This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3733235

A Multimodal Sensor Fusion Approach for Mental Health Detection Using Mobile, Wearable, and IoT Sensors

Youngji Koh, Gyuna Kim, Chanhee Lee, Panyu Zhang, Yunhee Ku, Inhwan Choi, Jewoo Ryu, and Uichin Lee

***Abstract*— Depression and anxiety are among the most prevalent mental health disorders, yet many cases remain undetected due to the lack of continuous and context-aware monitoring in everyday life. Prior work has demonstrated the potential of mobile and wearable devices for passive mental health sensing; however, their inconsistent usage limits coverage of in-home routines and context-rich behaviors. To address these limitations, we propose a multimodal sensing approach that integrates data from mobile phones, wearable devices, home IoT sensors, and smart speakers to capture daily behavioral patterns in real-world settings. We conducted a 30-day in-the-wild study with 20 single-person households aged 20 to 30, a group at elevated risk for mood disorders. We developed a multimodal pipeline that extracts and aligns features across all modalities and applies both machine learning and deep learning models. Our results show that multimodal integration consistently outperforms single-modality baselines in detecting both depression and anxiety. Deep learning models achieved AUROCs of 0.690 for depression and 0.634 for anxiety in generalized settings, while personalized tree-based models performed best, achieving up to 0.731 and 0.728, respectively. We publicly released our code and feature sets to support reproducibility and further research in pervasive mental health.**

***Index Terms*— Internet of Things, Mental health, Mobile devices, Multimodal sensing, Smart home systems, Wearable sensors**

I. Introduction

DEPRESSION and anxiety are among the most common and serious mental health disorders globally, affecting approximately one in eight people according to the World Health Organization [1]. However, over 70% of cases often go undiagnosed or untreated [2]. Traditional diagnostic methods primarily rely on self-reported questionnaires, such as the PHQ-9 and GAD-7, which assess mental disorder symptoms over the past two weeks. These methods rely on users' ability to retrospectively assess their symptoms, making them prone to recall bias [3]. In addition, users may not always be aware of or accurately report their own symptoms [4], which hinders timely and reliable detection.

To address these limitations, there has been growing research interest in passive sensing approaches using mobile and wearable devices for mental health detection [5], [6]. These devices have been used to monitor behavioral functioning, such as physical activities and social interactions, related to Activities of Daily Living (ADLs) [7], and to capture additional indicators like voice characteristics [8], which may serve as indirect markers of depression and anxiety. However, mobile and wearable sensing often suffers from incomplete data due to inconsistent usage [9], [10].

Beyond sensing coverage, depression and anxiety are clinically characterized by functional impairment in everyday life, including disruptions in sleep, eating, and daily routines. Such impairment is difficult to capture through mobile or wearable sensing alone, which primarily reflects short-term mobility or activity patterns. Given that individuals spend much of their time at home [11], capturing home-based daily routines is therefore critical, yet remains underexplored in prior mental health sensing studies.

To expand the scope of behavioral sensing beyond personal devices, recent work has highlighted the complementary potential of ambient sensing in home environments using Internet-of-Things (IoT) technologies [12], [13]. Building on these advances, our work focuses on inference-oriented mental health detection by integrating home IoT and mobile sensing to model passive, longitudinal behavioral patterns in everyday home settings. As illustrated in Figure 1, our study integrates multimodal passive sensing from mobile phones, wearable devices, and home IoT sensors with self-reported ESM data to compensate for modality-specific limitations. IoT sensors embedded in furniture, appliances, and environmental sensors enable low-burden continuous monitoring of daily routines without requiring active user engagement.

Despite these advantages, previous work has focused mainly on using home IoT detection to monitor cognitive or functional decline in older adults [14], [15], leaving its potential for detecting mood disorders such as depression and anxiety underexplored. To explore this potential, we focus on a younger population at-risk; individuals aged 20–30 living alone. This

(Corresponding author: Uichin Lee.)
Youngji Koh, Chanhee Lee and Uichin Lee are with the School of Computing, KAIST, Daejeon 34141, South Korea (e-mail: youngji@kaist.ac.kr; chanhee015@kaist.ac.kr; uclee@kaist.ac.kr).
Gyuna Kim, Panyu Zhange are with the Graduate School of Data Science, KAIST, Daejeon 34141, South Korea (e-mail: gyuna.kim@kaist.ac.kr; panyu@kaist.ac.kr).
Yunhee Ku, Inhwan Choi, Jewoo Ryu are with LG Electronics, Seoul 07336, South Korea (email: yunhee.ku@lge.com; inhwan.choi@lge.com; jewoo.ryu@lge.com).

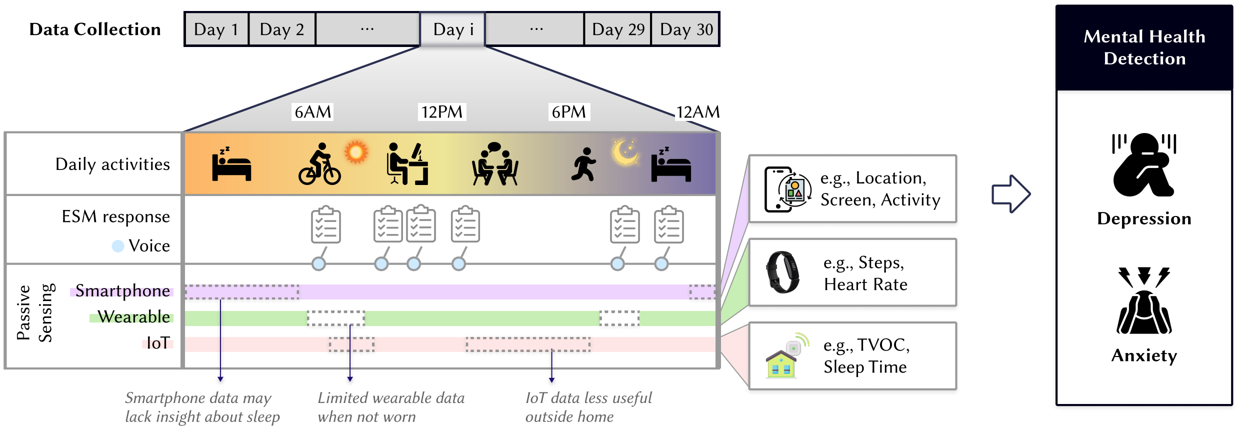


Fig. 1: Overview of the study design using 30-day multimodal passive sensing and self-reported data from the Experience Sampling Method (ESM) for mental health detection. Multimodal data helps compensate for modality-specific limitations.

group has been shown to experience elevated risks of depression and anxiety, in part due to increased social isolation and the increasing prevalence of single-person households [16], [17].

To better understand their behavioral patterns, we specifically examine Instrumental Activities of Daily Living (I-ADLs), which are strongly linked to independent functioning and mental health status [18]. According to the DSM-5 [19], disruptions in daily activity, eating, and sleep are core symptoms of depression and anxiety. These behavioral changes can be captured through home IoT sensing that reflects everyday domestic routines, including activity patterns, food-related behaviors, and sleep. We further leverage voice data collected through in-home smart speaker interactions as a promising modality for mental health detection.

While each modality contributes valuable information, none alone provides a complete picture of an individual's mental health state. To overcome these limitations, we propose a multimodal sensor fusion approach that integrates data from mobile phones, wearable devices, and home IoT sensors to detect depression and anxiety. Rather than merely hypothesizing the potential benefits of each sensing modality (e.g., IoT for domestic routine tracking and wearables for physical activity tracking), our study empirically investigates how their integration or absence affects detection performance in real-world conditions, and further analyzes the resulting features to interpret their clinical relevance. To support a comprehensive empirical analysis, we systematically evaluate several modeling approaches, including both machine learning and deep learning models appropriate for our feature-engineered representations, under generalized and personalized settings.

To the best of our knowledge, this is the first in-the-wild study to integrate mobile, wearable, home IoT, and voice data for depression and anxiety detection using off-the-shelf consumer devices. We conducted a 30-day study with 20 participants living alone, collecting continuous behavioral and physiological data and task-based voice recordings using off-the-shelf consumer devices. Our results show that incorporating home IoT and voice data provides complementary behavioral signals beyond mobile and wearable sensing alone, improving detection performance and supporting feature-level interpretation. To support reproducibility, we publicly release our code and extracted feature sets.[1]

II. Background and Related Works

In this section, we first review prior studies on mobile and wearable sensing for tracking depression and anxiety. We then discuss existing home IoT research, which has primarily focused on monitoring the cognitive health of older adults, and consider its potential for detecting mood disorders in broader populations.

A. Mobile and Wearable Sensing for Mental Health

People experiencing mood disorders, such as depression and anxiety, tend to exhibit changes in various daily activities, including sleep, eating habits, physical activity, and social interactions [20], [21]. With the advancement of ubiquitous computing technologies, mobile and wearable devices have enabled passive tracking of everyday behaviors that are indicative of mental health conditions such as depression and anxiety. Prior work has demonstrated that mobile and wearable sensing can passively capture behavioral signals related to mental health, including sleep, physical activity, eating behaviors, and social interactions [7], [22]. In addition, prior studies have explored speech-based depression and anxiety detection using acoustic and prosodic features, as well as emotion-aware multimodal ensemble models for video-based depression detection [23]–[25].

However, mobile and wearable sensing provides limited visibility into *in-home activities*, particularly when devices are not worn or carried. This limitation is critical for mental health monitoring, as key behavioral changes related to IADLs, sleep, and daily routines often occur within the home, and individuals with depression or anxiety tend to spend more time indoors [11]. As a result, conventional mobile and wearable sensing frequently suffers from data gaps, hindering continuous monitoring of everyday life. Moreover, these devices provide limited information about in-home environmental factors,

[1] `https://github.com/Kaist-ICLab/multimodal-mh-detection.git`

such as noise, temperature, and light, which are known to influence mental health [26].

