a comprehensive understanding of an individual's mental health, facilitating early detection of patterns, changes, or emerging issues. Consequently, accurate monitoring and understanding of daily mental health have become imperative for timely interventions and sustained mental well-being.

However, the field of daily mental health monitoring remains surprisingly underdeveloped, particularly regarding real-life applications [4]. The challenges faced by existing datasets and methods are not merely academic concerns but represent significant barriers to the effective and widespread adoption of mental health monitoring in everyday life. These challenges include the following.

- Real-world representation where a significant portion of existing datasets lack data derived from real-world settings, instead relying on artificial or laboratory conditions.
- Lack of self-annotation. many datasets do not use self-annotation [5-8], relying instead on experts' observation or clinical interpretation. This approach often fails to capture the subjective experience of the individual, crucial for a person-centered understanding and monitoring of mental health [4,9]. In addition, clinical assessments typically occur at discrete time points, potentially missing the dynamic, moment-to-moment fluctuations in mental states that individuals experience in their daily lives [10].
- Challenges in accessibility of monitoring data. Many studies use electroencephalography (EEG) [11,12] while providing valuable insights into brain activity and emotional states, requiring specialized equipment and expertise, making it impractical for daily monitoring. Similarly, facial expression data [13,14] capture often necessitates continuous video monitoring, posing substantial privacy and practicality challenges for everyday use.
- Limited modalities and single-model approach. Most available research focuses on a single modality [15,16], this overlooks the inherently multimodal nature of human emotional expression and mental states, reducing the systems' reliability.

To address the aforementioned challenges, our study adopts an innovative methodology aimed at forging more accurate, and efficacious tools for mental health monitoring. The contributions of our research are manifold, highlighted by the following key developments.

- Introducing a novel dataset collected from 298 individuals using noninvasive, everyday devices including wristband-type devices and smartphones, our dataset captures physiological signals: zero crossing mode (ZCM), proportional integration mode (PIM), and speech data. Participants provided self-annotated emotional states over 2 weeks, creating a rich, multimodal resource for understanding daily mental health dynamics.
- Developing a macro-micro framework for personalized daily mental health. Our framework develops a multimodal and multitask learning (MTL) strategy, innovatively built global emotion embeddings with individual personalization embedding.

In our research, the decision to focus on physiological signals and speech data, while excluding modalities such as facial expressions and EEG, was driven by several key considerations: Physiological signals have been shown to have a significant association with mental health and well-being. These signals, such as heart rate, skin conductance, and activity levels, can provide valuable insights into an individual's emotional and psychological state [17]. The relationship between physiological signals and mental health is complex and multifaceted. For example, changes in heart rate variability (HRV) have been linked to stress, anxiety, and depression [18]. Reduced HRV has been observed in individuals with mental health disorders, suggesting that it may serve as a potential biomarker for mental well-being [19].

In our research, we apply the wrist-worn device used in our study which is equipped with a highly sensitive piezoelectric accelerometer that can detect even the most subtle wrist movements, with a resolution as fine as 0.01 G/rad/s. This allows for the capture of a wide range of daily activities and movements that may be relevant to mental health assessment [20]. The device uses 2 key modes for processing the accelerometer data: ZCM and PIM [21].

In addition to physiological signals, our study also incorporates speech data as a key modality for assessing mental health. Speech provides a rich source of information about an individual's emotional state, cognitive functioning, and overall well-being [22]. There are several reasons why speech is a valuable tool for mental health assessment. First, speech carries emotional information through various features such as tone, pitch, and intonation. Changes in these features can reflect an individual's emotional state, such as increased monotonicity in speech being associated with depression [23]. Second, speech patterns and characteristics can provide insights into an individual's cognitive processes. For example, changes in speech fluency, coherence, and word choice have been linked to cognitive impairments and mental health conditions [24]. In addition, speech data can be collected noninvasively using readily available devices such as smartphones or voice recorders. This makes it a convenient and accessible modality for mental health assessment, especially in remote or telehealth settings [25].

Ecological momentary assessment (EMA) is a key methodological approach used in our study. EMA involves the repeated sampling of individuals' current behaviors and experiences in real time, in their natural environments [26]. While EMA offers several advantages, such as reducing recall bias and capturing the dynamics of mental states in real-world contexts [10], it also has limitations. These include potential reactivity (ie, the act of self-reporting influencing the very experiences being reported) and compliance issues [27]. In this study, we aim to mitigate these limitations through careful design and participant training, which will be discussed in the Methods section.

Furthermore, our study introduces the Dynamic Restrained Uncertainty Weighting (DRUW) fusion method, a novel approach for integrating multimodal data in the context of mental health monitoring. The DRUW fusion method adaptively weights the contribution of each modality based on its uncertainty and distinct characteristics, ensuring a balanced representation of the fused data. This method builds upon the principles of uncertainty weighting [28] and extends them to the multimodal fusion context. The key novelty of the DRUW fusion method lies in its ability to dynamically adjust the weighting of each modality based on the inherent uncertainty and complementary nature of the physiological signals and speech data [29].

By collecting and analyzing data on emotional states, speech characteristics, and physiological patterns, our study aims to contribute to the development of more effective, personalized, and accessible mental health interventions. The data collected in our study can contribute to better mental health outcomes in several ways.

Early Detection and Intervention

By correlating objective measures with subjective emotional states, we can develop tools for early detection of mental health issues, enabling timely interventions [30].

Personalized Treatment and Monitoring

Insights from our study can inform personalized treatment plans and monitoring strategies, tailoring interventions to individual needs [31].

Remote Monitoring and Telemedicine

Our use of wearable devices and speech analysis can contribute to remote monitoring tools, crucial for mental health support, especially in light of recent global events [32].

Reducing Stigma and Increasing Access

By demonstrating objective measures for mental health assessment, we can potentially reduce stigma and increase access to care, particularly for underserved populations [33].

Related Work

This section delves into various aspects of mental health monitoring research domain, including mental health data, EMA, personalization, and their real-world implications.

Recent research efforts, particularly in mental health detection and monitoring, have gained significant momentum. Key studies like the systematic review by Hickey et al [34] and Long et al [35] have critically evaluated the use of smart devices and wearable technologies. These investigations underline the capability of these devices in detecting stress, anxiety, and depression through physiological measures such as HRV, electrodermal activity, and EEG data. However, they also identify a notable gap in the availability of commercial depression-detecting devices, emphasizing the need for integrating multimodal data to enhance both accuracy and predictive power.

Recent advancements in multimodal data analysis have shown promising results in mental health diagnosis [36]. For instance, a study by Xu et al [36] proposed a measurement method for mental health based on dynamic multimodal feature recognition. This approach integrates various data sources, including physiological signals, speech patterns, and behavioral indicators, to provide a more comprehensive assessment of an individual's mental state. Similarly, Huckins et al [37] developed a multimodal machine learning approach that combines smartphone sensing data with self-reported mental health scores to predict changes in depression and anxiety among college students.

