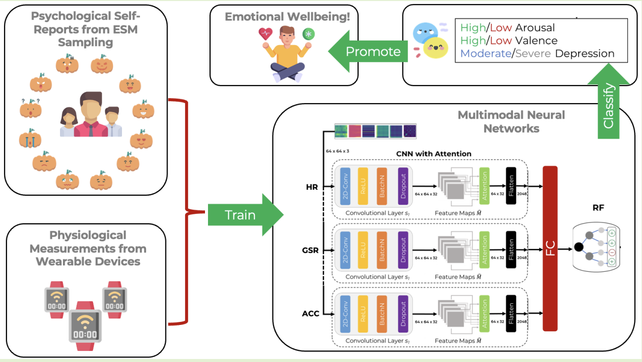
IEEE Sensors Council

Evaluating Multimodal Wearable Sensors for Quantifying Affective States and Depression With Neural Networks

Abdullah Ahmed, Jayroop Ramesh, *Student Member, IEEE*, Sandipan Ganguly, Raafat Aburukba, *Member, IEEE*, Assim Sagahyroon, *Senior Member, IEEE*, and Fadi Aloul, *Senior Member, IEEE*



***Abstract*—With the increasing proliferation of embedded sensors in wearable devices, there is potential for modeling individual emotional and mental state variations. The popular measure for the quantification of emotions outlines the affective states of arousal and valences, with high and low being the discrete categories of interest. Recent works explore the discernability of digital behavior differences between groups with and without mental disorders. However, the interaction between physiological states and affective states within a predominantly depressive population remains to be studied with the aid of wearables. Despite the pervasiveness of emotional state inference through the tracking of ubiquitous physiological trackers, such as heart rate, blood volume pulse, skin conductance, and motion, a dearth of work is noted in the exploration of physiological markers in single-modal and multimodal settings. This work provides an extensive evaluation of a convolutional neural network with an attention mechanism ensembled with a random forest algorithm to effectively leverage multiple raw signal-to-image transformations as feature inputs to predict depression severity and affective state. The proposed models are assessed on the Daily Ambulatory Psychological and Physiological recording for Emotion Research (DAPPER) dataset and achieve the sensitivity: specificity scores of 58.75%:45.59%, 62.34%:43.41%, and 49.43%:51.70% for predicting depression, valence, and arousal with a mixture of unimodality and bimodality applying continuous wavelet transforms and short-time Fourier transform to motion and skin-conductance readings, respectively. This work is envisioned as a preliminary study to contribute toward the monitoring of affective states among a depressed population by utilizing low-frequency sensor recordings with the DAPPER dataset.**

***Index Terms*— Affective computing, deep learning, depression, galvanic skin response (GSR), heart rate (HR) sensors, motion sensors, multimodal, unimodal, wearable devices.**

Manuscript received 4 July 2023; accepted 4 August 2023. Date of publication 11 August 2023; date of current version 2 October 2023. The associate editor coordinating the review of this article and approving it for publication was Dr. Qammer H. Abbasi. *(Abdullah Ahmed, Jayroop Ramesh, and Sandipan Ganguly contributed equally to this work.) (Corresponding author: Jayroop Ramesh.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Local Ethics Committee of the Department of Psychology, Tsinghua University.

Abdullah Ahmed is with the Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA 01003 USA (e-mail: amahmed@umass.edu).

Jayroop Ramesh, Raafat Aburukba, Assim Sagahyroon, and Fadi Aloul are with the Department of Computer Science and Engineering, American University of Sharjah, Sharjah, United Arab Emirates (e-mail: b00057412@aus.edu; raburukba@aus.edu; asagahyroon@aus.edu; faloul@aus.edu).

Sandipan Ganguly was with the Department of Computer Science, University College London, WC1E 6BT London, U.K. He is now with the Independent Researcher, Bengaluru, India (e-mail: sandipanganguly96@gmail.com).

Digital Object Identifier 10.1109/JSEN.2023.3303436

I. Introduction

THE affective states are often largely affected by interindividual differences in prior prevailing mental conditions and personalities, reflect responses to events and contextual stimuli, and offer a way of quantifying emotional dysregulation. There is also the rampant issue of true emotion concealment or suppression by individuals during subjective questionnaires, speech, and/or facial cue analysis, making it a challenge to diagnose during early stages [1]. To help with detection, many prominent biomarkers have been studied in relation to mental health, such as brain connectivity and heart rate (HR) variability [2], [3].

The constant observation and collection of this data with external disparate sensors pose a hurdle for patients' daily activity. In addition to the need for unobtrusive constant monitoring techniques, the utilization of long-term passive readings through a single wearable device is relatively indirect

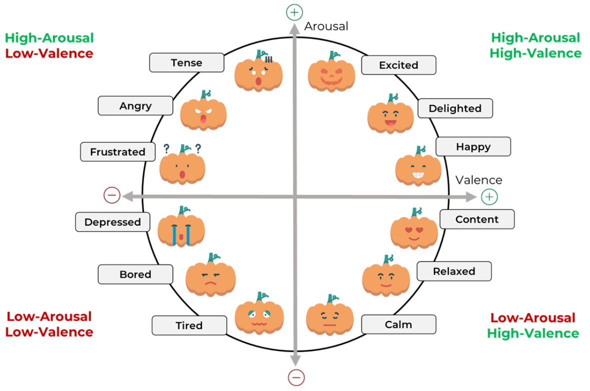


Fig. 1. Russell's 2-D valence-arousal affect model [10].

and more appealing to general users. Thus, this topic presents an interesting research challenge to undertake, where we study the underlying interplay between motion, HR, and skin conductance in predicting emotional states.

Multimodally studying seemingly orthogonal, yet related, markers brings extra dimensions for insight-enhancing analysis and diversity. Diversity is a conceptualization for the principle of multimodal intelligent systems representing the increase in gleaned insights in proportion to the increase in unique modalities [4]. Heart activity captured from wearables is hypothesized to encapsulate the following behavior [5]. Acute stress increases HR and respiration activity, triggering the fight-or-flight response of the autonomic nervous system (ANS). Skin-conductance sensors are expected to measure the level of sweat secretion through transient changes in the skin conductance brought about by affected mood and once again influenced by the sympathetic activation of the ANS. One study details how balance, stability, and posture quality degrade, during movement, with the rise of depressive symptoms [6]. Many exhibiting symptoms of depressive disorder tend to have impaired psychomotor skills and often fall into sedentary lifestyles [7], [8]. It also has been shown that a light increase in physical activity and physiotherapy has been proposed to alleviate mild to moderate depressive cases. This interplay, recorded in literature, between features of motion and depression gives ground to the utility of movement as a marker for depression.