To address these limitations, mental health detection should incorporate not only mobile and wearable data but also home IoT sensor data. By capturing passive interactions with appliances and monitoring ambient conditions, IoT sensors offer unique insights into daily routines that mobile and wearable sensors may miss. Despite growing interest in multimodal sensing, no prior studies have yet to fully integrate diverse sources of behavioral data across personal and ambient environments (i.e., mobile, wearable, home IoT sensors). Notably, home IoT sensing remains underexplored, despite its potential to improve performance or even function as a standalone modality.

B. Smart Home Sensing for Mental Health

Most existing work on home IoT sensing has focused on monitoring cognitive decline and functional impairment in older adults, leveraging sensors embedded in household appliances and furniture to capture changes in daily routines [14], [15]. For example, Lee and Dey [27] instrumented household appliances with IoT sensors to monitor activities such as pill-taking, phone usage, and the multi-step task of making a pot of coffee. Similarly, Andreadis et al. [28] attached plug-in sensors to kitchen appliances (e.g., microwaves, electric kettles) to monitor power consumption patterns associated with cooking behaviors in people living with dementia. In addition to appliance-based monitoring, motion sensors have been used to track in-home movement patterns as a proxy for daily functioning and to identify early signs of Mild Cognitive Impairment (MCI) and Alzheimer's Disease [29]. These sensing approaches suggest that changes in IADL-related behaviors, sleep patterns, and in-home activity levels may serve as behavioral indicators of mental health conditions. However, most prior work has emphasized cognitive decline and functional impairment in older adults, leaving mood disorders such as depression and anxiety comparatively underexplored in everyday home environments.

In addition to sensor-based data (e.g., mobile, wearable, and IoT sensors), the detection of depression and anxiety also requires the collection of self-reported data that reflect individuals' subjective emotional and psychological states. Prior studies have combined sensor data with self-reported measures using Experience Sampling Method (ESM) to capture momentary mental health states [7], [22]. More recently, smart speakers have been explored as a low-burden modality for collecting voice-based self-reports in home environments [30].

To summarize, in this study, we aim to integrate IoT sensor data from home environments with mobile and wearable data to detect depression and anxiety. We apply multimodal data fusion techniques to develop a detection model, compare model performance across different data sources, and analyze dominant features in generalized models to identify digital behavioral markers associated with mental health conditions.

III. DATA COLLECTION

In this section, we provide a comprehensive overview of our data collection methodology, detailing the types of data collected, devices used for data acquisition, as well as participant recruitment and demographic information.

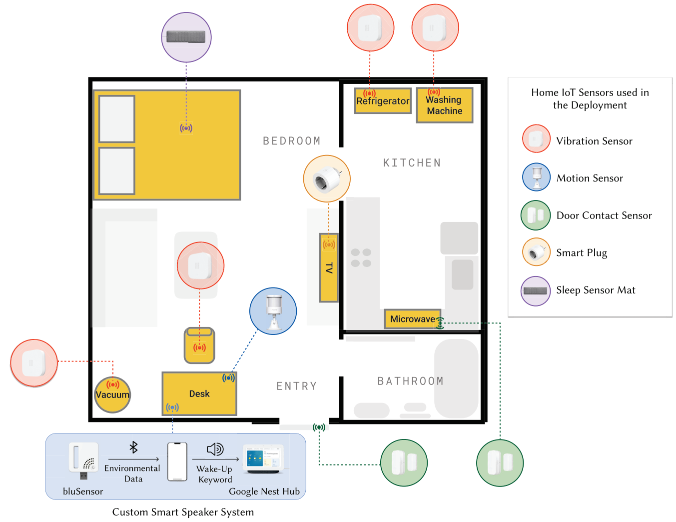


Fig. 2: Example deployment of home IoT sensors

A. Overview

Drawing on insights from prior research and established clinical frameworks, including the DSM-5 diagnostic criteria [19] for depression and anxiety disorders, we identified five lifestyle-related factors commonly associated with depression and anxiety: sleep, eating, physical activity, social interaction, and environmental conditions [20], [21]. To comprehensively capture these data types in real-world environments, we collected multimodal data from participants using commercial home IoT sensors, mobile and wearable devices, and a custom smart speaker system. Deployment examples of home IoT sensors and the smart speaker system are shown in Fig. 2. This study was approved by the Institutional Review Board (IRB approval No.KH2022-025).

B. Data Types

We collected data from three main sources: (1) home IoT sensors, (2) custom smart speaker systems, and (3) mobile and wearable devices. Below, we describe the types of data collected from each source.

1) Home IoT data: Home IoT sensors provided data on sleep patterns, food-related behaviors, and indoor physical activity.

Sleep patterns. Sleep is a critical factor in mental health and a key diagnostic criterion in the DSM-5 [19]. To unobtrusively and accurately measure sleep, we used a bed-mounted IoT sensor mat, which automatically tracks various sleep-related metrics. These include sleep duration, sleep efficiency, latency, number of awakenings, time spent in different sleep stages (light, deep, REM), heart rate, respiratory rate, snoring events, and an overall sleep score.

Food-related behaviors. Monitoring food-related behaviors (e.g., eating, cooking) is challenging due to the difficulty of automatically detecting food intake. Most studies relied on self-reported meal frequencies [31] or manual food logs [32].

In home environments, the use of kitchen appliances (e.g., refrigerator, coffee maker) has been employed as a proxy for eating behavior [27], [33]. Similarly, we monitored the usage frequency of two representative kitchen appliances (refrigerator and microwave) to approximate meal patterns. These appliances were selected due to their frequent and direct association with food storage and preparation.

Indoor physical activity. To capture physical activity in home settings, prior work has explored indicators such as indoor movements, household chores, and entry/exit events [29]. In our study, indoor physical activity was monitored through a motion sensor installed in the home, usage of household appliances, and front door open/close events indicating entry and exit. For household chores, we focused on the use of the vacuum cleaner and washing machine as proxies for domestic physical activity. In addition, television usage was tracked as a form of sedentary behavior, which may reflect passive engagement or reduced physical activity during leisure time.

2) Smart speaker data: Smart speakers collected three categories of data: self-reported mental health scores, voice recordings from structured tasks, and indoor environmental sensor data.

Self-reported mental health assessments. Participants responded to two standardized questionnaires, PHQ-2 [34] and GAD-2 [35], administered via the smart speaker, either through voice or touch interaction, to assess depressive and anxiety symptoms. The smart speaker served as the primary user-facing interface for both ESM delivery and response, while smartphones were used only for background triggering and audio recording.

Voice recordings. Participants engaged in a series of voice-based tasks, including describing recent activities and visual prompts (e.g., image cards), which were designed to elicit spontaneous speech rather than speech from a predefined script. The resulting recordings were then used to extract acoustic and linguistic features relevant to emotional and cognitive states.

Environmental conditions. In addition to supporting voice-based interactions, the smart speaker system also enabled contextual sensing by integrating environmental sensors. Prior research has shown that environmental factors such as light exposure, temperature, and humidity can significantly influence mental health [26]. Building on this, we equipped the smart speaker with sensors that continuously measured indoor light levels, temperature, humidity, carbon dioxide (CO_2), and total volatile organic compounds (TVOC). These environmental readings were used not only for passive data collection but also to support context-aware experience sampling prompts. Further implementation details are provided in Section III-C.2.

3) Mobile and wearable data: Smartphone and wearable devices captured data on mobility, social interaction, app usage, device context, physical activity, and sleep.

Mobile data. Smartphones captured a wide range of passive sensing data related to users' mobility, social interaction, and device usage. Mobility data included GPS location traces and physical activity states (e.g., walking, stationary, in-vehicle). Social interaction patterns were derived from call and SMS metadata, including frequency, duration, and timing of communication events (excluding content). Device usage patterns encompassed foreground app activity, screen on/off events, and device context such as battery level, charging status, and Wi-Fi scan history.

Wearable data. A commercial wrist-worn device was used to record both physiological and activity-related data. The captured metrics included step count, distance traveled, heart rate, and calorie expenditure, enabling monitoring of users' daily activity patterns. In addition, the device provided sleep-related information, including total sleep duration and time spent in different sleep stages (e.g., light, deep, REM).

C. Apparatus

To collect multimodal data in participants' home environments, we deployed a combination of commercial IoT sensors, a custom-developed smart speaker system, and a wrist-worn wearable device. Detailed descriptions of the data types collected from each source are available in our GitHub repository.[1]

1) Home IoT sensors: We installed five IoT sensing components to capture appliance usage, indoor activity, environmental context, and sleep, including vibration, contact, motion, and power consumption sensors, as well as a bed-mounted sleep tracking mat (Fig. 3). Specifically, four types of Aqara smart home sensors were deployed to monitor appliance interactions, in-home activity, and presence, while a Withings sleep tracking mat placed under the mattress continuously recorded sleep and physiological metrics (e.g., WASO, total sleep time, and sleep stages).

2) Custom smart speaker system: We developed a custom smart speaker system to collect environmental data, voice interactions, and self-reported mental health assessments in the home. The system delivered timely ESM prompts and captured in-situ voice responses, building on prior work in smart speaker-based mental health monitoring [30]. It consisted of three components: a smartphone-based speaker-triggering app, a multimodal smart speaker (Google Nest Hub 2nd Gen), and an air quality sensor (BluSensor BSP02AIQ).

ESM triggering strategy. ESM prompts were triggered using a hybrid strategy based on both contextual and temporal criteria. When the system inferred that the participant was present near the smart speaker and that the moment was suitable for interaction (e.g., reduced ambient noise and proximity to the device based on light, camera, and CO_2 signals), a context-aware prompt was issued. In addition, to ensure temporal coverage, prompts could also be delivered after a maximum interval of 90 minutes since the last query, regardless of inferred presence.