Building on this, the role of mental health datasets becomes crucial in understanding the complex and varied nature of mental health conditions across different populations. The comprehensive analysis of datasets, such as those examined during the COVID-19 pandemic [38], offers deep insights into the mental health effects of global crises on specific demographics, like the Bangladeshi population. These datasets are instrumental not only in assessing the prevalence and severity of mental health conditions across various groups but also in supporting longitudinal studies vital for tracking changes over time.

Recent studies have also focused on improving data collection methods for mental health monitoring. For example, Morshed et al [39] introduced a novel approach using passive sensing and machine learning to predict mood instability in bipolar disorder. This method leverages smartphone usage patterns and environmental data to provide continuous, unobtrusive monitoring of mental health states.

In this study, we apply EMA [40], which represents a method for recording participants' behavior, psychological state, and physical symptoms in real-time and at multiple time points. The primary advantage of EMA lies in its ability to minimize the biases often associated with retrospective recall in self-report data. Traditional self-report measures, which ask participants to remember and report past feelings, behaviors, or symptoms, can be influenced by memory distortions and subjective interpretations of past events, Thus, it reduces the likelihood of recall errors and increases the accuracy and reliability of the data collected.

Further, personalization in mental health monitoring systems is increasingly important [4]. Innovations in digital phenotyping [41] exemplify this trend. This is further advanced by groundbreaking approaches like those proposed by Gerczuk et al [42], using zero-shot personalization strategies for large speech foundation models in mood recognition.

The application of artificial intelligence and deep learning techniques in mental health monitoring has seen significant growth. A comprehensive review by Su et al [43] highlights the potential of deep learning models in analyzing multimodal data for mental health assessment. These advanced techniques allow for more nuanced interpretation of complex, high-dimensional data, potentially leading to more accurate and personalized mental health interventions.

In summary, the related work shows the dynamic nature of mental health detection and monitoring. However, bridging the gap between technological capabilities and personalized mental health care presents numerous challenges. The integration of multimodal data, advancements in data collection methods, and the application of sophisticated artificial intelligence techniques represent promising avenues for overcoming these challenges and improving mental health monitoring and diagnosis.

Methods

Methodology and Data Collection

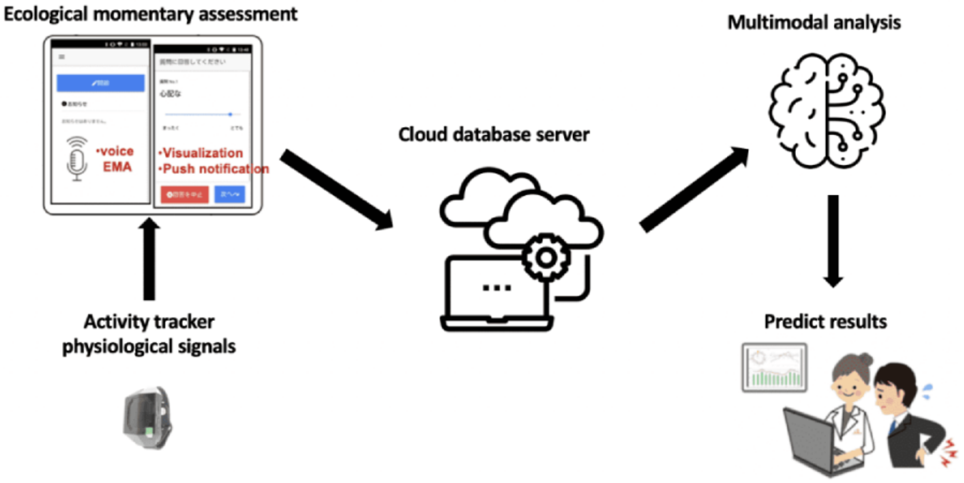
Overview

The study followed a 2-week data collection protocol, during which participants wore wrist-worn devices and used a smartphone app to record their speech and self-report their emotional states. The collected data included physiological signals from the wrist-worn devices and speech recordings from the smartphone app. To collect data, we developed a platform called Mental Healthcare Internet of Things (MHIT) system.

The MHIT System

The MHIT system, as shown in Figure 1, is a cloud-based platform specifically crafted to gather and analyze data from Internet of Things devices. This state-of-the-art system combines the collection of physical activity signals with speech data. The MHIT system is comprised of 2 key components: a cloud server (MHIT server) and a smartphone app (MHIT app).

Figure 1. The MHIT system designed for data collection. MHIT: Mental Healthcare Internet of Things.



Participants

A convenience sample of 298 Japanese office workers participated in our study. They were recruited by sending digital flyers. Those who agreed to participate in the study were asked to open a URL link on the flyer and complete a web-based registration form. Subsequently, the in-house wrist-worn device (ScienceNet device; ScienceNet Inc), survey progression guide, and informed consent form were mailed to them. Participants were instructed on the aim and procedure of the study through the survey progression guide. In addition, by scanning a QR code on the guide, they were able to watch a tutorial video to learn the use of the device and the MHIT app. After completing the informed consent form, they commenced the study. This recruitment process was conducted fully on the web, which contribute to achieve data collection from workers in different residential locations.

Annotation and EMA

This study used an EMA paradigm to capture participants' emotional states in real time, thus avoiding potential distortions of retrospective recall in self-report data. The EMA protocol involved the following.

Sampling Scheme

Participants were prompted to report their emotional states by using the MHIT app at randomly selected times within –10 minutes to +10 minutes of predetermined times (11 AM, 3 PM, and 7 PM). In addition, they were instructed to voluntarily complete the same EMA questionnaires when they woke up and went to bed.

Data Collection

At each EMA, participants self-reported the intensity of 9 different expressed emotions on a (0:100) visual-analogue scale (slider) displayed on the MHIT app. These 9 emotions correspond to the items of the Depression and Anxiety Mood Scale (DAMS) [44]. To prevent response bias and predetermination, the order of the DAMS items was randomized for each evaluation.

Speech Data

Before the mental state evaluation, participants recorded their voices by speaking, for example, “the current date and time are September 5, 2022, at 10:23 PM” on the MHIT app. The reasoning behind this is to keep the content emotionally neutral. Participants also recorded activities, and the actual time was recorded by the system.

Physiological Signals

The instrument is fitted with a sensitive piezoelectric accelerometer that detects minute wrist accelerations (as fine as 0.01 G/rad/s), capturing even the most subtle daily movements. The ZCM within the device tallies the instances the accelerometer’s signal traverses the 0 mark over a predefined duration, known as the epoch time. Conversely, the PIM assesses the integral of the root-mean-square for the triaxial accelerometer signals. For the purposes of this investigation, we have configured the epoch interval at 1 minute, aggregating 60 data points (representing 1 hour) prior to each participant’s DAMS entry within their routine activities. To ensure the integrity of our dataset, we have meticulously curated instances that comprise both ZCM and PIM recordings, each consisting of 60 data points.

