Empowering Mental Health Monitoring: Macro-Micro Personalization Framework for Multimodal-Multitask Learning

Song Meishu, Zijiang Yang, Andreas Triantafyllopoulos, Zixing Zhang, Hiroki Takeuchi, Toru Nakamura, Akifumi Kishi, Tetsuro Ishizawa, Kazuhiro Yoshiuchi, Bjoern Schuller, Yamamoto Yoshiharu

Submitted to: JMIR Mental Health
on: April 14, 2024

Table of Contents

Empowering Mental Health Monitoring: Macro-Micro Personalization Framework for Multimodal-Multitask Learning

Song Meishu[1]; Zijiang Yang[1]; Andreas Triantafyllopoulos[2]; Zixing Zhang[3]; Hiroki Takeuchi[1]; Toru Nakamura[4]; Akifumi Kishi[1]; Tetsuro Ishizawa[1]; Kazuhiro Yoshiuchi[1]; Bjoern Schuller[5]; Yamamoto Yoshiharu[1]

[1]The University of Tokyo Tokyo JP
[2]Technical University of Munich Munich DE
[3]Hunan University Changsha CN
[4]Osaka University Osaka JP
[5]Imperial College London London GB

Corresponding Author:
Yamamoto Yoshiharu
The University of Tokyo
Tokyo
Tokyo
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 non-invasive devices like wristbands and smartphones are capable of collecting a wide range of data, which has not yet been fully utilized for mental health monitoring.

Objective: The paper aims to introduce a novel dataset for Personalized Daily Mental Health Monitoring and a new Macro-Micro Framework. This framework is designed to employ multimodal and multitask learning strategies for improved personalization and prediction of emotional states in individuals.

Methods: Data was collected from 242 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 (CCC), resulting in a score of 0.503. This result demonstrates the framework's efficacy in predicting emotional states.

Conclusions: The paper 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 application, opening up new avenues for technology-based mental health interventions.

(JMIR Preprints 14/04/2024:59512)
DOI: https://doi.org/10.2196/preprints.59512

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Meishu Song, Zijiang Yang, Andreas Triantafyllopoulos, Zixing Zhang, Björn W. Schuller, Yoshiharu Yamamoto

Empowering Mental Health Monitoring: Macro-Micro Personalization Framework for Multimodal-Multitask Learning

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 non-invasive devices like wristbands and smartphones are capable of collecting a wide range of data, which has not yet been fully utilized for mental health monitoring.
Objective: The paper aims to introduce a novel dataset for Personalized Daily Mental Health Monitoring and a new Macro-Micro Framework. This framework is designed to employ multimodal and multitask learning strategies for improved personalization and prediction of emotional states in individuals.
Methods: Data was collected from 242 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 (CCC), resulting in a score of 0.503. This result demonstrates the framework's efficacy in predicting emotional states.
Conclusions: The paper 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 application, opening up new avenues for technology-based mental health interventions.

Keywords: Multimodal; Multitask; Daily Mental Health; Emotion

Introduction

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, utilizing 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:

1. **Real-World Representation**: A significant portion of existing datasets lack data derived from real-world settings, instead relying on artificial or laboratory conditions.

2. **Lack of Self-Annotation**: Many datasets do not employ self-annotation [5, 6, 7, 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].

3. **Challenges in Accessibility of Monitoring Data**: Many research utilizes Electroencephalogram (EEG) [11, 12] while providing valuable insights into brain activity and emotional states, requires 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.

4. **Limited Modalities and Single-Model Approach**: Most available research focus 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:

1. **Introducing a noval dataset**: Collected from 242 individuals using non-invasive, everyday devices including wristbands and smartphones, our dataset captures physiological signals: Zero Crossing Mode(ZCM) and Proportional Integration Mode(PIM) and speech data. Participants provided self-annotated emotional states over two weeks, creating a rich, multimodal resource for understanding daily mental health dynamics.

2. **Developing a Macro-Micro Framework for Personalized Daily Mental Health**: Our framework develops a multimodal and multitask learning strategy, innovatively built global emotion embeddings with an individual personalization embeddings.

In our research, the decision to focus on physiological signals and speech data, while excluding modalities like 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 employs two 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: Firstly, 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]. Secondly, 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 non-invasively 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 employed 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 (i.e., the act of self-reporting influencing the very experiences being reported) and compliance issues [27]. In our 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:

1. 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].

2. Personalized treatment and monitoring: Insights from our study can inform personalized treatment plans and monitoring strategies, tailoring interventions to individual needs [31].
3. 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].
4. 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 related work section delves into various aspects of mental health monitoring research domain, including mental health data, Ecological Momentary Assessment, 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 [34] and others [35] have critically evaluated the utility 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 (EDA), 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 H. 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 our 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 [42], employing 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 AI techniques represent promising avenues for overcoming these challenges and improving mental health monitoring and diagnosis.

Methodology and Data Collection

The study followed a two-week data collection protocol, during which participants wore wrist-worn devices and used a smartphone application 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 application. To collect data, we developed a platform called Mental Healthcare Internet of Things (MHIT) system. The study procedure involved the following steps: participant recruitment, orientation and consent, data collection, data preprocessing and analysis.