If a prompt was delivered while the participant was not at home or unavailable, it could be missed without retry and was treated as missing data. Upon meeting either trigger condition, the smartphone application activated the smart speaker using a predefined wake-up keyword (e.g., "OK Google, run the survey program"), initiating a voice-based ESM session. Participants received 8 to 15 prompts per day within a 10-hour window (e.g., 10 AM–8 PM) and were encouraged to respond to at least three prompts per day.

Task types. Each ESM session consisted of three tasks: (1) reporting the activity performed immediately before the prompt, (2) responding to standardized mental health questionnaires, and (3) describing an image card shown on the smart speaker.

We chose PHQ-2 and GAD-2 due to their brevity and suitability for high-frequency ESM. These instruments are validated screening tools for capturing momentary depressive and anxiety symptoms, and their simplicity minimizes participant burden during repeated daily surveys. Participants responded to the mental health questionnaires using a 4-point Likert scale (0 = Not at all, 3 = Very frequently). These questions allowed for either voice or touchscreen responses. For voice-based responses, participants verbally selected the corresponding numeric option (e.g., "zero" to "three"), consistent with prior smart speaker-based ESM designs.

3) Mobile sensing application and wearable device: Mobile data collection application. To collect passive smartphone sensing data, participants were required to install the phone logger to their phone, an open-source Android application [36], on their devices. The application operated continuously in the background throughout the study period. The app captured various behavioral and contextual data, including GPS-based mobility, physical activity states (e.g., walking, stationary), call and SMS metadata (frequency, duration), Wi-Fi scan history, foreground app usage, screen on/off events, and device status (e.g., battery level, charging). All data were stored locally and periodically uploaded to a secure research server for analysis.

Wearable device. To monitor participants' physical activity and physiological signals, we used the Fitbit Inspire 2, a commercially available wrist-worn activity tracker. Using the Fitbit API, we retrieved high-resolution data at 1-minute intervals, including step count, distance traveled, calorie expenditure, and heart rate. Sleep data were also collected, with each entry recorded as associated with the calendar date preceding the onset of sleep.

D. Participants

We recruited 20 individuals living in single-person households through online promotions on university platforms and local community websites. Participants had a mean age of 24.79 years (SD = 2.86) and included 13 males and 7 females. All participants lived in studio or one-bedroom apartments, enabling consistent sensor placement across similar residential layouts.

Eligibility criteria required participants to live alone, own key household appliances (e.g., bed, chair, microwave, refrigerator, washing machine, and vacuum cleaner), spend at least five hours per day at home excluding sleep, and use an Android smartphone (version 8.0 or higher) compatible with our mobile logging app. The five-hour threshold ensured sufficient in-home behavioral data for sensor-based analysis. Given recent reports of declining TV ownership among single-person households [37], TV usage was treated as optional. To capture a broad range of depressive and anxious symptoms in naturalistic settings, we did not screen participants based on clinical diagnoses, allowing inclusion of individuals with subclinical symptoms.

Participants were asked only to wear and charge the wearable device during waking hours and not to relocate installed sensors. Each participant completed a 30-day study between May 8 and June 13, 2023. Multimodal data were collected from mobile phones, wearables, home IoT devices, and smart speakers. To assess daily mental health status, participants received 8–15 PHQ-2 and GAD-2 prompts per day via smart speakers and were encouraged to respond at least three times daily. Participants responded on average 3.35 times per day (range: 1.76–6.84), indicating moderate compliance under a high-frequency prompting schedule.

IV. DATA ANALYSIS

This section presents the data analysis methodology employed for the modeling process, as illustrated in Fig. 3.

A. Preprocessing

To ensure data quality and reliable model training, we applied the following filtering criteria to the collected dataset. A total of 1,157 ESM samples with audio responses were collected from 20 single-person households via smart speaker-based ESM surveys. Participants with fewer than 30 ESM responses were excluded due to insufficient sample size. Additionally, we removed users whose binarized labels (PHQ-2 or GAD-2) were heavily imbalanced, where a single class accounted for more than 90% of responses. This step was taken to reduce class bias and ensure stable model learning. Following this filtering, 11 participants were retained for the depression analysis (606 samples), and 13 for the anxiety analysis (845 samples).

For label encoding, we binarized the scores to classify each ESM survey response as *depressed* or *non-depressed* and *anxious* or *non-anxious* based on the corresponding PHQ-2 and GAD-2 scores (each ranging from 0 to 6). While a cut-off score of 3 is commonly used in clinical contexts, a threshold of 2 has also been employed in prior studies to improve sensitivity [38], [39]. In line with this, we adopted a threshold of 2 to better capture subclinical symptoms, including those who experienced at least mild depressive or anxious mood. This decision aligns with the goal of our study, which focuses on everyday mental well-being and self-care rather than clinical diagnosis.

We then applied noise filtering and signal preprocessing strategies tailored to each data modality. For smartphone-derived data, we removed call events with negative or zero durations. Message events were encoded as binary indicators. For physical activity, accelerometer data were converted to signal magnitude using the X, Y, Z axes, and step count, distance, and calories were calculated between consecutive pedometer entries. Location data were processed by computing haversine distance between GPS coordinates, and clustering user locations into semantic places (e.g., home, work, and Google Map-based labels) via POI (Point-of-Interest)-based clustering. To extract phone usage features, mobile apps were re-categorized into predefined categories [40], and the duration

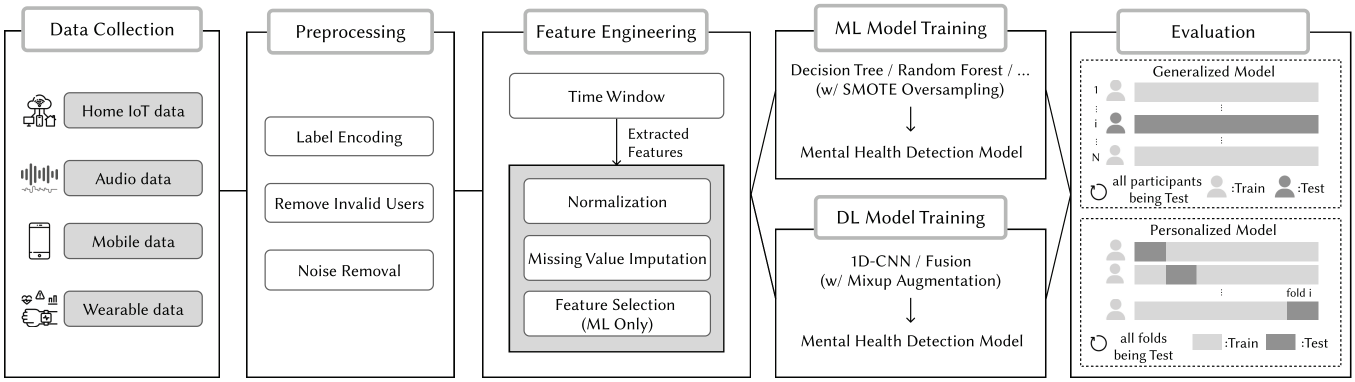


Fig. 3: Data Analysis Process

of each app usage session was computed. Installed apps were compared across consecutive timestamps using Jaccard similarity. Screen usage was measured by calculating the duration of each screen-on session. WiFi events were processed by computing cosine, Euclidean, and Manhattan distances between consecutive RSSI values, and Jaccard similarity between BSSID sets. Media activities (e.g., image or video access) were encoded as binary features.

For wearable data collected from young adult participants, biologically implausible heart rate values below 30 bpm or above 220 bpm were filtered out, consistent with a prior study that employed the same physiological limits [41]. For audio data, speaker diarization was applied to each smart speaker interaction recording to separate the participant's voice from the device's output. Using the Naver Clova API, we segmented speaker turns and extracted participant-spoken segments. We further trimmed silence and normalized amplitude using root mean square (RMS) energy threshold.

Home IoT data typically consist of discrete binary or timestamped events (e.g., microwave door openings), which are inherently robust to noise and do not require additional denoising. However, for TVs, simple event detection is not feasible due to continuous low-level power consumption (e.g., standby mode). Therefore, we used the Aqara smart plug to collect real-time power usage data and empirically tested operational thresholds to distinguish actual usage from background fluctuations. For instance, by analyzing observed usage patterns, we defined a minimum active power level to infer when the TV was truly in use, thereby reducing false activations and improving event accuracy.

To capture daily sleep patterns, we used both Fitbit (wrist-worn) and Withings (bed-mounted) devices. Although each device occasionally missed data due to the device not being worn (Fitbit) or the bed sensor not detecting sleep (Withings), the missing days rarely overlapped. For example, 15 out of 20 participants had sleep data recorded by Fitbit on fewer than 3 out of 30 nights, indicating that the wearable device was often not worn or failed to detect sleep during bedtime. However, there were also cases where Fitbit successfully recorded sleep when the bed sensor failed, highlighting the complementary nature of the two modalities. When combining Fitbit and Withings data, all participants had sleep data available for every single day. Based on this, we treated a day as valid if either device provided sleep data, leveraging the complementary coverage of the two modalities to reduce sparsity in longitudinal monitoring.

B. Feature Engineering

Sensor features were extracted and aggregated across this window aligned with the corresponding PHQ-2 and GAD-2 responses. Previous studies have commonly employed a 15-minute window for feature extraction [42]. However, such short windows are often unsuitable for sparse, event-driven data from home IoT sensors, particularly in single-person households where event frequency is often limited. To address this, prior work suggests that extending the window to approximately 80 minutes can alleviate sparsity issues [43]. Following a similar rationale, we adopted a 60-minute time window for feature extraction.