Annotation Scheme

The DAMS serves as a self-reported measure of an individual’s emotional state, providing a subjective assessment of their mental well-being. This scale, which encompasses nine distinct emotions—vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed, and worried—is used for comprehensively assessing mental health experiences pertinent to depression and anxiety. DAMS’s effectiveness in measuring depressive and anxious moods is particularly notable, as it uses a variety of descriptors, including adjectives, adjectival verbs, and phrases, to delineate depressive, anxious, and positive moods with high discriminant validity [44]. Moreover, its psychometric soundness has been established through methods such as parallel testing and test-retest evaluations [44], confirming its high convergent, discriminant validity, and reliability. The scale’s sensitivity to mood fluctuations is evidenced by the variance in scores observed between normal and stressful periods, underscoring its use in detecting mood changes. These features of DAMS, combined with its thorough statistical analysis across 9 emotional labels, confirmed it is a comprehensive choice for our study.

Previous research has shown that speech characteristics and patterns can reflect an individual’s emotional state. For example, depression has been associated with changes in prosody, such as reduced pitch variability and slower speaking rate [22]. Similarly, anxiety has been linked to increased vocal tension and higher fundamental frequency [45]. By analyzing speech features such as pitch, intonation, and speaking rate, we can potentially identify objective markers that correlate with the subjective emotional states reported through DAMS.

Physiological data, collected through wrist-worn accelerometers, can also provide an indirect measure of an individual’s emotional state. Studies have demonstrated that mood disorders, such as depression and anxiety, can influence an individual’s activity levels and patterns [46]. Depression, for instance, has been associated with reduced physical activity and increased sedentary behavior [47]. By examining the activity data captured by the accelerometers, we can explore potential correlations between the objective measures of physical activity and the subjective emotional states assessed by DAMS.

While speech and physical activity data do not directly measure the emotional states captured by DAMS, they can offer complementary and objective insights into an individual’s mental well-being. By combining these different modalities—self-reported mood, objective speech characteristics, and objective physiological signal patterns—we aim to develop a more comprehensive understanding of an individual’s mental health status.

Self-Annotation

Self-annotation is a cornerstone in daily mental health monitoring for important reasons such as capturing subjective emotional experience and ecological validity. Using the MHIT app, participants self-reported their emotional states 5 times daily over 2 weeks, using the 9 emotional states outlined in DAMS.

Capturing Subjective Emotional Experiences

Emotions are inherently subjective, and self-annotation allows individuals to express their emotional states based on personal experiences. This method ensures an authentic portrayal of their mental state, which is crucial for accurate mental health assessment.

Ecological Validity

By self-reporting in real-time within their usual environments, participants provide data that more accurately reflect their day-to-day emotional experiences, enhancing the ecological validity of our study.

Data Preprocessing

In our research, the preprocessing of collected data was a critical step for both speech and physical activity. This process involved several stages, each tailored to the specific nature of the data being processed.

Preprocessing of Speech Data

For speech data, audio files were standardized in terms of their sampling rate and format for subsequent analysis.

Data Cleansing

Any recordings that were unsuccessful or contained data anomalies were removed. This step was crucial to ensure the integrity and quality of the speech dataset.

Voice Activity Detection

We used algorithms to detect and eliminate silences in voice recordings. This focus on active speech segments helped in isolating meaningful data.

Denoising

Background noise within the recordings was reduced using digital signal processing techniques. While the specific method may vary depending on the characteristics of the noise and the recording environment, common approaches include spectral subtraction, Wiener filtering, or more advanced techniques such as deep learning-based noise suppression algorithms [48].

Preprocessing of Physical Activity Data

The preprocessing of physical activity data includes the following approaches.

Signal Cleaning

Similar to speech data, physical activity data were cleaned to remove any erroneous signals.

Signal Standardization

The raw data from the physical activity sensors were standardized to ensure consistency across different participants.

Normalization Process

The intensity ratings of the emotional states reported by participants were normalized to a uniform scale ranging from 0 to 1.

Multimodal Multitask Analysis

In transitioning from data collection to the analysis of daily mental health in our study, we shift our focus toward developing a robust multimodal multitask analysis framework. The initial step involves defining the analytical task, which in our case is predicting various mental health indicators as outlined by DAMS. To effectively achieve our goals in daily mental health monitoring, we need to address three pivotal questions.

- How to fuse different modalities: specifically, how do we integrate physical activities and speech data?
- How to achieve personalization: what strategies can we use to tailor the analysis to individual participants?
- How to balance different emotional states: how can we ensure that our analysis provides a balanced view of various emotional states?

To respond to these questions, our approach involves the introduction of a comprehensive framework architecture. We plan to detail each component of this framework, starting with multimodal fusion, then moving on to personalization, and concluding with multitask balancing. This sequence is carefully chosen by allowing multimodal fusion to initially integrate and align different data types (physiological signals and speech) after feature extraction, creating a whole picture of data for further analysis. Personalization subsequently adapts this integrated data to individual differences, ensuring the model accurately represents each participant's unique mental health profile. The final stage, multitask balancing, refines the network to efficiently manage multiple analytical tasks.

Framework

The proposed framework commences with a robust feature extraction phase. For physiological signals, PIM and ZCM are input into individual 2-layer feed-forward neural networks (FFNNs). Concurrently, speech signals are preprocessed through a specialized *wav2vec-l-emo* model [49], which is a pretrained model. These speech features are then similarly processed by a 2-layer FFNN. This standardization of feature dimensions across modalities primes the data for integration.

The fusion of these data streams is executed through a DRUW fusion block, effectively merging the standardized features from physiological signals and speech. This fusion process not only integrates the data but also applies DRUW fusion.

Upon fusion, the data advances into a specific emotional FFNN and a transformer layer, means, the combined features into a global emotional space. This space is not user-specific; rather, it serves as a shared domain, namely, macro space.

Personalization is introduced at the micro stage. Here, the framework uses additional FFNN layers tailored to individual users, enabling the selection of embedding elements pertinent to their unique emotional profiles. This adaptation leverages DRUW loss.

DRUW Multimodal Fusion

A fundamental question in our study of mental health monitoring is "how to fuse different modalities," specifically the integration of physical activity and speech data. To answer this, we have developed the DRUW fusion method.

The DRUW fusion formula can be represented as follows.

$$\mathcal{F}(\mathrm{w},\beta_1,\beta_2) = \frac{1}{\beta_1^2}\mathcal{M}_1(\mathrm{w}) + \frac{1}{\beta_2^2}\mathcal{M}_2(\mathrm{w})$$
$$+\log(1+\log\beta_1^2)+\log(1+\log\beta_2^2)$$
$$+|\phi-(|\log\beta_1|+|\log\beta_2|)|$$

where β_1 and β_2 are the uncertainty parameters for the physical activity data $\mathcal{M}_1(w)$ and the speech data $\mathcal{M}_2(w)$, respectively. The term ϕ acts as a constraint, similar to φ in the DRUW loss, the weighting of each modality in the fusion process is regulated, ensuring that neither modality is disproportionately represented in the fused data and maintaining a balanced integration. This adjustment, based on the uncertainty and distinct characteristics, ensures that each modality contributes appropriately to the combined dataset. Meanwhile, the complementary nature of these data types is capitalized upon by the DRUW fusion method. Physical activity data provides objective, quantifiable measures of movement and physiological responses, while speech data offers subjective insights into emotional states and mental well-being. Furthermore, a key advantage of the DRUW fusion method is its straightforward implementation.