MHIT

The MHIT system 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 two key components: a cloud server (MHIT Server) and a smartphone application (MHIT App).

Participants

A convenience sample of 242 Japanese office workers from an insurance company participated in our study. The participants had a mean age of 42.32 years with a standard deviation (SD) of 7.80 years. All participants were working from home during the study period, as mandated by their employer to mitigate the spread of COVID-19.

Participants attended an orientation session where they were informed about the study objectives, procedures, and the use of the MHIT system. They provided written informed consent and had the opportunity to ask questions about the study. Participants were trained on how to use the wrist-worn devices and the MHIT App. They received instructions on wearing the devices properly, recording their speech, and reporting their emotional states using the Depression and Anxiety Mood Scale (DAMS).

Annotation and Ecological Momentary Assessment (EMA)

Our study employed 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 using the DAMS five times daily at random intervals, with a minimum of 2 hours between each prompt. These prompts were delivered via the MHIT App.
- Data Collection: At each EMA prompt, participants self-reported the intensity of nine different expressed emotions on a [0:100] visual-analogue scale (slider) within the MHIT App. These nine emotions correspond to the items of the DAMS. 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, “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 zero 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 one minute, aggregating 60 data points (representing one 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 integral for comprehensively assessing mentalalth experiences pertinent to depression and anxiety. DAMS’s effectiveness in measuring depressive and anxious moods is particularly notable, as it employs 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 utility in detecting mood changes. These features of DAMS, combined with its thorough statistical analysis across nine 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 several important reasons:

1. **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.

2. **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.

Using the MHIT App, participants self-reported their emotional states five times daily over two weeks, employing the nine emotional states outlined in DAMS.

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.

1. **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 data set.

2. **Voice Activity Detection**: We employed algorithms to detect and eliminate silences in voice recordings. This focus on active speech segments helped in isolating meaningful data.

3. **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

Signal Cleaning: Similar to speech data, physical activity data was cleaned to remove any erroneous signals.
Signal Standardization: The raw data from the physical activity sensors were standardized to ensure consistency across different devices and 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 towards 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:

1. **How to Fuse Different Modalities**: Specifically, how do we integrate physical activities and speech data?

2. **How to Achieve Personalization**: What strategies can we employ to tailor the analysis to individual participants?

3. **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: multimodal fusion initially integrates and aligns 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.

Our Framework

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

The fusion of these data streams is executed through a Dynamic Restrained Uncertainty Weighting Fusion (DRUWF) block, effectively merging the standardized features from physiological signals and speech. This fusion process not only integrates the data but also applied DRUWF.

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, **Marco** space.

Personalization is introduced at the **Micro** stage. Here, the framework employs additional FFNN Layers tailored to individual users, enabling the selection of embedding elements pertinent to their unique emotional profiles. This adaptation leverages Dynamic Restrained Uncertainty Weighting Loss.

Dynamic Restrained Uncertainty Weighting 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:

$$F(w,\beta_1,\beta_2)=\frac{1}{\beta_1^2}M_1(w)+\frac{1}{\beta_2^2}M_2(w)$$

$$+\log\left(1+\log\beta_1^2\right)+\log\left(1+\log\beta_2^2\right)$$

$$+\left|\phi-\left(\left|\log\beta_1\right|+\left|\log\beta_2\right|\right)\right|$$

In this equation, β_1 and β_2 are the uncertainty parameters for the physical activity data $M_1(w)$ and the speech data $M_2(w)$, 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.

1. **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.

2. **Micro Personalization Space**: After Macro space, Micro layer is designed for each participant. This layer allows for the customization of the model based on micro-specific data. It adapts the general insights from the macro space to the nuances of each participant's emotional profile.

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:

$$P_i = \lambda \cdot E_{macro} + (1-\lambda) \cdot E_{micro,i}$$

In this equation:

1. P_i represents the personalized output for the i^{th} participant.

2. E_{macro} denotes the embeddings or features extracted from the macro emotional space.

3. $E_{micro,i}$ represents the embeddings or features specific to the i^{th} participant.

4. λ 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 like the diversity of the dataset or the specificity of the micro data.

Dynamic Restrained Uncertainty Weighting Loss For Multitask

Multitask Learning (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 employ 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:

$$L(w,\alpha_1,\alpha_2)=\frac{1}{\alpha_1^2}L_1(w)+\frac{1}{\alpha_2^2}L_2(w)$$

$$+\log\left(1+\log\alpha_1^2\right)+\log\left(1+\log\alpha_2^2\right)$$

$$+\left|\varphi-\left(\left|\log\alpha_1\right|+\left|\log\alpha_2\right|\right)\right|$$

In this equation, α_1 and α_2 are uncertainty parameters corresponding to different tasks in our model, with $L_1(w)$ and $L_2(w)$ 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 two-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	Percentage of Data	Number of Samples
Training	70%	4,340
Development	15%	931
Test	15%	929

Evaluation Metrics

In our study, the Concordance Correlation Coefficient (CCC) is employed 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:

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

where ρ represents the Pearson Correlation Coefficient between the predicted and actual values, σ_x and σ_y are the standard deviations 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:

1. **Pure Concatenation for Modality Fusion**: This approach linearly combines features from both physical activity and speech data without weighting or transformation.