All features were normalized using z-score normalization, with normalization parameters computed exclusively from the training data within each evaluation fold. To handle missing values, numerical features were interpolated and categorical features were forward-filled. For traditional machine learning models, feature selection involves removing zero-variance features and eliminating highly correlated feature pairs to reduce redundancy and mitigate the curse of dimensionality. Feature selection was performed independently within each training fold using only the training data, resulting in potentially different feature subsets across folds.fter feature extraction, the initial feature space comprised approximately 3,000 features, which was reduced through zero-variance removal and correlation-based feature filtering.

1) Home IoT & Environmental data: We extracted appliance usage features by aggregating the number of activation events within two time windows: (1) a 60-minute window preceding each label timestamp, and (2) a 12-hour contextual window (either morning or evening, based on label time). These windows allowed us to capture both short-term and extended behavioral patterns. For instance, microwave or vacuum cleaner usage was counted within these intervals as proxies for food preparation or cleaning activity.

Sleep features from the bed sensor were aggregated over the previous 24 hours. Extracted features included total time

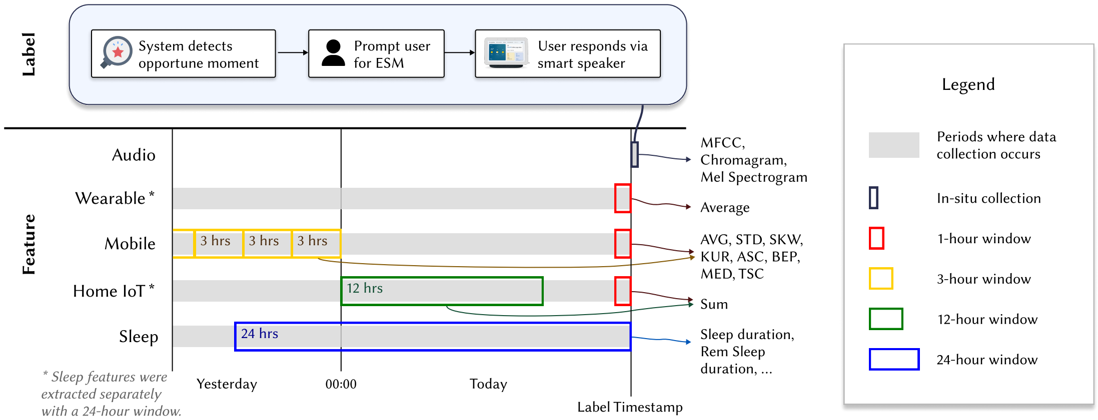


Fig. 4: Time window-based feature extraction approach for each sensing modality.

in bed, total sleep time, time in each sleep stage (light, deep, REM), and number of bed exits. These features were computed as sums, representing absolute behavioral quantities rather than momentary fluctuations.

Environmental features such as temperature, and humidity were averaged over a 15-minute window before each label, with both mean and standard deviation computed to reflect ambient stability and variability.

2) Audio data: Acoustic features used in speech emotion recognition are commonly grouped into prosodic, spectral, and voice quality [44]. Prosodic features describe variations in pitch, energy, and duration that often signal emotional states. Spectral features, derived from the resonant properties of the vocal tract, include measures such as Mel frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients (LPCC), which capture the short term distribution of energy across frequency bands. Voice quality features capture physiological aspects of phonation. In this study, we focused on spectral low level descriptors (LLDs), which are widely used in emotion recognition and remain computationally efficient and reliable for short, fragmented utterances with background noise [45]. To extract these features, we used the Librosa library to compute MFCC, mel spectrogram, and chromagram representations from each audio recording. All features were derived from Short Time Fourier Transform (STFT) based representations computed using a Hann window (frame size: 2048 samples, hop length: 512 samples). Frame level features were aggregated over time using mean pooling to generate a fixed length vector per ESM response. We also extracted utterance level duration and word count to quantify the amount of speech per recording.

3) Mobile data & Wearable data: We followed the mobile data pre-processing and feature extraction strategy used in prior work [36], [42]. Feature extraction was based on three time windows: the 60-minute window, the previous day's 3-hour epochs, and the current value window. Epoch-based windowing has been widely adopted for its effectiveness in capturing extended behavioral patterns. Following Zhang et al. [42], we segmented the previous day (06:00–24:00) into 3-hour epochs and computed summary statistics (mean and standard deviation) for each interval. Additionally, we used a 60-minute window preceding each label to extract immediate past features, and adopted the current window strategy from Kang et al. [36], which utilizes the most recently observed value before the label timestamp.

Using these windowing strategies, we derived a wide range of features from the mobile data, including social interaction, physical activity, context, phone usage, and system status. For social interaction, we extracted features such as call duration, call frequency, and message counts (sent, received, and total) to capture communication patterns. Physical activity features included raw accelerometer signals (X, Y, Z axes and computed magnitude), activity transitions (e.g., ENTER_WALKING, ENTER_STILL), and activity confidence scores (e.g., OnFoot, InVehicle, Tilting).

Contextual information was derived from location data, including total distance traveled, spatial clustering of visited locations, and semantic labels such as "home" and "work." For phone usage behavior, we computed both categorical and numeric features, including app usage types, session durations, and the Jaccard similarity between installed app lists over time. Screen activity was characterized by screen on/off events and their duration.

Additional behavioral and system-related features included power status (on/off events), network connectivity changes, battery level and temperature, mobile data traffic, and WiFi signal characteristics. In particular, for WiFi, we calculated distances between received signal strength indicators (RSSI) and measured the Jaccard similarity between consecutive BSSID sets. Media-related behavior was tracked by counting image, video, and general media access events. Lastly, system status indicators such as ringer mode, power saving state, and charging status were also included to capture contextual cues related to user state.

For wearable data, feature extraction was based on two time windows: a 60-minute window preceding each label timestamp for general physiological and behavioral signals (e.g., heart rate and activity levels), and a 24-hour window for aggregating sleep-related data recorded by the Fitbit.

C. Model Training

In our experiments, we evaluated a range of traditional machine learning (ML) and deep learning (DL) models to detect depression and anxiety from multimodal sensor data. Specifically, we trained seven traditional ML models: Decision Tree, Random Forest, AdaBoost, XGBoost, Linear Discriminant Analysis (LDA), k-Nearest Neighbor (kNN), and Support Vector Machine (SVM). Additionally, we implemented four baseline deep learning models on early-fused inputs: a one-dimensional convolutional neural network (1D CNN), a convolutional long short-term memory network (ConvLSTM), a long short-term memory network (LSTM), and a Transformer encoder. In addition, we implemented a modality-specific fusion model with separate encoders for each modality and audio-specific attention. Given that our inputs are feature-engineered tabular representations rather than raw time-series signals, we also evaluated deep learning models specifically designed for tabular data, including TabNet and TabTransformer.

For the ML and early-fusion DL models, features from four modalities were concatenated into a vector. For 1D-CNN input, each batch was reshaped from $[B, L]$ to $[B, 1, L]$ by adding a singleton channel, where B is the batch size and L is the number of input features, so convolution operated along the feature axis. In modality-specific fusion, the same reshaping was applied to each modality ($[B, L_r]$ to $[B, 1, L_r]$), where L_r denotes the number of features in modality r. Each modality was then processed by its corresponding 1D-CNN encoder; self-attention was additionally applied to voice features, and the resulting embeddings were concatenated for MLP classification [46], [47].The architecture can be formalized as follows. Each modality input $\mathbf{x}^{(r)} \in \mathbb{R}^{L_r}$ is processed by a stack of 1D convolutional layers with increasing channels. Each convolution is defined as:

$$h_{t,k}^{(l)} = \phi\left(\sum_{c}\sum_{\tau=0}^{2} W_{k,c,\tau}^{(l)} h_{t+\tau-1,c}^{(l-1)} + b_k^{(l)}\right), \tag{1}$$

where $\phi(\cdot)$ denotes ReLU, followed by batch normalization and max-pooling (stride 2), yielding a latent sequence $\mathbf{H}^{(r)} \in \mathbb{R}^{T\times d}$.

where $\phi(\cdot)$ denotes ReLU. Convolutions used a kernel size of 3, a stride of 1, and zero padding of 1 (“same” padding), followed by batch normalization and max-pooling (stride 2), producing $\mathbf{H}^{(r)}$.For voice data, we apply multi-head self-attention to the CNN output:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V, \tag{2}$$

with $Q = \mathbf{H}^{(v)}W_Q$, $K = \mathbf{H}^{(v)}W_K$, $V = \mathbf{H}^{(v)}W_V$.

Each modality branch produces an embedding $z^{(r)} \in \mathbb{R}^{32}$. For multimodal fusion, embeddings are concatenated and passed to a shallow MLP:

$$\hat{y} = \sigma\left(f\left([z^{(1)}; \ldots ; z^{(M)}]\right)\right), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, and the model is trained using binary cross-entropy loss.

To improve model training under limited labeled data, we adopted distinct strategies for ML and DL models. For ML models, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training partition within each cross-validation fold For DL models, we used Mixup augmentation [48] with $\alpha = 1.0$, which generates synthetic samples by linearly interpolating pairs of training instances and their labels. Mixup improves robustness by smoothing decision boundaries and reducing overfitting, which is useful for mental health detection where domain shifts and limited data are common. It promotes better generalization and helps maintain performance across diverse user populations.

D. Evaluation

To evaluate model performance, we used the AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) and accuracy. AUC-ROC was chosen for its robustness to class imbalance, while accuracy served as a complementary measure of overall classification performance.

In the generalized setting, we trained a single model across data from all participants to learn population-level patterns. For each fold in leave-one-subject-out (LOSO) cross-validation, data from one participant was held out as the test set, and the model was trained on the remaining participants. In contrast, the personalized models were trained separately for each participant using only their own data. We performed stratified 5-fold cross-validation within each individual’s dataset to evaluate the model’s ability to capture personalized behavioral-mood associations, ensuring that the class proportions in each fold matched the original data distribution.