Monitoring depression solely is not sufficient to attain a complete overview of an individual's mental state. The monitoring of immediate psychological affect, valence, and arousal is equally as essential. Affective valence is the measure of how positive or negative an experience is subjectively perceived, while affective arousal rates the activation level of the sympathetic nervous system or an approximation of the level of engagement when self-reported, which is illustrated in Fig. 1 [9].

Psychological affect can also be reliably monitored through motion expression and body movement quality [11]. Some subtle gestures like how a person shrugs their shoulder have been shown to relay affective information [12]. The work in [13] highlights the strong correlation between motion and valence/arousal by monitoring the affect conveyed by dancers through minute movements captured using motion capture technology.

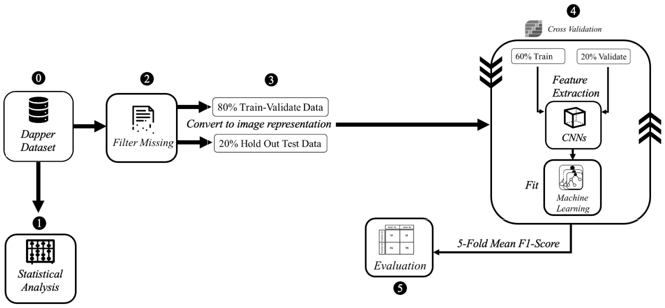


Fig. 2. Flow diagram of end-to-end process.

For unobtrusive constant monitoring of mental state, we propose the instrumentation of a framework that uses motion as the sole marker to derive a comprehensive overview of an individual mental health state using wearable sensors. Wearable sensors confer objectivity in naturalistic settings, which is paramount for a more accurate and complete representation of subconscious reactions and behaviors.

In summary, the primary contributions of this work are given as follows.

1) We aim to quantify and contrast the utility of unimodal and multimodal inference from wearables in providing insight relating to an individual's mental state in terms of high/low valence, high/low arousal, and moderate/severe depression within naturalistic settings.
2) We compare and contrast multiple raw signal-to-image transformations on 30-min long motion data preceding the ESM sampling questionnaire to leverage the spatiotemporal dependence characteristics of the signals.
3) We utilize a convolutional neural network (CNN) with an attention mechanism to highlight regions of interest in the extracted feature maps, connected to a random forest (RF) algorithm for the binary classification to avoid overfitting in the presence of scarce data.

This article is organized as follows. Section II explores the materials and methods. Section III presents the results. Section IV discusses the findings. Section V concludes the work and suggests possible future research directions.

II. Materials and Methods

Recent research has studied the connection between mood and the product of emotional states, such as creativity [14] using the popular experience sampling method (ESM) and daily reconstruction method (DRM) techniques. However, it is limited by the relative scarcity of physiological motion data to find relations between different states of emotions.

Fig. 2 summarizes a pipeline with the data extraction, data processing, data analysis, and evaluation stages part of this work.

A. Dataset

Our work utilizes the Daily Ambulatory Psychological and Physiological recording for Emotion Research (DAPPER) dataset [15] that is a collection of self-reported psychological data and physiological recordings through smartphone

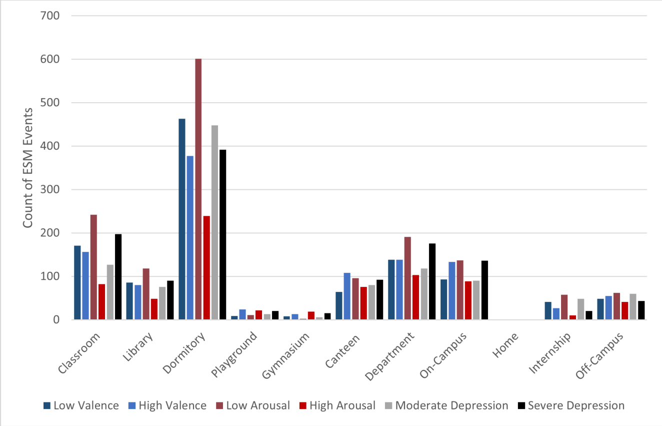


Fig. 3. ESM event count in different environment settings.

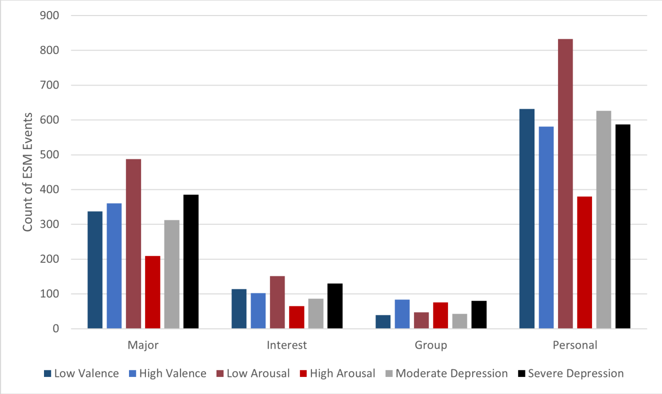


Fig. 4. ESM event count based on activity types.

apps and wearable wristbands to explore daily emotional experiences. It aims to actively encourage the study of emotions based on natural real-life scenarios as opposed to predetermined lab experiments. Laboratory-based methods for studying the psychological and physiological basis of human emotion are not authentic enough to simulate real-life scenarios in day-to-day activities. The conventional practice is to use images, videos, or text in controlled environments, which is not ideal to monitor individuals in a naturalistic setting over an extended period of time.

Data were collected during natural everyday activities from participants over a period of five consecutive days (9 A.M.–11 P.M.).

Based on the distribution of ESM events, as depicted by Fig. 3, the most common settings were dormitories, classrooms, and department buildings. Furthermore, the frequency of events was shown to be mainly personal, as demonstrated in Fig. 4, with the majority being related to low arousal, both moderate and severe depression, and both high and low valence.