Macro-Micro Personalization

A key aspect of our multimodal analysis framework is the implementation of macro-micro personalization.

To quantitatively define the macro-micro personalization approach, we can formulate the integration of the macro emotional space with the micro personalization layer. This can be represented as follows.

$$\mathcal{P}_i = \lambda \cdot \mathcal{E}_{macro} + (1-\lambda) \cdot \mathcal{E}_{micro,i}$$

where $\mathcal{P}_i$ represents the personalized output for the i^{th} participant, $\mathcal{E}_{macro}$ denotes the embeddings or features extracted from the macro emotional space, $\mathcal{E}_{micro,i}$ represents the embeddings or features specific to the i^{th} participant, and is a weighting factor that determines the balance between the influence of the macro and micro layers. It can be a fixed value or adaptively determined based on factors such as the diversity of the dataset or the specificity of the micro data.

Macro Emotional Space

Initially, we establish a macro emotional space that serves as a common ground for all participants. This space is built using FFNN-transformer embeddings, capturing generalized emotional patterns and trends observed across the entire participant pool. It reflects the shared aspects of emotional experiences and is crucial for understanding the broader context of mental health states.

Micro Personalization Space

After macro space, a micro layer is designed for each participant. This layer allows for the customization of the model based on microspecific data. It adapts the general insights from the macro space to the nuances of each participant's emotional profile.

DRUW Loss for Multitask

MTL is a crucial component in our study, particularly relevant to the diverse nature of mental health monitoring. MTL is a form of learning that involves training a model on several related tasks simultaneously. This approach is underpinned by the principle that "transfer should always be useful"; essentially, any pair of tasks should share some commonalities in their underlying distributions [50].

In addressing the critical challenge of "how to balance different emotional states" in our study, we use the DRUW loss [29], a solution developed in our previous work. This approach is particularly crucial in the context of multitask learning, where balancing the contribution of each task—especially when dealing with a spectrum of emotional states—is key to the overall model's performance. This method allows for the adaptive balancing of tasks, taking into account the varying degrees of complexity and uncertainty inherent in each task. The equation of the DRUW loss function is as follows.

$$\mathcal{L}(w,\alpha_1,\alpha_2) = \frac{1}{\alpha_1^2}\mathcal{L}_1(w) + \frac{1}{\alpha_2^2}\mathcal{L}_2(w)$$
$$+\log(1+\log\alpha_1^2) + \log(1+\log\alpha_2^2)$$
$$+|\varphi - (|\log\alpha_1| + |\log\alpha_2|)|$$

Where α_1 and α_2 are uncertainty parameters corresponding to different tasks in our model, with $\mathcal{L}_1(w)$ and $\mathcal{L}_2(w)$ representing the respective task-specific loss functions. The inclusion of φ serves as a constraint, regulating the sum of these weights to prevent trivial solutions and maintain the balance among tasks.

Evaluation

Experimental Setup

To construct a reliable evaluation scheme, the dataset is partitioned into training, development, and test sets based on time-dependent criteria, as outlined in Table 1. Given that the dataset is collected over a 2-week period, we allocate the first 70% of the data from each participant to the training set. The subsequent 15% forms the development set, and the remaining 15% constitutes the test set. This partitioning strategy ensures that the evaluation is robust and reflects the temporal dynamics of the data.

Table 1. Data partitioning.

Set	Data (%)	Samples
Training	70	4340
Development	15	931
Test	15	929

Evaluation Metrics

In this study, the concordance correlation coefficient (CCC) is used as a key evaluation metric. CCC considers both the scale and location shifts between the predicted and actual data, providing a comprehensive measure of the model's predictive performance. In our context, this metric is crucial for assessing the accuracy of our model in reflecting the true emotional experiences of participants.

The CCC is defined as follows.

$$\mathrm{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

Where ρ represents the Pearson correlation coefficient between the predicted and actual values, σ_x and σ_y are the SDs of the predicted and actual values, respectively, and μ_x and μ_y are their means.

Benchmark Model

To effectively evaluate our macro-micro framework, it is crucial to establish a benchmark for comparison. This benchmark model consists of the following key components: pure concatenation for modality fusion, 2-layer FFNN for presentation, and equal weight strategy for MTL. This benchmark model, with its straightforward concatenation, basic personalization, and uniform task weighting, provides a solid foundation for comparison.

Pure Concatenation for Modality Fusion

This approach linearly combines features from both physical activity and speech data without weighting or transformation.

2-Layer FFNN for Personalization

We designed FFNN to adapt the concatenated features to get personalized embeddings.

Equal Weight Strategy for MTL

In handling multiple tasks, we applied an equal weight strategy across all tasks.

Comparison Methods

Multimodal Fusion Techniques

In our exploration of multimodal fusion techniques, we first investigated the use of separate transformer embeddings for each modality. This approach aimed to capture unique features within physical activity and speech data independently before combining them. The results indicate that while this method was effective in isolating modality-specific characteristics, it also necessitated sophisticated alignment strategies during the fusion stage.

We also applied attention mechanisms, including solo attention [51] and postconcatenation attention. These techniques allowed the model to dynamically focus on the most informative features from each modality. The solo attention mechanism, applied before concatenation, proved particularly effective in enhancing the model's sensitivity to contextually relevant multimodal cues.

A more straightforward approach, the pure weighted method, involved assigning fixed weights to each modality during fusion. Despite its simplicity, this method displayed limitations in adaptability, especially in scenarios where the relative importance of each modality varied.

Besides, max fusion [52] was used to capture the most significant features across modalities by taking the maximum value across feature dimensions. This method was found to be particularly useful in scenarios where the dominant features in the data were more predictive of the outcome.

We tested gated fusion [53], which was designed for dynamic control over the contribution of each modality based on the data's contextual information. This adaptability resulted in improved performance, especially in complex scenarios where the relevance of each modality changed.