2. **Two Layer FFNN for Personalization**: We designed FFNN to adapt the concatenated features to get personalized embeddings.

3. **Equal Weight Strategy for Multitask Learning**: In handling multiple tasks, we applies an equal weight strategy across all tasks.

This benchmark model, with its straightforward concatenation, basic personalization, and uniform task weighting, provides a solid foundation for comparison.

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 indicated 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 post-concatenation 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 utilized 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 employed 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 pre-trained model. This approach facilitated efficient and effective personalization without the need for extensive retraining.

Multitask Learning Approaches

Utilizing 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, showed some improvements over the equal weight strategy but still lacked the dynamic adaptability required for more complex multitask scenarios.

Model Training Procedure

All models in our study were trained over 100 epochs using Stochastic Gradient Descent (SGD) with an initial learning rate set at 0.001, employing 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 five consecutive epochs, ensuring efficient optimization. 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 utilized 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.

Results

Our analysis investigated the efficacy of various multimodal fusion techniques and their capacity for personalization in assessing different emotional dimensions. The multimodal fusion approaches examined included *Basic*, *Max Fusion*, *Gated Fusion*, *Attention Fusion*, *Solo-Attention Fusion*, *Crossmodal 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*, *Multi-Outputs*, and our proposed method.

Table 2. The table demonstrates CCC for various emotional dimensions using different single-modal data types (Audio, Visual, and Text), with and without **Per**sonalization. The emotional dimensions covered are **Vig**orous, **Glo**omy, **Con**cerned, **Hap**py, **Unp**leasant, **Anx**ious, **Che**erful, **Dep**ressed and **Wor**ried. The “Mean” column represents the average CCC across these dimensions.

Single Model	Per	Vig	Glo	Con	Hap	Unp	Anx	Che	Dep	Wor	*Mean*
ZCM	No	.287	.187	.325	.341	.364	.179	.224	.282	.334	.281
	Yes	.407	.426	.499	.485	.487	.349	.322	.454	.535	.441
PCM	No	.130	.142	.368	.251	.374	.244	.006	.302	.291	.225
	Yes	.145	.315	.499	.368	.489	.272	.051	.493	.537	.341
Speech	No	.287	.187	.325	.341	.364	.179	.224	.282	.334	.281
	Yes	.407	.426	.499	.485	.487	.349	.322	.454	.535	.441

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 from 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.

Table 3. The table presents CCC 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 personalisation was used. The emotional dimensions covered are **Vig**orous, **Glo**omy, **Con**cerned, **Hap**py, **Unp**leasant, **Anx**ious, **Che**erful, **Dep**ressed and **Wor**ried. The "Mean" column represents the average CCC across these dimensions.

Multimodal Fusion	**Personalisation**	**Multitask**	**Vig**	**Glo**	**Con**	**Hap**	**Unp**	**Anx**	**Che**	**Dep**	**Wor**	***Mean***
Basic	Basic	Basic	.415	.404	.469	.452	.499	.212	.304	.472	.514	.416
Max Fusion	Basic	Basic	.451	.356	.464	.465	.501	.356	.325	.421	.509	.428
Gated Fusion	Basic	Basic	.417	.417	.495	.473	.497	.277	.317	.446	.537	.431
Attention Fusion	Basic	Basic	.414	.281	.524	.413	.554	.386	.317	.447	.574	.435
Solo-Attention Fusion	Basic	Basic	.403	.337	.488	.467	.513	.394	.298	.470	.540	.434
Crossmodal Attention	Basic	Basic	.436	.443	.483	.455	.497	.268	.323	.463	.527	.447
Proposed	Basic	Basic	.393	.401	.504	.435	.488	.469	.322	.465	.556	.449
Basic	Transformer	Basic	.434	.330	.547	.470	.548	.538	.325	.421	.578	.466
Basic	Adapter	Basic	.457	.329	.554	.463	.549	.543	.320	.432	.586	.472
Basic	Proposed	Basic	.460	.401	.554	.462	.544	.557	.351	.441	.585	.484
Basic	Basic	Equal	.420	.321	.521	.436	.511	.446	.298	.381	.563	.431
Basic	Basic	Multi-Output	.407	.399	.493	.470	.526	.469	.324	.467	.531	.454
Basic	Basic	Proposed	.434	.330	.547	.470	.548	.538	.325	.422	.579	.466
Proposed	Proposed	Basic	.414	.389	.564	.522	.556	.581	.364	.419	.589	.489
Basic	Proposed	Proposed	.454	.450	.549	.519	.554	.581	.358	.424	.583	.497
Proposed	Basic	Proposed	.449	.377	.525	.466	.554	.573	.296	.437	.543	.469
Proposed	Proposed	Proposed	.464	.464	.549	.538	.554	.581	.373	.420	.581	.503