It is assumed that participants have answered the questionnaires honestly without any bias, and the data collected were based on answers evoked by everyday activities instead of outlier events

The data were modeled with the following assumptions.

1) For binary segregation of arousal and valences, the Likert-scale reported scores of 1 and 2 are treated as low, whereas 3, 4, and 5 are considered as high. This threshold is in alignment with the works reported in [16] and [17].
2) Depression as per the Beck Depression Inventory-II (BDI-II) scale had scores in the range [20, 28] or [29, 63], corresponding to the categories of moderate and severe, respectively.
3) Signals of 1-Hz frequency are employed for robust and efficient performances in continuous monitoring situations.

Statistical analysis was performed to ascertain the differences between the groups belonging to high and low valence, high and low arousal, and moderate and severe depression with respect to *place* and *activity*. The null hypothesis is rejected if the p-value < 0.05 (i.e., the population does not have statistically significant differences) in the subsequent calculations. Groups belonging to all six classes for *place* are observed to deviate from the Gaussian distribution as per the Shapiro–Wilk test for normality. Thus, the nonparametric Wilcoxon rank sum test was applied, and differences were noted between the populations. Groups belonging to all six classes for *activity* are observed to follow the Gaussian distribution as per the Shapiro–Wilk test for normality. Thus, the parametric independent t-test was applied, and differences were noted between the populations. The alternate hypothesis was accepted in both cases.

The authors of DAPPER proposed to record psychological data through ESM and DRM surveys on the participant's mobile phone and assess it based on different scales, such as the positive and negative affect scale (PANAS), valence, and arousal. In addition to that, record physiological data through a custom-designed wristwatch, and organize it based on three-axis acceleration (ACCEL), galvanic skin response (GSR) signals, and photoplethysmography-derived HR (PPG-HR) signals. As mentioned in [15], the overall mean of physiological data (HR, GSR, and ACCEL) over the period of five days was steady throughout the day. The units of ACCEL, GSR, and PPG-HR are root mean square (rms) in m/s^2, micro-Siemens μS, and beats per minute (BPM), respectively. It was also validated through bivariate correlation matrices between emotion categories that there are similar patterns between the ESM and DRM data, which concurs with previous studies, as seen in [14] and [18]. The correlation between emotion categories and motion signals further adds to the validity of the data (for example, the positive correlation between PPG-HR fluctuations and inspired category).

The main observations of Shui et al. [15] can be summarized as follows.

1) The presentation of the DAPPER dataset that supports emotion research in authentic daily settings. DAPPER consists of the following.
    a) *ESM Data:* Self-reported thoughts, emotional sentiments, and actions over an extended period of time.
    b) *DRM Data:* Collection of how the participants spent their time and the emotions experienced during various activities from their day-to-day lives.

  c) *PPG-HR Data:* Determined by photoplethysmography signals.
  d) *GSR Data:* Determined by electrodermal activity sensors.
  e) *ACCEL Data:* Triaxial accelerometer data that determine changes in speed in relation to the Cartesian coordinate axes.

2) The association between psychological recordings (ESM) and physiological recordings (PPG-HR and GSR) that were collected through day-to-day activities over a long period of time was depicted through bivariate correlation matrices.

The data collection procedure followed to capture data from the participants is given as follows.

1) Pretest: The patients were asked to take a pretest to evaluate their traits before the start of the main experiments. The pretest was conducted using the WJX survey platform and was organized into the following sections: BDI-II, PANAS, and others.

2) Psychological Recordings: The psychological recordings were recorded on smartphones on the Psychorus questionnaire app. ESM and DRM were both utilized, and the questionnaires were sent to the participants' phones as push notifications.

The ESM questionnaire was divided into the following: daily event information, participants' personal assessment, a five-item ten-item personality inventory (TIPI-C) for personality, a ten-item PANAS scale, valence, and arousal. This was sent to participants at random time periods between 9 A.M. and 11 P.M. with a minimum 90-min break, six times per day. They were asked to fill up the questionnaire based on the events/activities in the last 30 min. The DRM questionnaire also had the ten-item PANAS scale, valence, arousal, and an additional section to describe events. This was sent to participants at 11 P.M. every night.

The ten emotional categories chosen for both questionnaires were active, afraid, attentive, determined, nervous, inspired, ashamed, alert, hostile, and upset.

3) Physiological Recordings: The physiological data were recorded by a custom wristband manufactured by Psychorus. ACCEL data were collected at 20-Hz across the x-, y-, and z-axes using an accelerometer sensor embedded in the wristband. Similarly, PPG-HR data were collected with a green light of 532-nm wavelength at 20 Hz. GSR, however, was sampled at a higher frequency of 40 Hz.

To minimize noise artifacts, allow for faster computation, and level data representation, the final signals were downsampled to 1-Hz rms using the 10-s time windows for PPG-HR downsampling and using the mean square of the triaxial raw ACCEL.

Both the psychological and physiological recordings were collected over a period of five days from Monday to Friday during the winter season. The participants' gender split was 64 male and 78 female, and the average age was 21.5 (between 18 and 31). Out of 142 patients, 88 participants recorded physiological data. Out of 88 records, one record was missing (i.e., invalid). The ethical approval process was Helsinki Standard, approved by the Local Ethics Committee of the Department of Psychology, Tsinghua University, and written consent was given by all the patients.

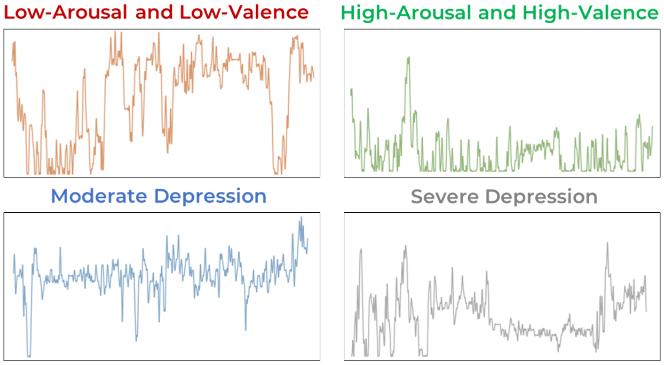


Fig. 5. Sample instances of PPG-HR representing the binary classes (x-axis: BPM against y-axis: time in seconds).