Personalization Strategies

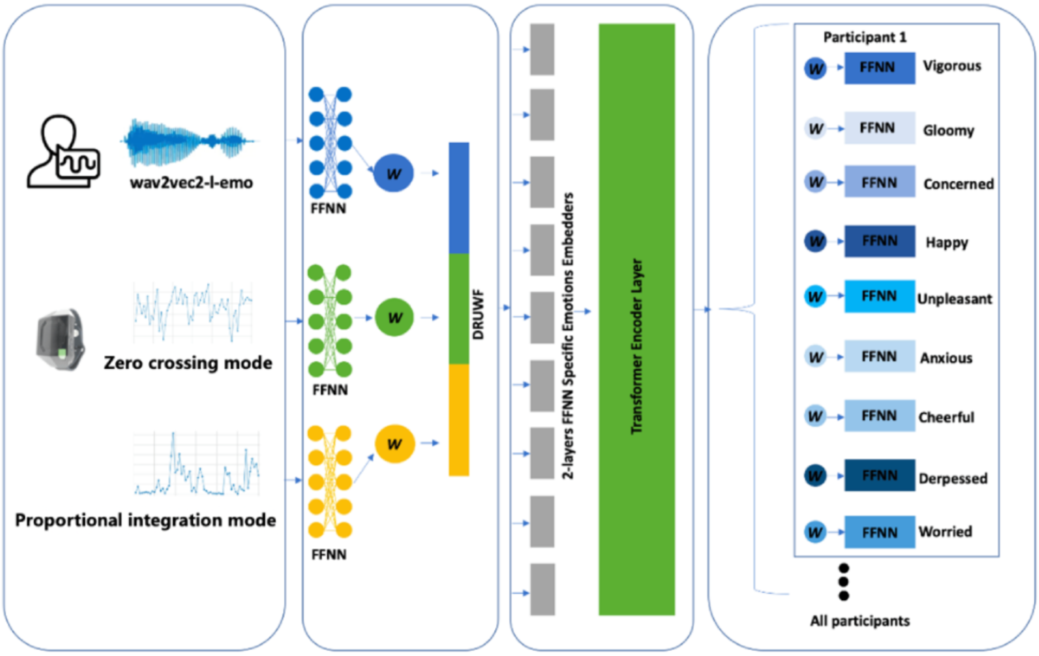
The FFNN was used as a baseline for personalization, adapting concatenated multimodal features to individual profiles. Comparatively, we used a transformer model without separate emotional FFNN for personalization.

We also applied an adapter [54] to compare. The adapter method involved integrating small, trainable modules into a pretrained model. This approach facilitated efficient and effective personalization without the need for extensive retraining.

MTL Approaches

Using a single embedding to produce outputs for multiple tasks with equal weighting provided a baseline for multitask performance. We also compared the pure weighted approach, assigning fixed weights to each task, which showed some improvements over the equal weight strategy but still lacked the dynamic adaptability required for more complex multitask scenarios. The architecture of the proposed model is given in Figure 2.

Figure 2. Macro-micro personalization framework for multimodal-multitask learning in mental health monitoring. W: weight; DRUW: Dynamic Restrained Uncertainty Weighting; DRUWF: Dynamic Restrained Uncertainty Weighting Fusion; FFNN: feed-forward neural network. The framework integrates physiological signals and speech data through feature extraction, DRUWF, macro layers with transformer encoder, and micro personalized layers.



Model Training Procedure

All models in our study were trained over 100 epochs using stochastic gradient descent with an initial learning rate set at 0.001, using a Nesterov momentum of 0.9 to enhance convergence. The learning rate was adaptively reduced by a factor of 0.9 if no improvement was observed on the development set after 5 consecutive epochs, ensuring efficient optimization. The training was conducted with a batch size of 16, balancing computational resources and model performance. Additionally, a weight decay of 0.0001 was applied to prevent overfitting. The final model configuration selected for evaluation on the test set was the one yielding the best performance on the development set, ensuring the reliability and robustness of our results. Before producing the final output, we used a sigmoid function to ensure that the predicted values ranged from 0 to 1. This adjustment was necessary because our labels had been normalized to a scale of 0 to 1.

Ethical Considerations

This study was approved by the Ethics Committee of the University of Tokyo (21-353). The study participants provided written informed consent.

Results

Overview

Our results first show CCC for various emotional dimensions using different single-modal data types (physical activity and audio), with and without personalization in Table 2. Our analysis also investigated the efficacy of various multimodal fusion techniques and their capacity for personalization in assessing different emotional dimensions in Table 3. The multimodal fusion approaches examined included basic, max fusion, gated fusion, attention fusion, solo-attention fusion, cross-modal attention, and a number of proposed methods. Each of these techniques was also analyzed in conjunction with various multitasking frameworks such as basic, transformer, adapter, equal, multioutputs, and our proposed method. In comparison between the 2 tables, multimodal results generally outperform single-modal ones.

The results indicate a differential impact on the CCC across emotional states and fusion methods. For instance, the max fusion approach yielded a CCC of 0.451 for vigorous, which was a notable improvement over the basic approach’s 0.415. However, this method seemed less effective for gloomy, with a CCC of 0.356. In contrast, the gated fusion technique exhibited a more consistent performance across different emotional states, with CCCs ranging from 0.277 for anxious to 0.537 for worried.

Of particular interest were the results from the proposed methods, which showed promising CCC values across several emotional states. The highest recorded CCC of the proposed methods was for worried, with a value of 0.581. Conversely, gloomy showed the lowest CCC at 0.377 using the same proposed methods.

Overall, the mean CCC values across all emotional states suggested that the proposed methods combined with our proposed multitasking framework outperformed the other techniques, achieving a mean CCC of 0.503. This mean value was computed by averaging CCCs across all the emotional states for each method. Notably, the proposed methods consistently yielded CCC values above the overall mean, underscoring their potential for enhancing emotion recognition tasks in multimodal settings.

In conclusion, our results underscore the importance of choosing the appropriate fusion and multitasking methods to maximize the agreement between predicted and actual emotional states. The proposed methods, when tailored for individual emotional dimensions, demonstrate significant promise for personalization, which is a critical aspect of effective emotional state prediction.

Table 2. CCC[a] for various emotional dimensions using different single-modal data types (physical activity and audio), with and without personalization. The emotional dimensions covered are vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed and worried.

Single model and personalization	Vigorous	Gloomy	Concerned	Happy	Unpleasant	Anxious	Cheerful	Depressed	Worried	Mean (SD)[b]
ZCM[c]										
No	0.287	0.187	0.325	0.341	0.364	0.179	0.224	0.282	0.334	0.281 (0.069)
Yes	0.407	0.426	0.499	0.485	0.487	0.349	0.322	0.454	0.535	0.441 (0.071)
PIM[d]										
No	0.130	0.142	0.368	0.251	0.374	0.244	0.006	0.302	0.291	0.225 (0.121)
Yes	0.145	0.315	0.499	0.368	0.489	0.272	0.051	0.493	0.537	0.341 (0.172)
Speech										
No	0.287	0.187	0.325	0.341	0.364	0.179	0.224	0.282	0.334	0.281 (0.069)
Yes	0.407	0.426	0.499	0.485	0.487	0.349	0.322	0.454	0.535	0.441 (0.071)

[a] CCC: concordance correlation coefficient.

[b] The “Mean” column represents the average CCC across the emotional dimensions.

[c] ZCM: zero crossing mode.

[d] PIM: proportional integration mode.