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 4. Mixed Linear Model Regression Results for Emotional Dimensions. The table summarizes the **Int**ercept, Regression **Coef**ficients, **Sta**ndard **Err**ors, **z**-values, **p**-values, Confidence Intervals, Group **Var**iances, and **Res**idual **Var**iances for each emotion studied. The emotional dimensions covered are **Vig**orous, **Glo**omy, **Con**cerned, **Hap**py, **Unp**leasant, **Anx**ious, **Che**erful, **Dep**ressed and **Wor**ried

Emo.	Int.	Coef.	Std. Err.	z	P>|z|	[0.025	0.975]	Group Var.	Res. Var.
Vig	.155	.549	.076	7.176	< 0.001	.399	.698	.039	.057
Glo	.087	.607	.064	9.512	< 0.001	.482	.733	.023	.067
Con	.153	.500	.073	6.813	< 0.001	.356	.644	.043	.057
Hap	.236	.446	.077	5.771	< 0.001	.295	.598	.030	.064
Unp	.071	.630	.066	9.534	< 0.001	.501	.760	.028	.061
Anx	.127	.577	.071	8.177	< 0.001	.439	.716	.035	.061
Che	.210	.477	.074	6.475	< 0.001	.333	.621	.022	.072
Dep	.044	.647	.060	10.758	< 0.001	.529	.765	.021	.062
Wor	.138	.539	.076	7.097	< 0.001	.390	.688	.048	.052

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, presented in Table 4, reveals two critical insights. Firstly, the highly significant within-individual associations (Q) across nine emotional scales underscore the model’s capability to capture nuanced emotional responses, indicating its robust predictive power. Secondly, 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.

Discussion

Our study introduces a novel dataset and a Macro-Micro Framework for personalized daily mental health monitoring, leveraging multimodal and multitask learning strategies. The results demonstrate the efficacy of our approach in predicting emotional states, with a mean CCC of 0.503 across nine 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 current model and suggests that additional factors, not captured in our current 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.

Conclusion

In conclusion, our study introduces a groundbreaking dataset and a Macro-Micro Framework that significantly advances personalized daily mental health monitoring. By leveraging multimodal and multitask learning 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, non-intrusive, and personalized practice.

References

[1] S. Rosenfield and D. Mouzon, "Gender and mental health," *Handbook of the Sociology of Mental Health*, pp. 277-296, 2013.

[2] W. K. Hou, F. T. T. Lai, M. Ben-Ezra and R. Goodwin, "Regularizing daily routines for mental health during and after the COVID-19 pandemic," *Journal of Global Health*, vol. 10, no. 2, 2020.

[3] K. T. Laird, E. E. Tanner-Smith, A. C. Russell, S. D. Hollon and L. S. Walker, "Comparative efficacy of psychological therapies for improving mental health and daily functioning in irritable bowel syndrome: A systematic review and meta-analysis," *Clinical Psychology Review*, vol. 51, pp. 142-152, 2017.

[4] M. Song, A. Triantafyllopoulos, Z. Yang, H. Takeuchi, T. Nakamura, A. Kishi, T. Ishizawa, K. Yoshiuchi, X. Jing, V. Karas, et al., "Daily mental health monitoring from speech: A real-world Japanese dataset and multitask learning analysis," in *Proc. ICASSP*, Rhode Island, Greece, pp. 1-5.

[5] S. Tripathi, S. Tripathi and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[6] J. Palotti, G. Narula, L. Raheem and H. Bay, "Analysis of emotion annotation strength improves generalization in speech emotion recognition models," in *Proc. CVPR*, Vancouver, BC, Canada, 2023, pp. 5828-5836.

[7] W. Fan, X. Xu, X. Xing, W. Chen and D. Huang, "LSSED: A large-scale dataset and benchmark for speech emotion recognition," in *Proc. ICASSP*, Toronto, ON, Canada, 2021, pp. 641-645.

[8] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 5115-5119.

[9] K. P. Truong, D. A. Van Leeuwen and F. M. G. De Jong, "Speech-based recognition of self-reported and observed emotion in a dimensional space," *Speech Communication*, vol. 54, no. 9, pp. 1049-1063, 2012.

[10] T. J. Trull and U. Ebner-Priemer, "Ambulatory assessment," *Annual Review of Clinical Psychology*, vol. 9, no. 1, pp. 151-176, 2013.