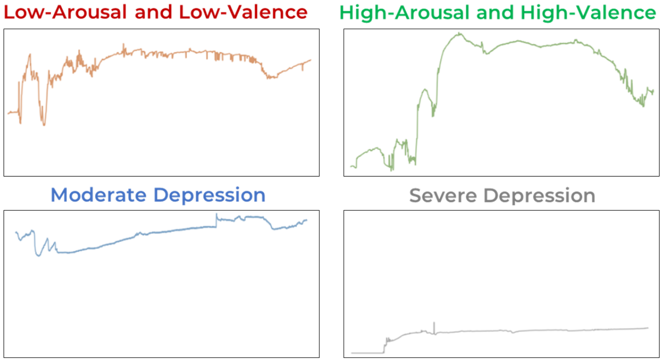


Fig. 6. Sample instances of GSR representing the binary classes (x-axis: conductance in μS against y-axis: time in seconds).

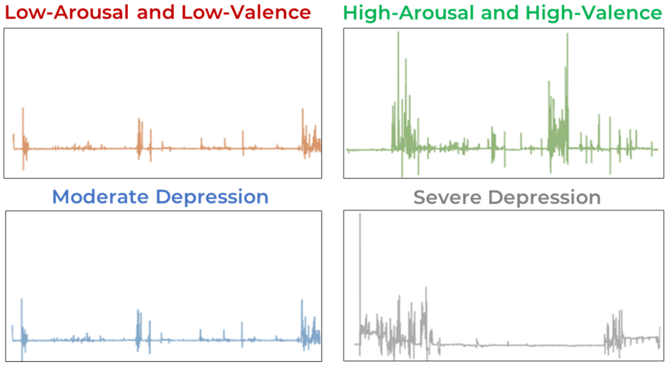


Fig. 7. Sample instances of ACCEL representing the binary classes (x-axis: rms in m/s^2 against y-axis: time in seconds).

Initially, the dataset consisted of 2249 signals, with an average duration of 1713.4 s. Since the ESM sampling method requires strictly considering the 30 min preceding the questionnaire, signals not meeting these criteria were discarded. This included signals with missing segments and shorter duration. The final number was 2034 signals with an average duration of 1800.00 s (30 min). To mitigate the amplitude scaling issue commonly prevalent in wearable devices' acquired signals, the Z-score normalization is performed on the signals, using the equation described in [19]. Sample instances of PPG-HR, GSR, and ACCEL data that represent each of the binary classes are illustrated in Figs. 5–7.

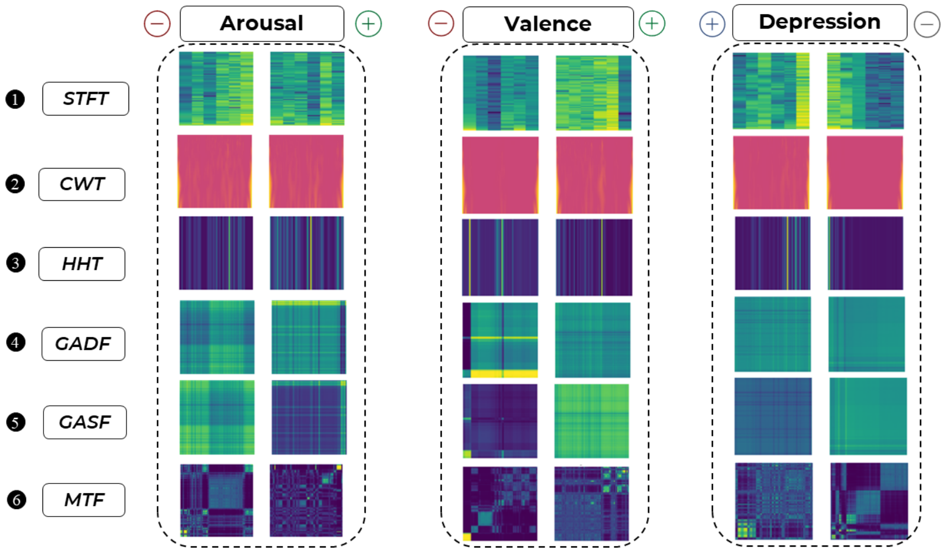


Fig. 8. Signal-to-image transformations for arousal (negative, positive), valence (negative, positive), and depression (moderate, severe) binary classes.

TABLE I
ESM Event Count

	Class 0	Class 1
Valence	1003	1031
Arousal	1361	673
Depression	960	1074

Postpreprocessing of the raw data, the distribution of ESM events based on valence, arousal, and depression is presented in Table I.

B. Preprocessing

Other than dropping erroneous records, the data had to be prepared for our inference framework. In this section, we describe the techniques used to extract the underlying signal features necessary for inference, and the output representations are shown in Fig. 8.

We rationalize the conversion of time-series signals into image-like representations of the time and frequency domains based on the following assertions. First, neural network architectures developed for computer vision achieved great success in classification performance by virtue of exploiting translational invariance through automated feature maps' extraction by receptive fields and learning with weight sharing [20]. By considering signals in such a way, we seek to leverage the abilities of CNNs. Second, a single derived image representation of fixed dimensions can implicitly capture spatial and temporal characteristics for the duration of 30 min (duration of one event preceding the ESM valence/arousal questionnaire). This adds elements of intensity, mean power, and dynamic ranges at different time points within a single window, which can be advantageous for improved classification discernability. Finally, the application of similar methods for capturing time and frequency relationships for human activity recognition from sensor data in [21], [22], and [23] encourages the exploration of these in our work.