Both model types used the same input feature sets and model architecture, but generalized models leveraged cross-subject variability, whereas personalized models aimed to capture within-subject behavioral regularities. Feature normalization and imputation were also applied separately within each evaluation context. All metrics were averaged across folds according to the corresponding evaluation setting.

E. Statistical Analysis

To assess the reliability of our results and account for small per-user sample sizes in personalized settings, we conducted a permutation test [49]. For each of 1,000 permutations, labels were shuffled within each participant and the full stratified 5-fold pipeline was rerun, including fold-specific imputation, feature selection, normalization, and SMOTE on training folds only. The empirical p-value was calculated as the proportion of permuted AUROCs at least as high as the observed AUROC.In addition, we evaluated performance differences across data-source configurations using paired Wilcoxon signed-rank tests, with Holm correction applied to multiple comparisons within each outcome and evaluation setting.

V. RESULTS

This section reports the overall performance of generalized and personalized models for detecting depression and anxiety. We then evaluate how data source combinations affect model performance and analyze key features driving the predictions.

A. How accurate is depression and anxiety detection?

For depression detection, the modality-specific fusion model performed best in the generalized setting, with an AUROC of

0.690 and an accuracy of 0.638. Among traditional machine learning models, k-Nearest Neighbors (kNN) achieved the highest performance, with an AUROC of 0.663. In contrast, tree-based models such as Random Forest, Decision Tree, and XGBoost showed relatively poor performance, with AUROC scores falling below 0.5, suggesting limited generalizability in this context.

In the personalized setting, overall performance significantly improved. XGBoost achieved the highest AUROC at 0.736, while Random Forest yielded the highest accuracy at 0.744. Most models showed clear gains in both metrics, indicating that personalization better captures individual behavioral patterns relevant to depression. Permutation tests confirmed that these performance improvements were statistically significant compared to the permutation-based null distribution ($p < 0.01$), indicating that the personalized models performed above chance levels.

For anxiety detection, the overall performance was generally lower than that for depression detection, particularly in the generalized setting. The 1D-CNN model outperformed others, achieving an AUROC of 0.634 and an accuracy of 0.672, followed by the modality-specific fusion model with slightly lower metrics (AUROC 0.617, accuracy 0.634). Nonetheless, the modality-specific fusion model still outperformed all traditional machine learning models in the generalized setting.

In the personalized setting for anxiety detection, the performance pattern resembled that of depression. Random Forest and XGBoost achieved the highest AUROC (0.728 and 0.692, respectively) and accuracy (0.748 and 0.705), again demonstrating that personalization significantly enhances predictive performance. Interestingly, unlike in the generalized setting, the modality-specific fusion model did not show performance advantages in the personalized context for either depression or anxiety. Consistent with the depression results, permutation tests showed that the personalized anxiety models achieved statistically significant performance relative to chance ($p < 0.01$).

Overall, these results highlight two key findings: (1) 1D-CNN-based deep learning models (i.e., 1D-CNN and modality-specific fusion models) are particularly effective in generalized settings, where inter-individual variability is high, and (2) traditional machine learning models, especially tree-based approaches (i.e., Random Forest and XGBoost), can perform competitively when trained in a personalized manner that leverages individual behavioral signatures.

TABLE I: Performance of generalized and personalized models for depression detection

Model	AUROC	Accuracy
Generalized Model		
Decision Tree	0.431 (0.11)	0.459 (0.10)
Random Forest	0.482 (0.17)	0.524 (0.14)
AdaBoost	0.482 (0.18)	0.471 (0.09)
XGBoost	0.464 (0.14)	0.490 (0.08)
LDA	0.491 (0.10)	0.502 (0.13)
kNN	0.663 (0.13)	0.574 (0.11)
SVM	0.617 (0.15)	0.584 (0.12)
1D-CNN	0.543 (0.17)	0.559 (0.18)
GRU	0.510 (0.15)	0.539 (0.14)
LSTM	0.495 (0.12)	0.577 (0.13)
ConvLSTM	0.448 (0.12)	0.430 (0.21)
TabNet	0.571 (0.15)	0.485 (0.14)
TabTransformer	0.503 (0.11)	0.488 (0.12)
Modality-Specific Fusion*	**0.690 (0.11)**	**0.638 (0.15)**
Personalized Model		
Decision Tree	0.610 (0.15)	0.660 (0.16)
Random Forest	0.731 (0.15)	**0.744 (0.11)**
AdaBoost	0.676 (0.13)	0.683 (0.13)
XGBoost	**0.736 (0.15)**	0.701 (0.15)
LDA	0.644 (0.16)	0.661 (0.17)
kNN	0.667 (0.12)	0.716 (0.11)
SVM	0.533 (0.22)	0.705 (0.13)
1D-CNN	0.679 (0.16)	0.660 (0.10)
GRU	0.609 (0.26)	0.660 (0.19)
LSTM	0.604 (0.25)	0.698 (0.15)
ConvLSTM	0.497 (0.23)	0.692 (0.14)
TabNet	0.530 (0.29)	0.502 (0.20)
TabTransformer	0.671 (0.26)	0.655 (0.22)
Modality-Specific Fusion*	0.617 (0.15)	0.634 (0.12)

*The modality-specific fusion model uses separate 1D-CNN encoders for smartphone, wearable, and IoT data, and an attention-based 1D-CNN for voice data.

B. Which data sources contribute most to detection?

To examine which data sources and behavioral features contribute most to detecting depression and anxiety, we conducted two complementary analyses: ablation studies and feature attribution analysis.

First, we conducted ablation experiments using the best-performing models identified in earlier sections. Specifically, we used the modality-specific fusion model for generalized depression detection, the 1D-CNN model for generalized anxiety detection, and Random Forest for all personalized models. The configurations were selected based on sensing context to represent three levels of multimodal integration: four single-modality baselines, two within-context pairs (IoT + Voice for home sensing and Mobile + Wearable for personal-device sensing), and the full cross-context multimodal configuration.n the generalized depression model, the modality-specific fusion model achieved the highest performance when all data sources were combined, with an AUROC of 0.690 and an accuracy of 0.638. Using IoT and voice data in combination resulted in lower performance (AUROC 0.546, accuracy 0.553), compared to the combination of mobile and wearable data (AUROC 0.584, accuracy 0.577). When used as single sources, both IoT and voice data yielded even poorer results, with AUROC scores of 0.519 and 0.491, respectively. Additionally, paired Wilcoxon signed-rank tests showed that using all data sources improved AUROC over Mobile (ΔAUROC = 0.111, adjusted $p = 0.032$) and Mobile + Wearable (ΔAUROC = 0.106, adjusted $p = 0.034$), supporting the benefit of broader multimodal fusion.

A similar trend was observed in the personalized model. Using all data sources yielded the highest performance, with an AUROC of 0.731 and an accuracy of 0.744. Among two-source combinations, mobile and wearable data achieved better performance (AUROC 0.704, accuracy 0.740) than the combination of IoT and voice data. However, the improvements over Mobile (ΔAUROC = 0.030) and Mobile + Wearable (ΔAUROC = 0.028) were not statistically significant after

TABLE II: Performance of generalized and personalized models for anxiety detection

Model	AUROC	Accuracy
Generalized Model		
Decision Tree	0.520 (0.07)	0.569 (0.09)
Random Forest	0.550 (0.13)	0.539 (0.08)
AdaBoost	0.531 (0.12)	0.564 (0.11)
XGBoost	0.598 (0.16)	0.599 (0.11)
LDA	0.500 (0.12)	0.512 (0.11)
kNN	0.611 (0.10)	0.579 (0.15)
SVM	0.588 (0.18)	0.660 (0.15)
1D-CNN	**0.634 (0.17)**	**0.672 (0.12)**
GRU	0.542 (0.14)	0.511 (0.15)
LSTM	0.552 (0.12)	0.553 (0.12)
ConvLSTM	0.461 (0.10)	0.486 (0.22)
TabNet	0.525 (0.12)	0.511 (0.10)
TabTransformer	0.525 (0.10)	0.536 (0.11)
Modality-Specific Fusion*	0.617 (0.15)	0.634 (0.12)
Personalized Model		
Decision Tree	0.620 (0.12)	0.684 (0.14)
Random Forest	**0.728 (0.15)**	**0.748 (0.15)**
AdaBoost	0.625 (0.14)	0.671 (0.15)
XGBoost	0.692 (0.16)	0.705 (0.15)
LDA	0.661 (0.17)	0.689 (0.16)
kNN	0.690 (0.13)	0.710 (0.14)
SVM	0.673 (0.16)	0.723 (0.13)
1D-CNN	0.700 (0.14)	0.697 (0.13)
GRU	0.593 (0.21)	0.647 (0.20)
LSTM	0.561 (0.24)	0.660 (0.19)
ConvLSTM	0.461 (0.10)	0.486 (0.22)
TabNet	0.493 (0.22)	0.492 (0.17)
TabTransformer	0.627 (0.26)	0.702 (0.19)
Modality-Specific Fusion*	0.650 (0.16)	0.657 (0.12)

*The modality-specific fusion model uses separate 1D-CNN encoders for smartphone, wearable, and IoT data, and an attention-based 1D-CNN for voice data.