[11] N. S. Suhaimi, J. Mountstephens and J. Teo, “EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities,” *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, p. 8875426, 2020.
[12] R. Jenke, A. Peer and M. Buss, “Feature extraction and selection for emotion recognition from EEG,” *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327-339, 2014.
[13] B. C. Ko, “A brief review of facial emotion recognition based on visual information,” *Sensors*, vol. 18, no. 2, p. 401, 2018.
[14] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon and A. C. Sobieranski, “A survey on facial emotion recognition techniques: A state-of-the-art literature review,” *Information Sciences*, vol. 582, pp. 593-617, 2022.
[15] M. El Ayadi, M. S. Kamel and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.
[16] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, “A comprehensive review of speech emotion recognition systems,” *IEEE Access*, vol. 9, pp. 47795–47814, 2021.
[17] S. Massaro and L. Pecchia, “Heart rate variability (HRV) analysis: A methodology for organizational neuroscience,” *Organizational Research Methods*, vol. 22, no. 1, pp. 354-393, 2019.
[18] A. H. Kemp, D. S. Quintana, M. A. Gray, K. L. Felmingham, K. Brown and J. M. Gatt, “Impact of depression and antidepressant treatment on heart rate variability: A review and meta-analysis,” *Biological Psychiartry*, vol. 67, no. 11, pp. 1067-1074, 2010.
[19] T. P. Beauchaine and J. F. Thayer, “Heart rate variability as a transdiagnostic biomarker of psychopathology,” *International Journal of Psychophysiology*, vol. 98, no. 2, pp. 338-350, 2015.
[20] J. Torous, P. Staples, M. Shanahan, C. Lin, P. Peck, M. Keshavan and J. P. Onnela, “Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder,” *JMIR Mental Health*, vol. 2, no. 1, p. e3889, 2015.
[21] M. J. Mathie, A. C. Coster, N. H. Lovell and B. G. Celler, “Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement,” *Physiological Measurement*, vol. 25, no. 2, p. R1, 2004.
[22] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10-49, 2015.
[23] J. C. Mundt, A. P. Vogel, D. E. Feltner and W. R. Lenderking, “Vocal acoustic biomarkers of depression severity and treatment response,” *Biological Psychiatry*, vol. 72, no. 7, pp. 580-587, 2012.
[24] A. S. Cohen and B. Elvevåg, “Automated computerized analysis of speech in psychiatric disorders,” *Current Opinion in Psychiatry*, vol. 27, no. 3, pp. 203-209, 2014.
[25] J. Torous, M. V. Kiang, J. Lorme and J. P. Onnela, “New tools for new research in psychiatry: a scalable and customizable platform to empower data driven

smartphone research," *JMIR Mental Health*, vol. 3, no. 2, p. e5165, 2016.
[26] S. Shiffman, A. A. Stone and M. R. Hufford, "Ecological momentary assessment," *Annual Review of Clinial Psychology*, vol. 4, no. 1, pp. 1-32, 2008.
[27] J. Fahrenberg, M. Myrtek, K. Pawlik and M. Perrez, "Ambulatory assessment - Monitoring behavior in daily life settings," *European Journal of Psychological Assessment*, vol. 23, no. 4, pp. 206-213, 2007.
[28] M. Kendall, "The advance theory of statistics," 1943.
[29] M. Song, Z. Yang, A. Triantafyllopoulos, X. Jing, V. Karas, J. Xie, Z. Zhang, Y. Yamamoto and B. W. Schuller, "Dynamic restrained uncertainty weighting loss for multitask learning of vocal expression," *arXiv preprint arXiv:2206.11049*, 2022.
[30] P. Dagum, "Digital biomarkers of cognitive function," *NPJ Digital Medicine*, vol. 1, no. 1, p. 10, 2018.
[31] T. R. Insel, "Digital phenotyping: technology for a new science of behavior," *JAMA*, vol. 318, no. 13, pp. 1215-1216, 2017.
[32] J. Torous, K. J. Myrick, N. Rauseo-Ricupero and J. Firth, "Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow," *JMIR Mental Health*, vol. 7, no. 3, p. e18848, 2020.
[33] D. C. Mohr, H. Riper and S. M. Schueller, "A solution-focused research approach to achieve an implementable revolution in digital mental health," *JAMA Psychiatry*, vol. 75, no. 2, pp. 113-114, 2018.
[34] B. A. Hickey, T. Chalmers, P. Newton, C. Lin, D. Sibbritt, C. S. McLachlan, R. Clifton-Bligh, J. Morley and S. Lal, "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review," *Sensors*, vol. 21, no. 10, p. 3461, 2021.
[35] N. Long, Y. Lei, L. Peng, P. Xu, P. Mao et al., "A scoping review on monitoring mental health using smart wearable devices," *Mathematical Biosciences and Engineering*, vol. 19, no. 8, pp. 7899-7919, 2022.
[36] H. Xu, X. Wu and X. Liu, "A measurement method for mental health based on dynamic multimodal feature recognition," *Frontiers in Public Health*, vol. 10, p. 990235, 2022.
[37] J. F. Huckins, A. W. DaSilva, W. Wang, E. Hedlund, C. Rogers, S. K. Nepal, J. Wu, M. Obuchi, E. I. Murphy, M. L. Meyer, 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," *Journal of Medical Internet Research*, vol. 22, no. 6, p. e20185, 2020.
[38] R. Das, M. R. Hasan, S. Daria and M. R. Islam, "Impact of COVID-19 pandemic on mental health among general Bangladeshi population: A cross-sectional study," *BMJ Open*, vol. 11, no. 4, p. e045727, 2021.
[39] M. B. Morshed, K. Saha, R. Li, S. K. D'Mello, M. De Choudhury, G. D. Abowd and T. Plötz, "Prediction of mood instability with passive sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1-21, 2019
[40] H. Takeuchi, K. Suwa, A. Kishi, T. Nakamura, K. Yoshiuchi and Y. Yamamoto, "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 and uHealth*, vol. 10, no. 10, pp. e39150, 2022.
[41] J. Melcher, R. Hays and J. Torous, "Digital phenotyping for mental health of college students: A clinical review," *BMJ Mental Health*, vol. 23, no. 4, pp. 161-166, 2020.
[42] M. Gerczuk, A. Triantafyllopoulos, S. Amiriparian, A. Kathan, J. Bauer, M. Berking and B. W. Schuller, "Zero-shot personalization of speech foundation models for depressed mood monitoring," *Patterns*, vol. 4, no. 11, 2023.
[43] C. Su, Z. Xu, J. Pathak and F. Wang, "Deep Learning in Mental Health Outcome Research: A Scoping Review," *Translational Psychiatry*, vol. 12, no. 1, p. 116, 2020.
[44] I. Fukui, "The depression and anxiety mood scale (DAMS): Scale development and validation," *Japanese Journal of Behavior Therapy*, vol. 23, p. 83, 1997.
[45] A. Satt, S. Rozenberg and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1089-1093.
[46] C. Burton, B. McKinstry, A. S. Tătar, A. Serrano-Blanco, C. Pagliari and M. Wolters, "Activity monitoring in patients with depression: A systematic review," *Journal of Affective Disorders*, vol. 145, no. 1, pp. 21-28, 2013.
[47] B. Helgadóttir, Y. Forsell and Ö. Ekblom, "Physical activity patterns of people affected by depressive and anxiety disorders as measured by accelerometers: A cross-sectional study," *PLoS One*, vol. 10, no. 1, p. e0115894, 2015.
[48] P. C. Loizou, "Speech enhancement: Theory and practice," 2013.
[49] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben and B. W. Schuller, "Dawn of the Transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transcations of Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745-10759, 2023.
[50] Y. Zhang, Y. Wei and Q. Yang, "Learning to multitask," *Advances in Neural Information Processing Systems*, vol. 31, pp. 5771-5782, 2018.
[51] M. Song, Z. Yang, A. Triantafyllopoulos, T. Nakamura, Y. Zhang, Z. Ren, H. Takeuchi, A. Kishi, T. Ishizawa, K. Yoshiuchi, et al., "Crossmodal transformer on multi-physical signals for personalised daily mental health prediction," in *Proc. ICDM Workshop*, Shanghai, China, 2023, pp. 1299-1305.
[52] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281-286, 2002.
[53] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu and M. Yang, "Gated fusion network for single image dehazing," in *Proc. CVPR*, Salt Lake City, UT, USA, 2018, pp. 3253-3261.
[54] W. Dong, D. Yan, Z. Lin and P. Wang, "Efficient adaptation of large vision transformer via adapter re-composing," in *Proc. NeurIPS*, New Orleans, LA, USA, 2023.