1) Short-Time Fourier Transform (STFT): STFT is a method for performing time–frequency analysis of a time-domain signal by dividing signals into shorter windows and computing Fourier Transform on each window (256 segments in this work, with an overlap of 128) separately then subsequently aggregated [24]. We represent STFT numerically, with $x(t)$ as the original signal, while $w(t-T)$ as a t-centered window tapering function [25]

$$X(\tau,\omega)=\int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-i\omega t}dt. \tag{1}$$

2) Continuous Wavelet Transform (CWT): CWT, like STFT, allows for time–frequency analysis of time-domain signals through the shifting of predefined analytic wavelets across the time axis of the original signal. Alongside the varying of the shifting parameter, the scale of the wavelets is varied for optimal capturing of frequency characteristics of a signal. The *morl* wavelet is used in this work. CWT is expressed mathematically using the following equation [26]:

$$X_w(a,b)=\frac{1}{|a|^{1/2}}\int_{-\infty}^{\infty} x(t)\bar{\psi}\left(\frac{t-b}{a}\right)dt. \tag{2}$$

3) Hilbert–Huang Transform (HHT): HHT performs power distribution analysis on a given signal through its decomposition, within the time domain, to its intrinsic mode functions (IMFs), using empirical mode decomposition (EMD). Instantaneous frequency, defined in the following equation, is obtained through the application of Hilbert spectral analysis (HSA) on each IMF, ultimately preserving signal spectral characteristics [27]:

$$\omega(t)=\frac{d\varphi(t)}{dt}. \tag{3}$$

4) Gramian Angular Summation/Difference Fields (GASF/GADF): Unlike the mathematical tools discussed above, GASF/GADF is a framework to encode time-series data as images for better time-series feature extraction in CNNs. The first step is to normalize the time-series data

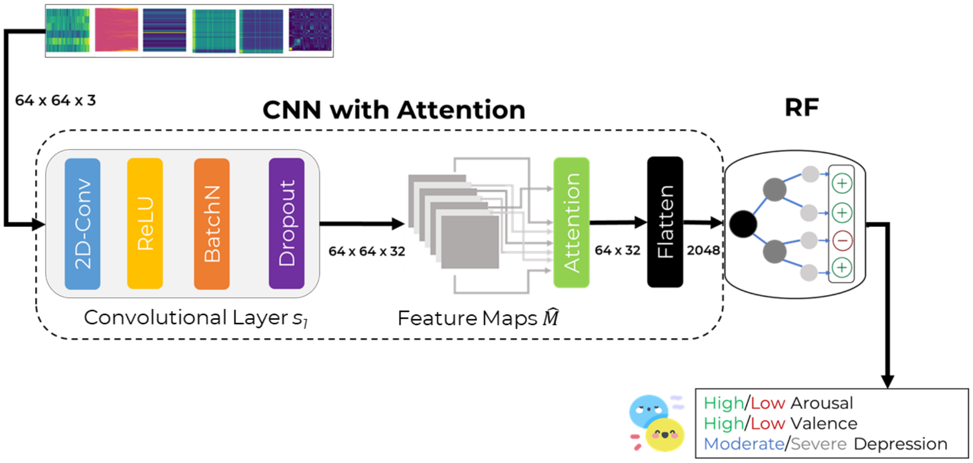


Fig. 9. Architectural pipeline of the proposed model (unimodal).

then to represent the data in polar form to preserve temporal relations by using the angular cosine on the normalized data to get the angle, while the radius is the timestamp of the particular datapoint (divided by a constant factor N) [20]

$$\tilde{x}_0^i = \frac{x_i \min(X)}{\max(X) - \min(X)} \tag{4}$$

$$\phi = \arccos(\tilde{x}_i), \ 0 \le \tilde{x}_i \le 1, \ \tilde{x}_i \in \tilde{X}, \ r = \frac{t_i}{N}, \ t_i \in \mathbb{N}. \tag{5}$$

Finally, GASF is defined by the trigonometric summing of all points, while GADF is defined by their difference, as shown in the following equations:

$$\begin{aligned} \text{GASF} &= [\cos(\phi_i + \phi_j)] \\ &= \tilde{X}' \cdot \tilde{X} - \sqrt{I - \tilde{X}^2} \cdot \sqrt{I - \tilde{X}^2} \\ \text{GADF} &= [\sin(\phi_i - \phi_j)] \\ &= \sqrt{I - \tilde{X}^2} \cdot \tilde{X} - \tilde{X}' \cdot \sqrt{I - \tilde{X}^2}. \end{aligned} \tag{6}$$

5) Markov Transition Fields (MTFs): Another time-series visualization algorithm that is prominently used as a preprocessing step for computer vision deep neural network applications is the MTF. MTF relies on splitting the data longitudinally into quantile bins to create a Markov transition matrix (MTM) denoting the probability of a transition from one bin to another. Subsequently, MTF is defined by adding a temporal dependence to MTM, according to the following equation:

$$M = \begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ w_{ij|x_2 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_2 \in q_i, x_n \in q_j} \\ \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix} \tag{7}$$

where $W_{i,j}$ corresponds to a single MTM value [20].

C. Ensemble Deep Learning

In this work, a deep convolutional network with an attention mechanism [28] is proposed for the binary classification of high/low valence, high/low arousal, and moderate/severe depression states. From an overview perspective, the deep learning architecture consists of five layers, including the attention layer, as depicted in Fig. 9. During each set of experiments, the raw signal-to-image transformations processed by STFT, CWT, HHT, GASF, GADF, or MTF are fed as input into the model. The dimensions of the images are 64 × 64 × 3 and span the 30-min duration denoting the time period of relevance prior to the ESM sampling questionnaire.

Suppose that the input images are represented by $X = \{x_{1,1}, \ldots, x_{h,b}\}$, where $x(i, j) \in \mathbb{R}^{hxb}$, and h and b are height and width of the image. Consider then the input of the 2-D convolutional layer s_1 as a single image $x(i, j)$, which, when convolved with the kernel $w(i, j)$ of size $a \times b$, obtains the feature map $m(i, j)$ using

$$\begin{aligned} m(i, j) &= x(i, j)^* w(i, j) \\ &= \sum_{u=-at=-b}^{a} \sum_{v}^{b} x(u, v) \cdot w(i - u, j - v). \end{aligned} \tag{8}$$

The 2-D convolution is performed on the input images with a stride of 1 and a kernel size of 3 × 3 with 32 filters. Batch normalization was integrated after this layer for the normalization of the activation function, rectified linear unit (ReLU). To reduce overfitting, both the $L1$ kernel regularizer and a single dropout layer with a value of 0.3 were added. Let the set of feature maps across after the batch normalization and dropout layers be denoted as $\hat{M}^{s_1} = \{\hat{m_k}^{s_1}, \ldots, \hat{m_n}^{s_1}\}$, where k indicates a spatial location of n total locations within the layer. We forego a second convolutional layer with pooling layers and fully connected layers in favor of the attention mechanism and the RF, respectively, in the complete model.