Holm correction (both adjusted $p = 0.332$).In anxiety detection, using all data sources led to the best performance (AUROC 0.634 in the generalized model and 0.728 in the personalized model), followed by mobile data (AUROC 0.599 and 0.705, respectively). For the generalized anxiety model, the improvement over Mobile was small and not statistically significant (ΔAUROC = 0.035, $p = 0.227$). In the personalized setting, using all data sources significantly improved AUROC over Mobile (ΔAUROC = 0.023, adjusted $p = 0.025$) and Mobile + Wearable (ΔAUROC = 0.038, adjusted $p = 0.0022$).Second, we used SHapley Additive exPlanations (SHAP), a model-agnostic framework for explaining individual predictions to interpret the model's predictions and identify dominant features. SHAP assigns a contribution score to each feature for each prediction, representing the weighted marginal impact of that feature on the model's output. These scores enable a quantitative assessment of feature importance. Using SHAP summary plots, we visualized the distribution of Shapley values across features. In these plots, the x-axis indicates the Shapley value (i.e., the impact on the model's output), while each point represents a data instance and is colored by the original feature value (with red indicating higher values and blue indicating lower values). For example, when red points are concentrated on the right side of the plot, this suggests that high feature values contribute positively to the prediction. Conversely, if blue points appear on the right, it indicates that lower feature values have a positive influence.

TABLE III: Depression detection performance across representative data source configurations.

Data Source Combination	AUROC	Accuracy
Generalized Model (Modality-specific Fusion)		
IoT	0.519 (0.12)	0.487 (0.13)
Voice	0.491 (0.09)	0.458 (0.18)
Mobile	0.579 (0.13)	0.603 (0.18)
Wearable	0.546 (0.11)	0.547 (0.15)
IoT + Voice	0.546 (0.08)	0.553 (0.15)
Mobile + Wearable	0.584 (0.15)	0.577 (0.17)
Mobile + Wearable + IoT + Voice*	**0.690 (0.11)**	**0.638 (0.15)**
Personalized Model (Random Forest)		
IoT	0.669 (0.17)	0.719 (0.15)
Voice	0.593 (0.15)	0.714 (0.09)
Mobile	0.701 (0.15)	0.738 (0.13)
Wearable	0.580 (0.12)	0.706 (0.15)
IoT + Voice	0.685 (0.15)	0.714 (0.11)
Mobile + Wearable	0.704 (0.14)	0.740 (0.13)
Mobile + Wearable + IoT + Voice	**0.731 (0.15)**	**0.744 (0.11)**

*Significant improvements over Mobile and Mobile + Wearable after Holm correction (adjusted $p = 0.032$ and 0.034, respectively).

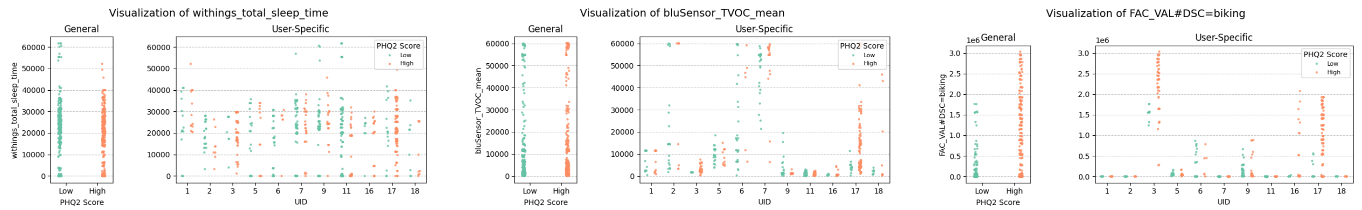
TABLE IV: Anxiety detection performance across representative data source configurations.

Data Source Combination	AUROC	Accuracy
Generalized Model (Conv 1D)		
IoT	0.584 (0.10)	0.574 (0.15)
Voice	0.515 (0.13)	0.504 (0.14)
Mobile	0.599 (0.20)	0.621 (0.13)
Wearable	0.485 (0.12)	0.465 (0.21)
IoT + Voice	0.560 (0.07)	0.590 (0.13)
Mobile + Wearable	0.588 (0.20)	0.624 (0.17)
Mobile + Wearable + IoT + Voice	**0.634 (0.17)**	**0.672 (0.12)**
Personalized Model (Random Forest)		
IoT	0.649 (0.22)	0.696 (0.19)
Voice	0.613 (0.13)	0.674 (0.16)
Mobile	0.705 (0.15)	0.745 (0.16)
Wearable	0.553 (0.15)	0.669 (0.15)
IoT + Voice	0.635 (0.19)	0.681 (0.18)
Mobile + Wearable	0.690 (0.18)	0.735 (0.17)
Mobile + Wearable + IoT + Voice*	**0.728 (0.15)**	**0.748 (0.15)**

*Significant improvements over Mobile and Mobile + Wearable after Holm correction (adjusted $p = 0.025$ and 0.0022, respectively).

As shown in Figure 6, we extracted the top 20 most dominant features for a generalized depression detection model using the modality-specific fusion method. Among the top-ranked features, we observed a diverse set of variables from mobile, wearable, IoT, and voice data. Key features included bluSensor_TVOC_mean (indoor air quality), sleep-related metrics (e.g., withings_total_timeinbed, withings_total_sleep_time), voice features (MFCC, duration, word_count), and mobile activity indicators such as FAC_VAL#DSC=biking.

To further understand how these features relate to depressive symptoms, we analyzed the distribution of each feature on a per-user basis when participants exhibited high versus low PHQ-2 scores. For the total sleep time feature, participants with higher PHQ-2 scores tended to have shorter total sleep durations. While this pattern was generally observed, some



(a) Total sleep time (b) Indoor air quality (TVOC) (c) Bicycling activity

Fig. 5: Distributions of behavioral and environmental features for individual participants, grouped by high and low PHQ-2 scores.

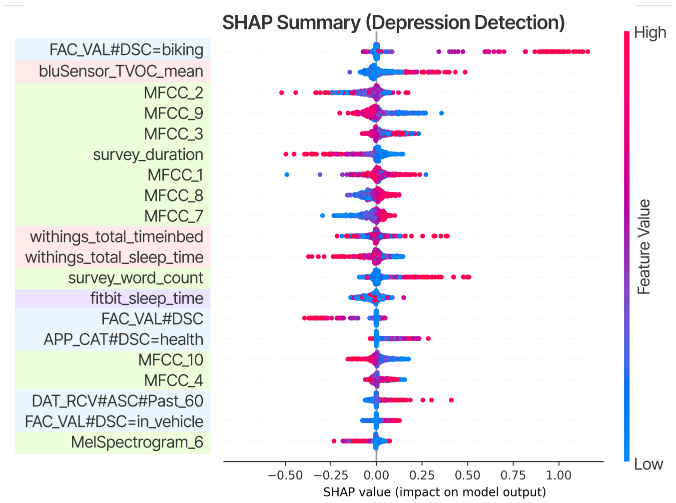


Fig. 6: Top 20 most important features for depression detection, based on SHAP values from the modality-specific fusion model

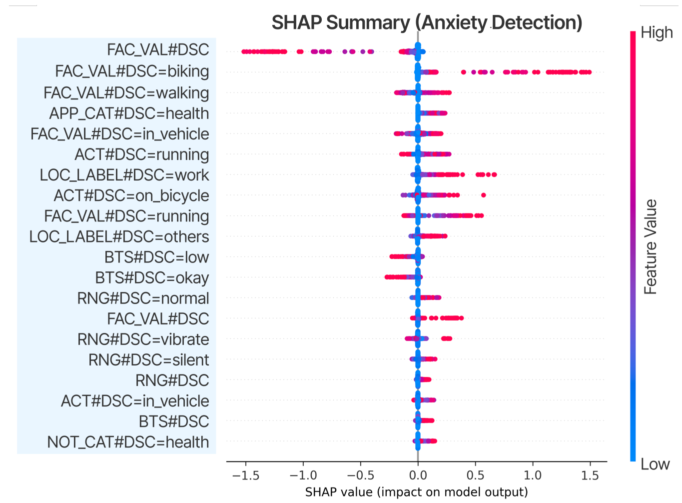


Fig. 7: Top 20 most important features for anxiety detection, based on SHAP values from the CNN model

individuals showed inconsistent relationships between sleep duration and PHQ-2 scores. For instance, UID 1 exhibited high variability in sleep duration, with no clear distinction between high and low PHQ-2 periods (Figure 5a). TVOC (Total Volatile Organic Compounds), a marker of indoor air pollution, also appeared as a significant predictor. In some participants, such as UID 17, higher PHQ-2 scores were accompanied by elevated TVOC values (Figure 5b). In mobile data, the feature FAC_VAL#DSC=biking (bicycling activity) was identified as one of the top-ranked predictors for depression detection. This feature was primarily associated with a subset of participants, such as UIDs 3, 16, and 17, who showed higher frequencies of biking during high PHQ-2 periods. In contrast, other participants, including UIDs 1, 2, and 7, recorded no biking activity during the study period (Figure 5c). These patterns indicate considerable individual differences in activity levels related to this feature.

For the generated anxiety detection model, we extracted the top 20 most important features using the 1D-CNN model. Most of the top-ranked features were derived from mobile data, such as fitness activity and app usage, whereas IoT-based features did not appear among the highest contributors (Figure 7).

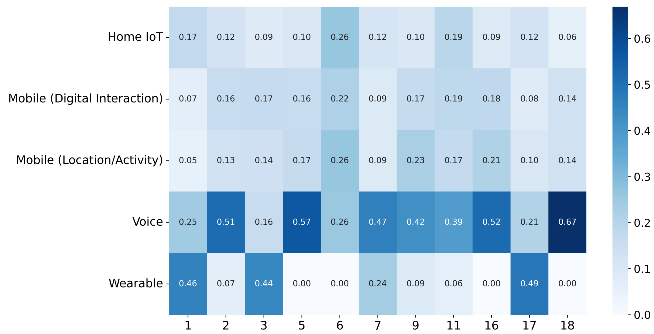
To better understand the modality-level contribution within personalized models, we analyzed the relative importance of each data source using SHAP based on the Random Forest model, which demonstrated the highest performance across all personalized detection tasks. Figure 8 illustrates the within-user modality contributions, where SHAP values were aggregated and normalized for each participant.