Dear Reviewers,

We sincerely appreciate your thoughtful comments and suggestions on our manuscript. Your feedback has been invaluable in improving the quality and clarity of our paper. We have carefully addressed each of your concerns and made substantial revisions to the manuscript. Here's a summary of the major changes:

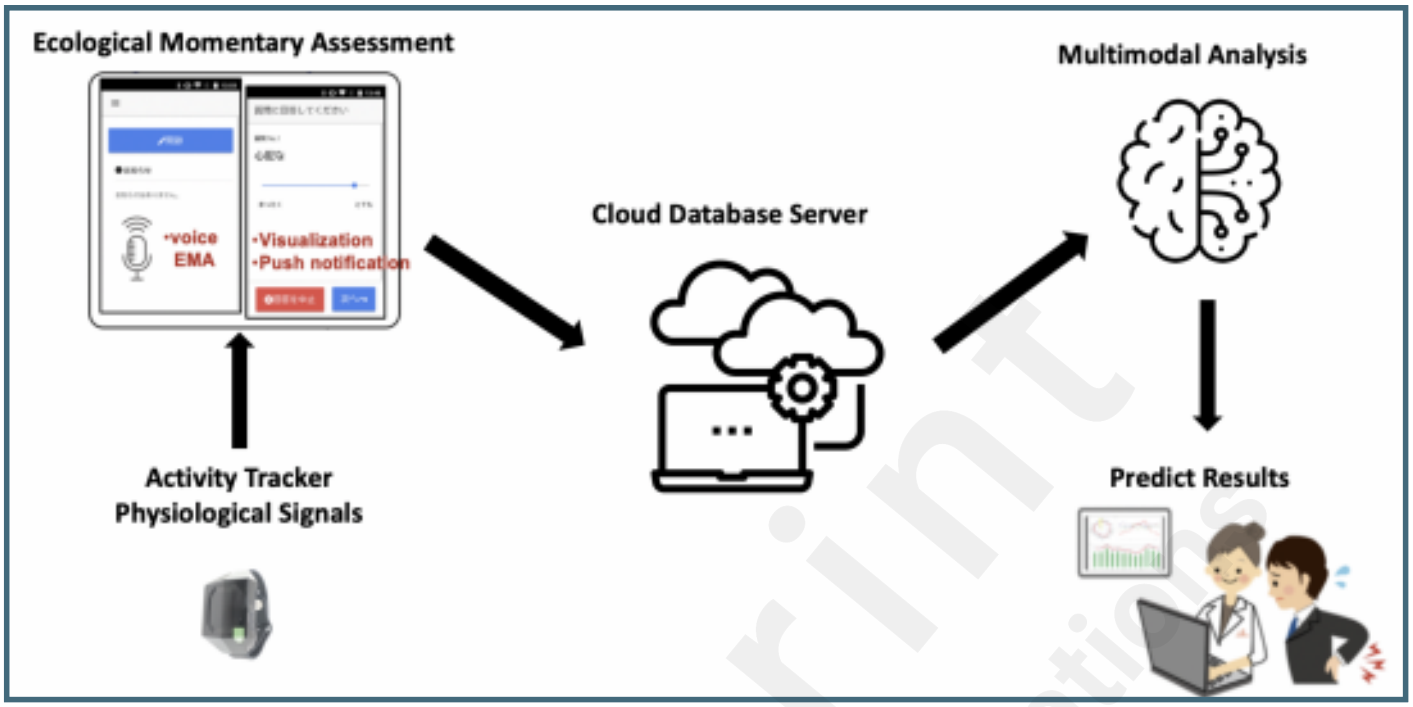
1. Introduction:
    - We have thoroughly rewritten and reorganized the introduction, providing a more comprehensive background on our chosen modalities.
    - We've elaborated on how physical activity affects mental health and explained the connection between physiological data and emotions.
    - We've added detailed sections on Ecological Momentary Assessment (EMA), the Depression and Anxiety Mood Scale (DAMS), and our newly developed deep learning methods.
2. Related Work:
    - We've expanded the Related Work section to include more recent research literature on multimodal data for mental health diagnosis, as suggested.
3. Privacy Considerations:
    - We've reviewed and included recent research on privacy implications of voice and speech analysis, justifying our choice of speech data over facial expression data.
4. Methodology:
    - We've reorganized the Methods section and added more detailed content on participants and the EMA paradigm.
    - We've clarified the correlation between DAMS and speech & physical activity data, illustrating how these different data types are related.