Table 3. CCC[a] for various emotional dimensions, measured using participant-dependent partitions on our dataset. The table also demonstrates the effectiveness of different fusion methods and indicates whether personalization was used. The emotional dimensions covered are vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed, and worried.

Multimodal fusion	Personalization	Multitask	Vigorous	Gloomy	Concerned	Happy	Unpleasant	Anxious	Cheerful	Depressed	Worried	Mean (SD)[b]
Basic	Basic	Basic	0.415	0.404	0.469	0.452	0.499	0.212	0.304	0.472	0.514	0.416 (0.099)
Max fusion	Basic	Basic	0.451	0.356	0.464	0.465	0.501	0.356	0.325	0.421	0.509	0.428 (0.067)
Gated fusion	Basic	Basic	0.417	0.417	0.495	0.473	0.497	0.277	0.317	0.446	0.537	0.431 (0.086)
Attention fusion	Basic	Basic	0.414	0.281	0.524	0.413	0.554	0.386	0.317	0.447	0.574	0.435 (0.102)
Solo-attention fusion	Basic	Basic	0.403	0.337	0.488	0.467	0.513	0.394	0.298	0.470	0.540	0.434 (0.082)
Cross-modal attention	Basic	Basic	0.436	0.443	0.483	0.455	0.497	0.268	0.323	0.463	0.527	0.447 (0.084)
Proposed	Basic	Basic	0.393	0.401	0.504	0.435	0.488	0.469	0.322	0.465	0.556	0.449 (0.069)
Basic	Transformer	Basic	0.434	0.330	0.547	0.470	0.548	0.538	0.325	0.421	0.578	0.466 (0.09)
Basic	Adapter	Basic	0.457	0.329	0.554	0.463	0.549	0.543	0.320	0.432	0.586	0.472 (0.098)
Basic	Proposed	Basic	0.460	0.401	0.554	0.462	0.544	0.557	0.351	0.441	0.585	0.484 (0.080)
Basic	Basic	Equal	0.420	0.321	0.521	0.436	0.511	0.446	0.298	0.381	0.563	0.431 (0.090)
Basic	Basic	Multioutput	0.407	0.399	0.493	0.470	0.526	0.469	0.324	0.467	0.531	0.454 (0.067)
Basic	Basic	Proposed	0.434	0.330	0.547	0.470	0.548	0.538	0.325	0.422	0.579	0.466 (0.095)
Proposed	Proposed	Basic	0.414	0.389	0.564	0.522	0.556	0.581	0.364	0.419	0.589	0.489 (0.091)
Basic	Proposed	Proposed	0.454	0.450	0.549	0.519	0.554	0.581	0.358	0.424	0.583	0.497 (0.079)
Proposed	Basic	Proposed	0.449	0.377	0.525	0.466	0.554	0.573	0.296	0.437	0.543	0.469 (0.091)
Proposed	Proposed	Proposed	0.464	0.464	0.549	0.538	0.554	0.581	0.373	0.420	0.581	0.503 (0.075)

[a] CCC: concordance correlation coefficient.

[b] The “Mean” column represents the average CCC across the emotional dimensions.

Statistical Validation

We also conducted a statistical analysis to complement the CCC results from our deep learning model, crucial for validating the model’s reliability and generalizability across different datasets and conditions. Our mixed linear model analysis [55], presented in Table 4, reveals 2 critical insights. First, the highly significant within-individual associations (Q) across 9 emotional scales underscore the model’s capability to capture nuanced emotional responses, indicating its robust predictive power. Second, the observation of group and residual variances highlights the variability that the model does not account for, signaling areas that require further refinement. This unexplained variability invites a deeper investigation into potential factors, such as the model’s sensitivity to specific data characteristics or the necessity for incorporating a more diverse range of training data. Understanding these elements can guide targeted improvements in the model’s architecture and training process, ultimately enhancing its accuracy and applicability in personalized mental health monitoring.

Table 4. Mixed linear model regression results for emotional dimensions[a].

Emotional dimension	Intercept	Regression coefficient	SE	*z* score	*P* value>|*z*|	95% CI	Group variance	Residual variance
Vigorous	0.155	0.549	0.076	7.176	<.001	0.399-0.698	0.039	0.057
Gloomy	0.087	0.607	0.064	9.512	<.001	0.482-0.733	0.023	0.067
Concerned	0.153	0.500	0.073	6.813	<.001	0.356-0.644	0.043	0.057
Happy	0.236	0.446	0.077	5.771	<.001	0.295-0.598	0.030	0.064
Unpleasant	0.071	0.630	0.066	9.534	<.001	0.501-0.760	0.028	0.061
Anxious	0.127	0.577	0.071	8.177	<.001	0.439-0.716	0.035	0.061
Cheerful	0.210	0.477	0.074	6.475	<.001	0.333-0.621	0.022	0.072
Depressed	0.044	0.647	0.060	10.758	<.001	0.529-0.765	0.021	0.062
Worried	0.138	0.539	0.076	7.097	<.001	0.390-0.688	0.048	0.052

[a] The table summarizes the intercept, regression coefficients, standard errors, *z* scores, *P* values, CIs, group variances, and residual variances for each emotion studied. The emotional dimensions covered are vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed, and worried.

Discussion

Overview

This study introduces a novel dataset and a macro-micro framework for personalized daily mental health monitoring, leveraging multimodal and MTL strategies. The results demonstrate the efficacy of our approach in predicting emotional states, with a mean CCC of 0.503 across 9 emotional dimensions.

The proposed macro-micro framework, which combines macro-level emotion transformer embeddings with micro-level personalization layers, shows superior performance compared to traditional approaches. This suggests that incorporating both general emotional patterns and individual-specific adaptations is crucial for accurate mental health monitoring. The effectiveness of our DRUW fusion method in integrating multimodal data further underscores the importance of adaptive weighting strategies in handling diverse data types.

Our findings align with previous studies highlighting the potential of multimodal approaches in mental health monitoring. However, our work extends beyond existing research by incorporating personalization at both macro and micro levels, addressing a critical gap in current mental health technology.

The high significance of within-individual associations across emotional scales, as revealed by our mixed linear model analysis, validates the model’s capability to capture nuanced emotional responses. This has important implications for the development of personalized mental health interventions, as it suggests that our model can detect subtle changes in an individual’s emotional state over time.

However, the observed group and residual variances in our statistical analysis indicate that there is still unexplained variability in emotional states. This highlights a limitation of our model and suggests that additional factors, not captured in our framework, may influence daily emotional states. These could include external stressors, social interactions, or physiological factors not measured in our study.

Another limitation is the reliance on self-reported emotional states, which, while valuable for capturing subjective experiences, may be subject to reporting biases. Future research could explore the integration of objective measures of emotional state, such as facial expression analysis or additional physiological markers, to complement self-reports.

Looking ahead, several avenues for future research emerge from our findings. First, expanding the dataset to include a more diverse range of participants and longer monitoring periods could enhance the generalizability of our model. Second, investigating the incorporation of additional modalities, such as sleep patterns or social media activity, could provide a more comprehensive picture of mental health. Finally, exploring the application of our framework in clinical settings could help bridge the gap between research and practical mental health interventions.