For depression detection, we observed a high degree of heterogeneity in modality usage across individuals. While some users, such as UID 18, showed a dominant dependency on voice data, others relied more heavily on different sources, such as UID 1 with wearable and UID 6 with a balanced reliance across multiple sources. Anxiety detection exhibited a more consistent reliance on specific modalities across participants. Voice data emerged as a dominant source for a majority of participants, with UIDs 1, 2, 4, 7, and 8 all showing voice contributions exceeding 0.50. However, individual variations still existed, such as UID 3, who showed a heavy reliance on wearable data.

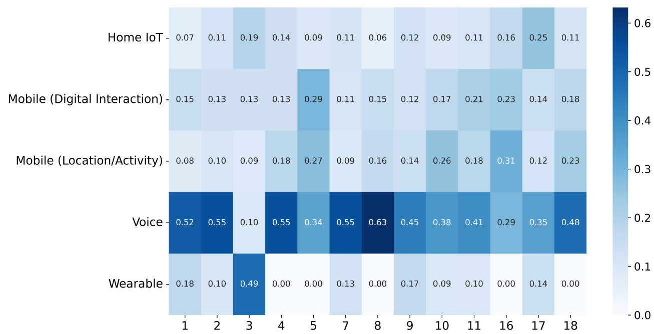
VI. Discussion

A. Analysis of Model Performance

Across our experiments, no single model consistently performed well across all users and settings. In the generalized setting, a deep learning model achieved the highest classification performance, suggesting its potentialo capture complex



(a) Depression (PHQ-2)



(b) Anxiety (GAD-2)

Fig. 8: User-specific relative SHAP contributions across modalities for depression (PHQ-2) and anxiety (GAD-2) detection.

patterns by integrating heterogeneous data sources such as audio and behavioral features [50].

However, the overall performance of generalized models remained limited. Consistent with findings from prior studies [42], [51], this limited performance may be partly attributed to domain shifts arising from interpersonal differences in behavior and environment. In particular, the below-0.5 AUROCs of tree-based models may reflect difficulty in transferring patterns learned from training participants to held-out participants under domain shifts. The lower generalized anxiety performance may partly reflect differences in the relevance of sensing features across outcomes [52]. Although our dataset included heart rate and diverse behavioral and contextual features, anxiety detection may benefit from additional physiological signals such as heart rate variability and electrodermal activity [53], which were not available. Their absence may have limited the representation of anxiety-related variation, partly explaining the lower performance. Given the small cohorts (11 participants for depression and 13 for anxiety), statistical power and generalizability remain limited, and overfitting remains possible. We therefore interpret these findings as preliminary until validated in larger cohorts.In the personalized setting, deep learning models performed poorly due to data scarcity, while tree-based models such as Random Forest and XGBoost outperformed them by effectively selecting informative features [54]. To improve deep learning performance in such settings, techniques like pre-training or transfer learning can be considered.

B. Ablation Study & Interpretation of Dominant Features

Across both generalized and personalized settings, ablation results showed consistent performance gains when multiple data modalities were combined. Although IoT and voice data underperformed mobile data when used in isolation, they contributed complementary information when fused with other modalities. To better understand which behavioral signals drove these gains and how different modalities contributed to model predictions, we conducted feature-level interpretation using SHAP.

1) Feature-Level Interpretability in Clinical Contexts: Although depression and anxiety scores were highly correlated in our dataset ($r = 0.83$), the generalized models relied on different dominant feature categories for each condition.

For depression, SHAP analysis identified sleep-related, voice-related, and environment-related features as key contributors. Sleep-related features (e.g., time in bed and total sleep duration) align with DSM-5 criteria describing sleep disturbance as a core depressive symptom [19]. Voice features (e.g., MFCCs and speech duration) may reflect reduced expressiveness or psychomotor slowing, consistent with prior speech-based depression studies [23], [24]. Environment-related features, particularly TVOC, emerged as important contributors, suggesting links between indoor conditions, depressive risk [55], and accompanying behavioral patterns such as prolonged indoor stay and reduced ventilation.

In contrast, anxiety-related patterns were more prominently reflected in mobile-derived features, consistent with clinical characterizations of anxiety as involving heightened arousal and motor tension. Activity and mobility fluctuations captured via smartphones may therefore serve as observable correlates of anxiety-related distress, highlighting a more modality-specific expression compared to depression.

2) Individual Differences in Modality Contributions: While population-level feature analysis highlights dominant signals in generalized models, it can obscure substantial heterogeneity across individuals. For example, one of the top-ranked mobile features for depression detection, FAC_VAL#DSC=biking, was primarily driven by a subset of participants (e.g., UID 3, 16, and 17), whereas other participants showed no biking activity throughout the study period. This illustrates that globally important features may reflect highly specific behavioral patterns present only in certain individuals.

To better characterize inter-individual differences at a higher level of abstraction, we analyzed modality-level contributions for each individual. As shown in Figure 8, the relative importance of data modalities varied substantially across users, indicating that individuals relied on distinct sensing modalities to infer depressive states.

For anxiety detection, population-level feature ranking was dominated by mobile-derived features, reflecting patterns broadly consistent across users. However, within-user analysis revealed that voice features also contributed meaningfully for many individuals once models were personalized, suggesting that signals not prominent at the population level may still be informative for specific users.

Overall, multimodal data support generalized models by accommodating inter-individual variability and personalized models by highlighting user-specific informative modalities, reinforcing the need for multimodal sensing beyond a single disorder-specific marker.

C. Toward Detection as Clinical Decision Support Systems

This study demonstrates the feasibility of detecting mental health conditions using commercially available mobile, wearable, and home IoT devices in real-world settings. While these results are promising, future systems should move beyond detection toward identifying early warning signs and supporting timely interventions within clinical workflows.

In this context, we adopt a *human-in-the-loop* perspective, where sensing-based outputs are not treated as autonomous clinical decisions but as decision-support cues for clinicians and patients. For example, longitudinal behavioral summaries derived from approximately two weeks of sensing data (e.g., sleep regularity, voice activity, indoor stay or ventilation patterns) could support symptom assessment prior to clinical visits, reducing reliance on retrospective self-report [56]. This is particularly relevant for depression and anxiety, where changes in daily functioning are core diagnostic considerations [19] yet are often difficult for patients to consistently articulate.

Following treatment or counseling, similar behavioral summaries may help contextualize symptom changes by tracking shifts in daily routines, such as stabilization of sleep patterns or changes in vocal expressiveness. These signals are not intended as standalone indicators of treatment efficacy, but as quantitative cues to support clinician interpretation.

Finally, deployment in clinical settings requires careful attention to privacy and data governance, given the sensitivity of mental health data. Privacy-aware system design, including considerations of data access, consent, and secondary exposure risks, remains essential in IoT-based health monitoring [57], [58].

VII. Conclusion

We investigated the feasibility of detecting depression and anxiety using a multimodal sensor fusion approach that integrates data from mobile phones, wearable devices, and home IoT sensors. Through a one-month data collection study involving 20 single-person households, we showed that combining multiple data sources improves detection performance compared to using a single source alone. Our results show that 1D-CNN models performed best in generalized settings by capturing diverse signals from behavior, physiology, environment, and speech. In contrast, tree-based models outperformed in personalized settings by fitting individual-level patterns. In particular, we found that home IoT data, such as sleep and environmental sensing, proved valuable for detecting depression, demonstrating the potential of ambient sensing in mental health monitoring. This work contributes to expanding the scope of mental health sensing beyond traditional mobile and wearable data by incorporating unobtrusive home IoT sensing.

References

[1] P. Cuijpers, A. Javed, and K. Bhui, "The who world mental health report: a call for action," *The British Journal of Psychiatry*, vol. 222, no. 6, pp. 227–229, 2023.

[2] World Health Organization, "Who highlights urgent need to transform mental health and mental health care," 2022, accessed: 2025-03-14. [Online]. Available: https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care

[3] A. Althubaiti, "Information bias in health research: definition, pitfalls, and adjustment methods," *Journal of multidisciplinary healthcare*, pp. 211–217, 2016.

[4] R. M. Epstein, P. R. Duberstein, M. D. Feldman, A. B. Rochlen, R. A. Bell, R. L. Kravitz, C. Cipri, J. D. Becker, P. M. Bamonti, and D. A. Paterniti, ""i didn't know what was wrong:" how people with undiagnosed depression recognize, name and explain their distress," *Journal of general internal medicine*, vol. 25, pp. 954–961, 2010.

[5] Y. Hu, J. Chen, J. Chen, W. Wang, S. Zhao, and X. Hu, "An ensemble classification model for depression based on wearable device sleep data," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 5, pp. 2602–2612, 2023.

[6] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 140–148, 2014.

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

[8] M. Tlachac, R. Flores, M. Reisch, R. Kayastha, N. Taurich, V. Melican, C. Bruneau, H. Caouette, J. Lovering, E. Toto *et al.*, "Studentsadd: rapid mobile depression and suicidal ideation screening of college students during the coronavirus pandemic," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–32, 2022.

[9] A. K. Dey, K. Wac, D. Ferreira, K. Tassini, J.-H. Hong, and J. Ramos, "Getting closer: an empirical investigation of the proximity of user to their smart phones," in *Proceedings of the 13th international conference on Ubiquitous computing*, 2011, pp. 163–172.

[10] H. Jeong, H. Kim, R. Kim, U. Lee, and Y. Jeong, "Smartwatch wearing behavior analysis: a longitudinal study," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–31, 2017.