5. Model Architecture and Training:
    - We've explained our approach using both pre-trained (for speech features) and non-pre-trained (for physiological data) components.
6. Learning Approach and Data Labeling:
    - We've clarified that our model uses supervised learning, with labels derived from self-reported DAMS scores.
    - We've detailed our time-based data split approach (70% training, 15% validation, 15% test), which respects the temporal nature of the data.
7. Practical Applications:
    - We've added information on how we redirect subjects to seek help if needed based on the collected data.
    - We've explained how the data collected could contribute to improving mental health outcomes.
8. Discussion:
    - We've added a comprehensive Discussion section that was previously missing.
    - In this section, we've carefully considered the limitations of our framework and analyses, as suggested.

- We've also interpreted our results in the context of existing literature and discussed implications for future research and applications.

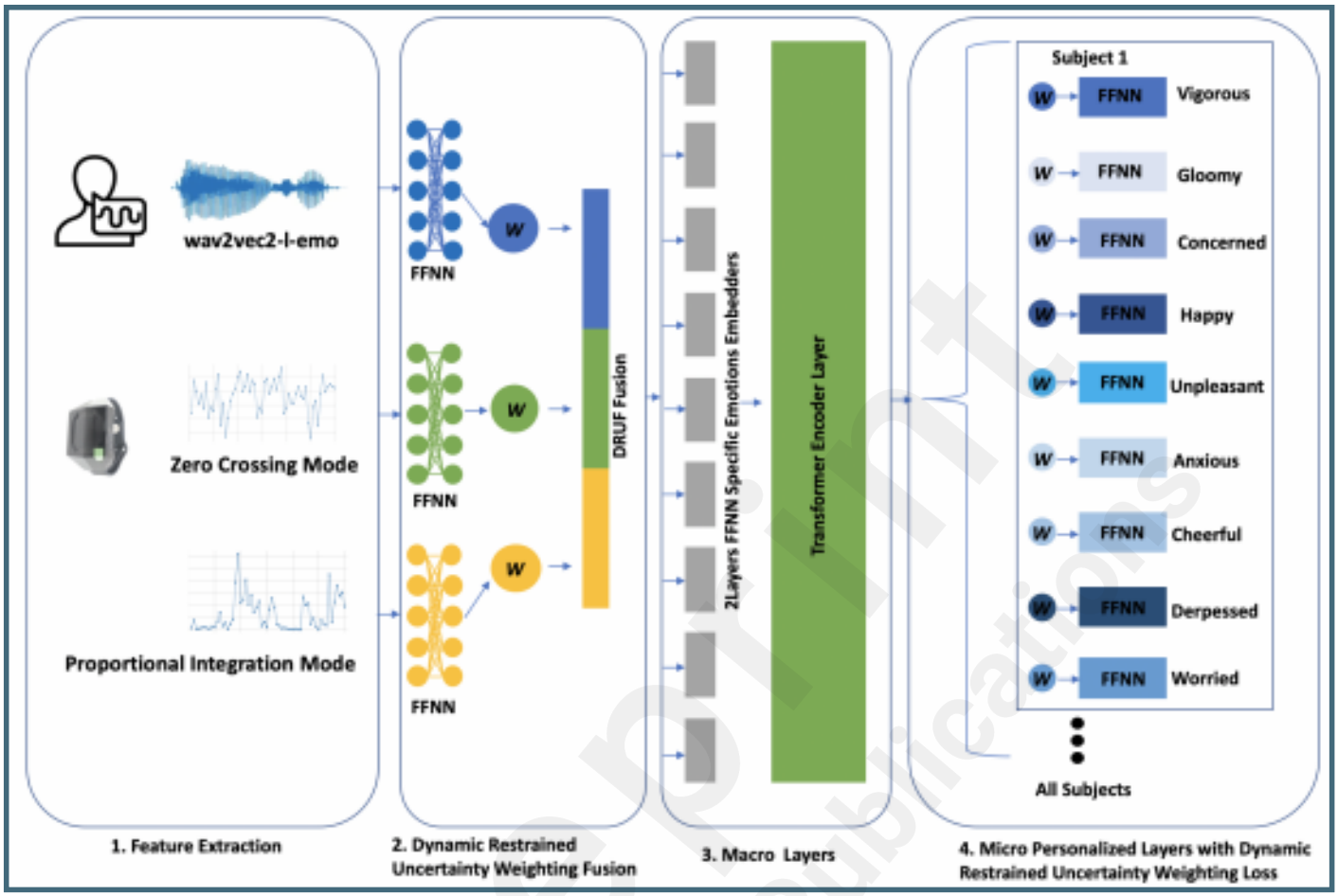
We believe these revisions have significantly strengthened our manuscript and addressed all of your concerns. The paper now provides a clearer, more comprehensive presentation of our research methodology, theoretical foundations, potential real-world applications, and limitations.