The learned feature maps are then propagated to the attention layer for additional emphasis on the key aspects of the representations. The attention layer uses the compatibility scores of the dot product of $\hat{m_k}^{s_1}$ and the global feature vector $\boldsymbol{g}$, which has the entire input image as support as defined as follows:

$$c_i^s = \langle \ell_i^s, \boldsymbol{g} \rangle, \quad i \in \{1, \ldots, n\}. \tag{9}$$

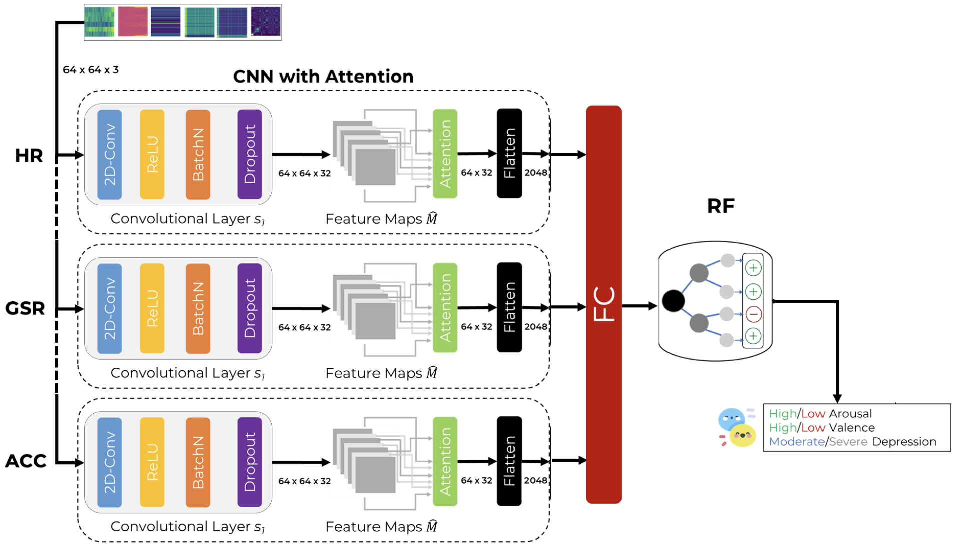


Fig. 10. Architectural pipeline of the proposed model (multimodal).

The relative magnitude of the scores is contingent on the strength of activation of $\hat{m_k}^{s_1}$ and the alignment between this feature map and $\boldsymbol{g}$. The scores are normalized to the range (0, 1) by the softmax operation in

$$a_i^s = \frac{\exp\left(c_i^s\right)}{\sum_j^n \exp\left(c_j^s\right)}, \quad i \in \{1, \ldots, n\}. \tag{10}$$

The final output of the attention mechanism for the convolutional layer s_1 with feature map $\hat{m_k}^{s_1}$ is denoted by $\boldsymbol{g}_a^s$, which is the weighted combination of the feature maps for this layers and the weights are specified by a

$$\boldsymbol{g}_a^s = \sum_{i=1}^{n} a_i^s \cdot \hat{m_k}^{s_1}. \tag{11}$$

It is hypothesized based on [29] and [30] that the convolutional layer learns a hidden representation of the general regions of interest across the images, and the attention mechanism weighted features derived from this representation indicate the relevant intermittent or latent activation accompanying mood states during motion or the exhibition of certain physiological markers.

The weights of the model are trained using a fully connected layer comprising 24 neurons with a sigmoid activation function. After training for 25 epochs, with early stopping and learning rate reduction on the plateau of 3 epochs, only the feature extraction layers are maintained as input for RF.

For the multimodal cases, as shown in Fig. 10, the convolutional channels are replicated three times for each modality, and the features from each pipeline are fed to a fully connected layer of 16 neurons before being passed through a sigmoid activation for final classification during training of the feature extraction component.

For both unimodal and multimodal cases, random search and grid search were utilized to select the ideal number of neurons, depth of the networks, and activation functions, with F1-score being the optimization criterion. It is worth mentioning that kernel sizes varied between 5 × 5, 20, and 20, while the numbers of filters varied between 1 and 128.

To identify the machine learning classifier most suitable for classifying the feature representations only, we compare RF with support vector machines (SVMs) and K-nearest neighbors (KNNs) [31]. To see if changes to the feature extraction components drastically affect model performance, we adapt the CNN layers with spectral normalization (SN) [32] for stability of the weights and also extend the module with long-short term memory (LSTM) cells and bidirectional LSTM (BiLSTM) as well [33]. These are independently incorporated prior to the attention mechanism, and the goal is to observe if the models are consistently learning similar latent space representations in the presence of architectural variations.

III. Results

The training and test instances are selected using hold-out cross-validation, where 75%, 15%, and 10% of motion signals are used for training, testing, and validation. When adapting the weights of the CNN without the RF component so that it operates as a feature extractor, 50% of the original training set was used for training and the remaining 25% for validation. The standard evaluation metrics considered are accuracy, sensitivity, specificity, and F1-score.

While inferring users depressive states using each of HR (PPG-HR), GSR, and accelerometer signals yielded highest performance results when CWT, STFT, and HHT signal-to-image preprocessing transforms were employed, relatively, with an average accuracy and F1-score values of 51.88 ± 0.57 and 51.78 ± 0.39, as seen in Table II, resulting models exhibit moderate, yet balanced, performance with minimal overfitting yielding average performance values of 50.33 ± 1.15 accuracy, 56.92 ± 1.76 sensitivity, 43.13 ± 1.78 specificity, and 50.02 ± 1.21 F1-score.