[11] L. MacLeod, B. Suruliraj, D. Gall, K. Bessenyei, S. Hamm, I. Romkey, A. Bagnell, M. Mattheisen, V. Muthukumaraswamy, R. Orji *et al.*, "A mobile sensing app to monitor youth mental health: observational pilot study," *JMIR mHealth and uHealth*, vol. 9, no. 10, p. e20638, 2021.

[12] S. Zamani, R. Sinha, M. Nguyen, and S. Madanian, "Enhancing emotional well-being with iot data solutions for depression: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[13] A. Moglia, K. Georgiou, B. Marinov, E. Georgiou, R. N. Berchiolli, R. M. Satava, and A. Cuschieri, "5g in healthcare: from covid-19 to future challenges," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4187–4196, 2022.

[14] A. Alberdi, A. Weakley, M. Schmitter-Edgecombe, D. J. Cook, A. Aztiria, A. Basarab, and M. Barrenechea, "Smart home-based prediction of multidomain symptoms related to alzheimer's disease," *IEEE journal of biomedical and health informatics*, vol. 22, no. 6, pp. 1720–1731, 2018.

[15] P. P. Morita, K. S. Sahu, and A. Oetomo, "Health monitoring using smart home technologies: scoping review," *JMIR mHealth and uHealth*, vol. 11, p. e37347, 2023.

[16] D. Wu and F. Liu, "Assessment of the relationship between living alone and the risk of depression based on longitudinal studies: a systematic review and meta-analysis," *Frontiers in psychiatry*, vol. 13, p. 954857, 2022.

[17] P. N. Cohen, "The rise of one-person households," *Socius*, vol. 7, p. 23780231211062315, 2021.

[18] D. J. Cook, M. Schmitter-Edgecombe, L. Jönsson, and A. V. Morant, "Technology-enabled assessment of functional health," *IEEE reviews in biomedical engineering*, vol. 12, pp. 319–332, 2018.

[19] D. American Psychiatric Association, A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013, vol. 5, no. 5.

[20] M. C. Piotrowski, J. Lunsford, and B. N. Gaynes, "Lifestyle psychiatry for depression and anxiety: Beyond diet and exercise," *Lifestyle Medicine*, vol. 2, no. 1, p. e21, 2021.

[21] J. Firth, M. Solmi, R. E. Wootton, D. Vancampfort, F. B. Schuch, E. Hoare, S. Gilbody, J. Torous, S. B. Teasdale, S. E. Jackson *et al.*, "A meta-review of "lifestyle psychiatry": the role of exercise, smoking, diet and sleep in the prevention and treatment of mental disorders," *World psychiatry*, vol. 19, no. 3, pp. 360–380, 2020.

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

[23] 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.

[24] L.-M. Wadle, U. W. Ebner-Priemer, J. C. Foo, Y. Yamamoto, F. Streit, S. H. Witt, J. Frank, L. Zillich, M. F. Limberger, A. Ablimit *et al.*, "Speech features as predictors of momentary depression severity in patients with depressive disorder undergoing sleep deprivation therapy: ambulatory assessment pilot study," *JMIR mental health*, vol. 11, p. e49222, 2024.

[25] Z. Lin, Y. Wang, Y. Zhou, F. Du, and Y. Yang, "Mlm-eoe: Automatic depression detection via sentimental annotation and multi-expert ensemble," *IEEE Transactions on Affective Computing*, vol. 16, no. 4, pp. 2842–2858, 2025.

[26] J. Radua, A. Pertusa, and N. Cardoner, "Climatic relationships with specific clinical subtypes of depression," *Psychiatry research*, vol. 175, no. 3, pp. 217–220, 2010.

[27] M. L. Lee and A. K. Dey, "Sensor-based observations of daily living for aging in place," *Personal and Ubiquitous Computing*, vol. 19, pp. 27–43, 2015.

[28] S. Andreadis, T. G. Stavropoulos, G. Meditskos, and I. Kompatsiaris, "Dem@ home: Ambient intelligence for clinical support of people living with dementia," in *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13*. Springer, 2016, pp. 357–368.

[29] A. Akl, J. Snoek, and A. Mihailidis, "Unobtrusive detection of mild cognitive impairment in older adults through home monitoring," *IEEE journal of biomedical and health informatics*, vol. 21, no. 2, pp. 339–348, 2015.

[30] J. Lim, Y. Koh, A. Kim, and U. Lee, "Exploring context-aware mental health self-tracking using multimodal smart speakers in home environments," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–18.

[31] M. B. Morshed, S. S. Kulkarni, K. Saha, R. Li, L. G. Roper, L. Nachman, H. Lu, L. Mirabella, S. Srivastava, K. De Barbaro *et al.*, "Food, mood, context: examining college students' eating context and mental well-being," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 4, pp. 1–26, 2022.

[32] F. Bentley, K. Tollmar, P. Stephenson, L. Levy, B. Jones, S. Robertson, E. Price, R. Catrambone, and J. Wilson, "Health mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 20, no. 5, pp. 1–27, 2013.

[33] A. Kurze, A. Bischof, S. Totzauer, M. Storz, M. Eibl, M. Brereton, and A. Berger, "Guess the data: data work to understand how people make sense of and use simple sensor data from homes," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing systems*, 2020, pp. 1–12.

[34] B. Löwe, K. Kroenke, and K. Gräfe, "Detecting and monitoring depression with a two-item questionnaire (phq-2)," *Journal of psychosomatic research*, vol. 58, no. 2, pp. 163–171, 2005.

[35] K. Kroenke, R. L. Spitzer, J. B. Williams, P. O. Monahan, and B. Löwe, "Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection," *Annals of internal medicine*, vol. 146, no. 5, pp. 317–325, 2007.

[36] S. Kang, W. Choi, C. Y. Park, N. Cha, A. Kim, A. H. Khandoker, L. Hadjileontiadis, H. Kim, Y. Jeong, and U. Lee, "K-emophone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels," *Scientific data*, vol. 10, no. 1, p. 351, 2023.

[37] Korea Communications Commission. (2024) Kcc announces 2024 broadcast media consumption behavior survey results. Press release. [Online]. Available: https://www.kcc.go.kr/user.do?boardId=1058&page=E04010000&dc=E04010000&boardSeq=65045&mode=view

[38] A. J. Hughes, K. M. Dunn, T. Chaffee, J. J. Bhattarai, and M. Beier, "Diagnostic and clinical utility of the gad-2 for screening anxiety symptoms in individuals with multiple sclerosis," *Archives of physical medicine and rehabilitation*, vol. 99, no. 10, pp. 2045–2049, 2018.

[39] S. T. Sedeeq, M. J. M. AlTamimi, E. Hamed, and M. A. Syed, "Diagnostic accuracy of the patient health questionnaire 2 (phq-2) in qatar's primary care settings," *Primary Health Care Research & Development*, vol. 23, p. e5, 2022.

[40] T. Stütz, T. Kowar, M. Kager, M. Tiefengrabner, M. Stuppner, J. Blechert, F. H. Wilhelm, and S. Ginzinger, "Smartphone based stress prediction," in *International conference on user modeling, adaptation, and personalization*. Springer, 2015, pp. 240–251.

[41] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen, and D. Kotz, "Evaluating the reproducibility of physiological stress detection models," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 4, pp. 1–29, 2020.

[42] P. Zhang, G. Jung, J. Alikhanov, U. Ahmed, and U. Lee, "A reproducible stress prediction pipeline with mobile sensor data," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 8, no. 3, pp. 1–35, 2024.

[43] J. Alikhanov, P. Zhang, Y. Noh, and H. Kim, "Design of contextual filtered features for better smartphone-user receptivity prediction," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 11 707–11 722, 2023.

[44] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[45] M. R. Ahmed, S. Islam, A. M. Islam, and S. Shatabda, "An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition," *Expert Systems with Applications*, vol. 218, p. 119633, 2023.

[46] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Multilevel sensor fusion with deep learning," *IEEE sensors letters*, vol. 3, no. 1, pp. 1–4, 2018.

[47] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2020.

[48] H. Zhang, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[49] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance." *Journal of machine learning research*, vol. 11, no. 6, 2010.

[50] K. Yang, C. Wang, Y. Gu, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves, "Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1082–1097, 2021.

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

[52] C. A. Stamatis, J. Meyerhoff, Y. Meng, Z. C. C. Lin, Y. M. Cho, T. Liu, C. J. Karr, T. Liu, B. L. Curtis, L. H. Ungar *et al.*, "Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: a longitudinal cohort study," *Npj Mental Health Research*, vol. 3, no. 1, p. 1, 2024.

[53] M. Elgendi, K. Markov, H. Liu, M. De Vos, K. Khalaf, A. Khandoker, H. F. Jelinek, D. Goswami, and C. Menon, "Wearable devices for anxiety assessment: a systematic review," *Communications Medicine*, vol. 6, no. 1, p. 20, 2026.

[54] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[55] X. Zhang, L. Ding, F. Yang, G. Qiao, X. Gao, Z. Xiong, and X. Zhu, "Association between indoor air pollution and depression: a systematic review and meta-analysis of cohort studies," *BMJ open*, vol. 14, no. 5, p. e075105, 2024.

[56] A. A. Stone, S. Shiffman, J. E. Schwartz, J. E. Broderick, and M. R. Hufford, "Patient compliance with paper and electronic diaries," *Controlled clinical trials*, vol. 24, no. 2, pp. 182–199, 2003.

[57] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in internet of things: The road ahead," *Computer networks*, vol. 76, pp. 146–164, 2015.

[58] R. Abbasi, A. K. Bashir, A. Mateen, F. Amin, Y. Ge, and M. Omar, "Efficient security and privacy of lossless secure communication for sensor-based urban cities," *IEEE Sensors Journal*, vol. 24, no. 5, pp. 5549–5560, 2023.