Supplementary Files

Untitled.



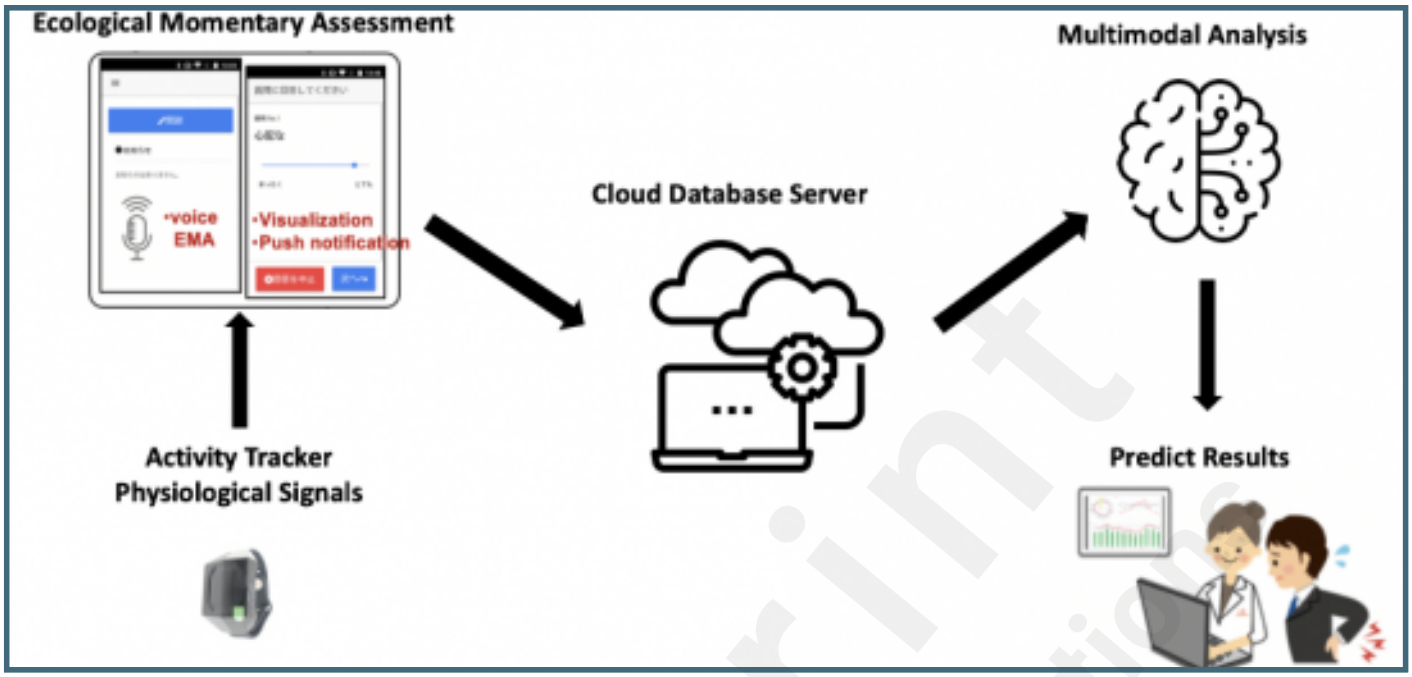
Untitled.



Main Paper.
URL: http://asset.jmir.pub/assets/2a2736f81158e8cf644616609bf613aa.docx

Figures

MHIT system designed for data collection.



Our Macro-Micro Framework.