On the other hand, valence and arousal respective inference unimodal models tended to overfit as seen in the highly contrasted (and constantly fluctuating) specificity and sensitivity metrics in Tables III and IV, with average arousal

TABLE II
DEPRESSION - CNN-ATTENTION-RF MODEL PERFORMANCE COMPARISONS EXPRESSED AS PERCENTAGE

Method	Accuracy	Sensitivity	Specificity	F1-Score
PPG-HR				
STFT	50.837	56.338	44.655	50.497
CWT	51.967	59.26	44.284	51.772
HHT	49.95	54.532	44.988	49.76
GASF	51.031	58.706	42.704	50.705
GADF	49.067	56.115	41.308	48.712
MTF	49.115	56.573	40.838	48.701
GSR				
STFT	51.279	52.947	49.837	51.392
CWT	50.934	54.453	47.136	50.795
HHT	49.213	54.087	43.906	48.996
GADF	50.393	54.093	46.638	50.365
GASF	50.738	52.996	48.376	50.686
MTF	50.294	57.298	42.613	49.956
ACCEL				
STFT	50.146	53.627	46.261	49.944
CWT	49.755	56.01	42.845	49.427
HHT	52.407	58.745	45.592	52.168
GASF	49.607	54.817	43.849	49.333
GADF	48.969	55.914	41.448	48.681
MTF	48.279	53.517	41.961	48.239

TABLE III
AROUSAL - CNN-ATTENTION-RF MODEL PERFORMANCE COMPARISONS EXPRESSED AS PERCENTAGE

Method	Accuracy	Sensitivity	Specificity	F1-Score
PPG-HR				
STFT	56.048	82.108	14.442	48.275
CWT	58.507	86.640	14.161	50.281
HHT	57.718	86.504	11.758	49.131
GASF	57.768	85.950	12.670	49.310
GADF	59.488	89.295	11.761	50.528
MTF	59.784	88.848	13.500	51.174
GSR				
STFT	57.031	81.380	18.004	49.692
CWT	60.028	89.304	13.165	51.234
HHT	59.538	88.747	12.722	50.735
GADF	58.753	86.151	15.044	50.598
GASF	59.292	87.261	14.656	50.959
MTF	58.996	88.990	10.974	49.982
ACCEL				
STFT	65.144	08.196	93.381	50.789
CWT	64.897	05.100	94.489	49.795
HHT	65.339	06.091	94.658	50.374
GADF	64.995	02.642	95.810	49.226
GASF	64.701	04.340	94.568	49.454
MTF	62.130	06.842	93.105	49.973

TABLE IV
VALENCE - CNN-ATTENTION-RF MODEL PERFORMANCE COMPARISONS EXPRESSED AS PERCENTAGE

Method	Accuracy	Sensitivity	Specificity	F1-Score
PPG-HR				
STFT	81.735	98.006	03.726	50.866
CWT	81.415	99.703	00.510	50.107
HHT	81.267	99.578	00.294	49.936
GASF	81.218	99.578	00.000	40.789
GADF	81.464	99.881	00.000	49.941
MTF	81.565	99.703	01.372	50.538
GSR				
STFT	80.286	98.244	00.859	49.552
CWT	81.170	99.209	01.443	50.326
HHT	81.515	99.818	00.584	50.201
GADF	81.368	99.704	00.303	50.004
GASF	81.563	99.882	00.494	50.188
MTF	81.416	99.638	00.850	50.244
ACCEL				
STFT	50.491	49.428	51.703	50.566
CWT	47.887	46.213	49.844	48.029
HHT	46.707	45.617	47.959	46.788
GADF	47.592	46.38	49.478	47.929
GASF	49.410	48.242	50.787	49.515
MTF	49.705	47.641	51.849	49.745

model specificity and sensitivity values of 59.69 ± 39.47 and 40.49 ± 39.21, respectively. Similarly, for valence inference models, the average specificity and sensitivity metrics are 82.3 ± 25.32 and 17.32 ± 23.99.

The effects of multimodality on the detection of depression, arousal, and valence were tabulated and highlighted in Tables V–VII. The average performance metrics for multimodal inference of depressive states are 49.51 ± 1.46 accuracy, 65.0 ± 10.0 sensitivity, 35.0 ± 10.0 specificity, and 50.0 ± 0.0 for F1-score. Analogously, arousal inferring models exhibited 55.86 ± 3.87 accuracy, 71.59 ± 18.98 sensitivity, 30.75 ± 20.69 specificity, and 51.17 ± 1.31 for F1-score. Finally, valence prediction models followed similar trade when incrementing input data modality, as average recorded performance values were 74.58 ± 5.85 accuracy, 87.21 ± 9.08 sensitivity, 15.06 ± 11.09 specificity, and an F1-score 51.13 ± 1.35.

We report additional experiments to support one of our theories that leveraging feature representations from CNN components and classifying with a simpler model can reduce overfitting for certain types of data as in this work. With additions of SN, LSTM, BiLSTM, or replacements of SVM and KNN instead or RF, we notice that the performance of the models tends to become more balanced. However, the differences between each of these variants are negligible and show that the data that we used, i.e., the 30-min window containing emotional stimuli and preceding an ESM reporting event, were not particularly effective for the tasks.

After intent analysis highlighted results, we cannot arrive at a conclusion in favor of multimodal networks. Out of the aforementioned permutations, only arousal-oriented models retained reliability within multimodal settings. Despite that, models during experimentations consistently degraded in reliability with each added modality. Even arousal-oriented prediction models quickly overfit with their specificity plummeting to 0 and sensitivity to 100, meaning that the model was prone to an increasing amount of false positives during our trial runs.

IV. DISCUSSION

A. Contributions

In this work, we utilize wearable data acquired during the ESM method, with the intention of capturing a "snapshot" of short events during the day associated with likely external stimuli and aggregating the derived measures to capture more complex interactions between the mind and body. The participants of the data collection procedure are all students, and this establishment of physiological or behavioral baselines for this group of patients essentially paves the way for recognizing signs and symptoms of underlying emotional and mental states via significant deviations from mined patterns. Some studies show valence and its consistent relation to

TABLE V

DEPRESSION - MULTIMODAL MODEL PERFORMANCE COMPARISONS EXPRESSED AS PERCENTAGE

Model	Method	Accuracy	Sensitivity	Specificity	F1
PPG-HR & GSR					
CNN	CWT-STFT	50.439	80.000	20.000	50.000
PPG-HR & ACCEL					
CNN	CWT-HHT	48.576	60.000	40.000	50.000
GSR & ACCEL					
CNN	STFT-HHT	47.984	60.000	40.000	50.000
PPG-HR & GSR & ACCEL					
CNN	CWT-STFT-HHT	51.036	60.000	40.000	50.000
CNN+Attn-RF	CWT-STFT-HHT	50.296	48.347	52.480	50.413
CNN+SN+Attn-RF	CWT-STFT-HHT	49.167	48.232	50.399	49.316
CNN+LSTM+Attn-RF	CWT-STFT-HHT	48.723	49.061	48.328	48.694
CNN+BiLSTM+Attn-RF	CWT-STFT-HHT	48.179	47.637	48.952	48.295
CNN+Attn-SVM	CWT-STFT-HHT	49.851	30.970	71.894	51.432
CNN+Attn-KNN	CWT-STFT-HHT	49.067	29.802	71.009	50.406