Conclusions

In conclusion, this study introduces a groundbreaking dataset and a macro-micro framework that significantly advances personalized daily mental health monitoring. By leveraging multimodal and MTL strategies, we have demonstrated a robust model capable of predicting emotional states. The statistical analysis further validates the model’s reliability, highlighting its potential for wider application in the mental health domain.

Moving forward, our focus will be on expanding the dataset, incorporating additional modalities, and refining our model to address these variances, with the ultimate goal of making daily mental health monitoring a more accessible, nonintrusive, and personalized practice.

Acknowledgments

The authors acknowledge support by the Japan Science and Technology Agency (JST) Mirai Program (grant 21473074), JST FORESTO (grant JPMJFR2156) Program, and the JST MOONSHOT Program (grant JPMJMS229B).

Conflicts of Interest

None declared.

References

1. Rosenfield S, Mouzon D. Gender and mental health. In: Aneshensel CS, Phelan JC, Bierman A, editors. Handbook of the Sociology of Mental Health. Dordrecht, Netherlands. Springer; 2013:277-296.
2. Hou W, Lai FT, Ben-Ezra M, Goodwin R. Regularizing daily routines for mental health during and after the COVID-19 pandemic. J Glob Health. 2020;10(2):020315. [FREE Full text] [doi: 10.7189/jogh.10.020315] [Medline: 33110518]
3. Laird KT, Tanner-Smith EE, Russell AC, Hollon SD, Walker LS. Comparative efficacy of psychological therapies for improving mental health and daily functioning in irritable bowel syndrome: a systematic review and meta-analysis. Clin Psychol Rev. 2017;51:142-152. [FREE Full text] [doi: 10.1016/j.cpr.2016.11.001] [Medline: 27870997]
4. Song M, Triantafyllopoulos A, Yang Z, Takeuchi H, Nakamura T, Kishi A, et al. Daily mental health monitoring from speech: a real-world Japanese dataset and multitask learning analysis. 2023. Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); June 10, 2023:1-5; Rhodes, Greece. [doi: 10.1109/icassp49357.2023.10096884]
5. Tripathi S, Beigi H. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. arXiv. Preprint posted online on April 16, 2018. [doi: 10.48550/arXiv.1804.05788]
6. Palotti J, Narula G, Raheem L, Bay H. Analysis of emotion annotation strength improves generalization in speech emotion recognition models. 2023. Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 17, 2023:5828-5836; Vancouver, BC, Canada. [doi: 10.1109/cvprw59228.2023.00619]
7. Fan W, Xu X, Xing X, Chen W, Huang D. LSSED: a large-scale dataset and benchmark for speech emotion recognition. 2021. Presented at: ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); June 06, 2021:641-645; Toronto, ON, Canada. [doi: 10.1109/icassp39728.2021.9414542]
8. Bertero D, Fung P. A first look into a convolutional neural network for speech emotion detection. 2017. Presented at: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); March 05, 2017:5115-5119; New Orleans, LA. [doi: 10.1109/icassp.2017.7953131]
9. Truong KP, van Leeuwen DA, de Jong FMG. Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Commun. 2012;54(9):1049-1063. [doi: 10.1016/j.specom.2012.04.006]
10. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol. 2013;9:151-176. [FREE Full text] [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
11. Suhaimi NS, Mountstephens J, Teo J. EEG-based emotion recognition: a state-of-the-art review of current trends and opportunities. Comput Intell Neurosci. 2020;2020:8875426. [FREE Full text] [doi: 10.1155/2020/8875426] [Medline: 33014031]
12. Jenke R, Peer A, Buss M. Feature extraction and selection for emotion recognition from EEG. IEEE Trans Affective Comput. 2014;5(3):327-339. [doi: 10.1109/taffc.2014.2339834]
13. Ko BC. A brief review of facial emotion recognition based on visual information. Sensors. 2018;18(2):401. [FREE Full text] [doi: 10.3390/s18020401] [Medline: 29385749]
14. Canal FZ, Müller TR, Matias JC, Scotton GG, de Sa Junior AR, Pozzebon E, et al. A survey on facial emotion recognition techniques: a state-of-the-art literature review. Inf Sci. 2022;582:593-617. [doi: 10.1016/j.ins.2021.10.005]
15. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. 2011;44(3):572-587. [doi: 10.1016/j.patcog.2010.09.020]
16. Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E. A comprehensive review of speech emotion recognition systems. IEEE Access. 2021;9:47795-47814. [doi: 10.1109/access.2021.3068045]
17. Massaro S, Pecchia L. Heart rate variability (HRV) analysis: a methodology for organizational neuroscience. Organizational Res Methods. 2016;22(1):354-393. [doi: 10.1177/1094428116681072]
18. Kemp AH, Quintana DS, Gray MA, Felmingham KL, Brown K, Gatt JM. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. Biol Psychiatry. 2010;67(11):1067-1074. [doi: 10.1016/j.biopsych.2009.12.012] [Medline: 20138254]
19. Beauchaine TP, Thayer JF. Heart rate variability as a transdiagnostic biomarker of psychopathology. Int J Psychophysiol. 2015;98(2):338-350. [doi: 10.1016/j.ijpsycho.2015.08.004] [Medline: 26272488]

20. Kim J, Nakamura T, Kikuchi H, Yoshiuchi K, Sasaki T, Yamamoto Y. Covariation of depressive mood and spontaneous physical activity in major depressive disorder: toward continuous monitoring of depressive mood. IEEE J Biomed Health Inform. 2015;19(4):1347-1355. [doi: 10.1109/JBHI.2015.2440764] [Medline: 26054079]
21. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep. 2003;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
22. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
23. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry. 2012;72(7):580-587. [FREE Full text] [doi: 10.1016/j.biopsych.2012.03.015] [Medline: 22541039]
24. Cohen AS, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. Curr Opin Psychiatry. 2014;27(3):203-209. [FREE Full text] [doi: 10.1097/YCO.0000000000000056] [Medline: 24613984]
25. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health. 2016;3(2):e16. [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
26. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol. 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
27. Fahrenberg J, Myrtek M, Pawlik K, Perrez M. Ambulatory assessment—monitoring behavior in daily life settings. Eur J Psychol Assess. 2007;23(4):206-213. [doi: 10.1027/1015-5759.23.4.206]
28. Kendall M, Stuart A. The Advance Theory of Statistics. London. Charles Griffin and Company, Ltd.; 1943.
29. Song M, Yang Z, Triantafyllopoulos A, Jing X, Karas V, Xie J, et al. Dynamic restrained uncertainty weighting loss for multitask learning of vocal expression. arXiv. Preprint posted online on June 22, 2022. [doi: 10.48550/arXiv.2206.11049]
30. Dagum P. Digital biomarkers of cognitive function. NPJ Digit Med. 2018;1:10. [FREE Full text] [doi: 10.1038/s41746-018-0018-4] [Medline: 31304295]
31. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
32. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health. 2020;7(3):e18848. [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
33. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry. 2018;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
34. Hickey BA, Chalmers T, Newton P, Lin CT, Sibbritt D, McLachlan CS, et al. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: a systematic review. Sensors. 2021;21(10):3461. [FREE Full text] [doi: 10.3390/s21103461] [Medline: 34065620]
35. Long N, Lei Y, Peng L, Xu P, Mao P. A scoping review on monitoring mental health using smart wearable devices. Math Biosci Eng. 2022;19(8):7899-7919. [FREE Full text] [doi: 10.3934/mbe.2022369] [Medline: 35801449]
36. Xu H, Wu X, Liu X. A measurement method for mental health based on dynamic multimodal feature recognition. Front Public Health. 2022;10:990235. [FREE Full text] [doi: 10.3389/fpubh.2022.990235] [Medline: 36620271]
37. Huckins JF, daSilva AW, Wang W, Hedlund E, Rogers C, Nepal SK, et al. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res. 2020;22(6):e20185. [FREE Full text] [doi: 10.2196/20185] [Medline: 32519963]
38. Das R, Hasan MR, Daria S, Islam MR. Impact of COVID-19 pandemic on mental health among general Bangladeshi population: a cross-sectional study. BMJ Open. 2021;11(4):e045727. [FREE Full text] [doi: 10.1136/bmjopen-2020-045727] [Medline: 33837107]
39. Morshed MB, Saha K, Li R, D'Mello SK, De Choudhury M, Abowd GD, et al. Prediction of mood instability with passive sensing. 2019. Presented at: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; September 09, 2019:1-21; New York, NY. [doi: 10.1145/3351233]
40. Takeuchi H, Suwa K, Kishi A, Nakamura T, Yoshiuchi K, Yamamoto Y. The effects of objective push-type sleep feedback on habitual sleep behavior and momentary symptoms in daily life: mHealth intervention trial using a health care internet of things system. JMIR Mhealth Uhealth. 2022;10(10):e39150. [FREE Full text] [doi: 10.2196/39150] [Medline: 36201383]
41. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health. 2020;23(4):161-166. [FREE Full text] [doi: 10.1136/ebmental-2020-300180] [Medline: 32998937]
42. Gerczuk M, Triantafyllopoulos A, Amiriparian S, Kathan A, Bauer J, Berking M, et al. Zero-shot personalization of speech foundation models for depressed mood monitoring. Patterns. 2023;4(11):100873. [FREE Full text] [doi: 10.1016/j.patter.2023.100873] [Medline: 38035199]
43. Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry. 2020;10(1):116. [FREE Full text] [doi: 10.1038/s41398-020-0780-3] [Medline: 32532967]
44. Fukui I. The Depression and Anxiety Mood Scale (DAMS): scale development and validation. Jpn J Behav Ther. 1997;23:83-93. [FREE Full text] [doi: 10.24468/jjbt.23.2_83]

45. Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. 2017. Presented at: Interspeech 2017; August 10, 2017:1089-1093; Stockholm, Sweden. [doi: 10.21437/interspeech.2017-200]
46. Burton C, McKinstry B, Tătar AS, Serrano-Blanco A, Pagliari C, Wolters M. Activity monitoring in patients with depression: a systematic review. J Affect Disord. 2013;145(1):21-28. [FREE Full text] [doi: 10.1016/j.jad.2012.07.001] [Medline: 22868056]
47. Helgadóttir B, Forsell Y, Ekblom Ö. Physical activity patterns of people affected by depressive and anxiety disorders as measured by accelerometers: a cross-sectional study. PLoS One. 2015;10(1):e0115894. [FREE Full text] [doi: 10.1371/journal.pone.0115894] [Medline: 25585123]
48. Loizou PC. Speech Enhancement: Theory and Practice. 2nd Edition. Boca Raton, FL. CRC Press; 2013.
49. Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Burkhardt F, Eyben F, et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Trans Pattern Anal Mach Intell. 2023;45(9):10745-10759. [doi: 10.1109/TPAMI.2023.3263585] [Medline: 37015129]
50. Zhang Y, Wei Y, Yang Q. Learning to multitask. 2018. Presented at: Advances in Neural Information Processing Systems; November 21, 2001:5771-5782; Cambridge, MA. [doi: 10.7551/mitpress/11474.003.0014]
51. Song M, Yang Z, Triantafyllopoulos A, Nakamura T, Zhang Y, Ren Z, et al. Crossmodal transformer on multi-physical signals for personalised daily mental health prediction. 2023. Presented at: IEEE ICDM 2023—23rd IEEE International Conference on Data Mining; January 11, 2023:1299-1305; Shanghai, China. [doi: 10.1109/icdmw60847.2023.00167]
52. Kuncheva LI. A theoretical study on six classifier fusion strategies. IEEE Trans Pattern Anal Machine Intell. 2002;24(2):281-286. [doi: 10.1109/34.982906]
53. Ren W, Ma L, Zhang J, Pan J, Cao X, Liu W, et al. Gated fusion network for single image dehazing. 2018. Presented at: CVPR 2018; June 05, 2018:3253-3261; Salt Lake City, UT. [doi: 10.1109/cvpr.2018.00343]
54. Dong W, Yan D, Lin Z, Wang P. Efficient adaptation of large vision transformer via adapter re-composing. 2023. Presented at: 2023 Conference on Neural Information Processing Systems (NeurIPS 2023); December 10, 2023; New Orleans, LA. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html
55. Kalaian HA, Raudenbush SW. A multivariate mixed linear model for meta-analysis. Psychol Methods. 1996;1(3):227-235. [doi: 10.1037//1082-989x.1.3.227]

Abbreviations

CCC: concordance correlation coefficient
DAMS: Depression and Anxiety Mood Scale
DRUW: Dynamic Restrained Uncertainty Weighting
EEG: electroencephalography
EMA: ecological momentary assessment
FFNN: feed-forward neural network
HRV: heart rate variability
MHIT: Mental Healthcare Internet of Things
MTL: multitask learning
PIM: proportional integration mode
ZCM: zero crossing mode.

Edited by J Torous; submitted 14.04.24; peer-reviewed by M Bagewadi Ellur, L Henry, X Wu; comments to author 22.05.24; revised version received 16.07.24; accepted 03.08.24; published 18.10.24

Please cite as:
Song M, Yang Z, Triantafyllopoulos A, Zhang Z, Nan Z, Tang M, Takeuchi H, Nakamura T, Kishi A, Ishizawa T, Yoshiuchi K, Schuller B, Yamamoto Y
Empowering Mental Health Monitoring Using a Macro-Micro Personalization Framework for Multimodal-Multitask Learning: Descriptive Study
JMIR Ment Health 2024;11:e59512
URL: https://mental.jmir.org/2024/1/e59512
doi: 10.2196/59512
PMID: 39422993