TABLE VI

AROUSAL - MULTIMODAL MODEL PERFORMANCE COMPARISONS EXPRESSED AS PERCENTAGE

Model	Method	Accuracy	Sensitivity	Specificity	F1
PPG-HR & GSR					
CNN	MTF-CWT	53.245	63.478	37.094	50.286
PPG-HR & ACCEL					
CNN	MTF-STFT	53.589	60.543	42.508	51.526
GSR & ACCEL					
CNN	CWT-STFT	55.063	62.338	43.410	52.874
PPG-HR & GSR & ACCEL					
CNN	MTF-CWT-STFT	61.553	100.000	00.000	50.000
CNN+Attn-RF	MTF-CWT-STFT	54.966	72.292	27.804	50.048
CNN+SN+Attn-RF	MTF-CWT-STFT	54.181	73.262	23.912	48.589
CNN+LSTM+Attn-RF	MTF-CWT-STFT	55.262	73.498	26.260	49.879
CNN+BiLSTM+Attn-RF	MTF-CWT-STFT	54.721	67.453	33.903	50.678
CNN+Attn-SVM	MTF-CWT-STFT	61.553	100.000	0.00	50.000
CNN+Attn-KNN	MTF-CWT-STFT	55.504	68.912	34.102	51.507

TABLE VII

VALENCE - MULTIMODAL MODEL PERFORMANCE COMPARISONS EXPRESSED AS PERCENTAGE

Model	Method	Accuracy	Sensitivity	Specificity	F1
PPG-HR & GSR					
CNN	STFT-CWT	74.812	87.360	14.159	50.759
PPG-HR & ACCEL					
CNN	STFT-STFT	70.904	80.484	25.711	53.097
GSR & ACCEL					
CNN	CWT-STFT	69.863	80.997	20.365	50.681
PPG-HR & GSR & ACCEL					
CNN	STFT-CWT-STFT	82.747	100.00	00.000	50.000
CNN+Attn-RF	STFT-CWT-STFT	81.103	97.547	2.082	49.814
CNN+SN+Attn-RF	STFT-CWT-STFT	80.733	96.373	5.611	50.992
CNN+LSTM+Attn-RF	STFT-CWT-STFT	81.356	98.055	1.515	49.785
CNN+BiLSTM+Attn-RF	STFT-CWT-STFT	79.976	96.508	5.711	48.54
CNN+Attn-SVM	STFT-CWT-STFT	82.745	100.00	0.000	50.000
CNN+Attn-KNN	STFT-CWT-STFT	78.213	93.780	3.730	48.755

PPG-HR, particularly among men, in response to pleasant, neutral, and unpleasant stimuli [34]. Skin-conductance-related features show both positive and negative associations with arousals [34], [35]. However, the relationship between autonomic functioning described by these sensors and affective emotions is not as linear and can change drastically across the type of stimuli and setting of the measurements [36], and studies such as ours contribute to ongoing research. ACCEL is relatively more general in its application, as it can capture anxious gesticulation, nervous fidgeting, and frequency of activities such as sitting/walking/running.

It appears that empirically STFT and CWT preserve temporal dependencies to a greater extent than their counterparts. High sensitivity with much lower specificity or vice versa suggests probably overfitting for the positive and negative classes, respectively. It is worth mentioning that the initial training approaches involved a standard 2-D-CNN based on the LeNet architecture the application of transfer learning with the VGG16, ResNet50, and InceptionV3 models [37]. These models produced results such as a sensitivity of 100% and a specificity of 0%, showing signs of extreme overfitting alongside unstable and diverging training and validation losses.

By employing a simpler structure with the proposed CNN-RF with attention, a considerable increase in the discriminability between the classes is observed. The rationale behind the modification of the architecture is rooted in three primary reasons. First, the attention mechanism serves to prioritize the most relevant regions of interest within the feature maps extracted from the input images by the convolutional filters. This likely reduces variance in the presence of potential sampling bias. Second, 30-min long physiological signals can prove to be computationally expensive to process, which is lessened by the drastic reduction in the number of pertinent features exiting the attention mechanism. Finally, ensemble techniques incorporating "bagging" and "boosting" have the ability to fare reasonably when the size of the data is comparatively small.

According to Leeflang et al. [38], based on the perceived domain importance of either class, screening with high sensitivity is possible, at the risk of false positives taking into account that a reasonable threshold of specificity is maintained (40̃%). This is because the models tend to detect higher valence/arousal states considerably better than lower valence/arousal states in general but often struggle to differentiate between instances of high and low. We purport that this is because of interindividual differences stemming from their reactionary behaviors within the context of changing places and activities. This is in line with [39], as it suggests that no conclusive statistical relationship between valence and arousal was found, especially when analyzing intra-individualistic findings. This insinuates that nontransferability of information could be the underlying cause impeding the learnability of such relationships, as observed in our findings. Conversely, the depression states had roughly the same level of differentiability, indicating that not all individuals exhibit similar or consistent patterns in the motion dimension. Although some studies show that accelerometer-based movement for psychologically distressed individuals is lower in intensity [40], others [41] indicate that patients suffering from manic depressive disorders can have competitively more movement activity. Moreover, concordance with [42] is noted as well, where individuals with depression have only minor and mostly stable fluctuations of valence compared to a healthy control group.

To minimize the role of confounding variables within the context of people, events, or even the relationship between ephemeral mood (valence/arousal) and chronic depression, statistical importance between *place* and *activity* was computed. With significance values below $p = 0.05$, it can be surmised that place and activity could have in fact had a bearing on the level of subconscious response activation in a substantially meaningful way. The reliability of the self-ratings can be influenced by simple subjectivity, or events preceding the ESM questionnaire time, carrying over feelings of nonchalance, agitation, or impatience while reporting. A core limitation stems from the fact that the data appears to be quite similar in nature, suggesting that only a causal relationship might exist and might after all be inferior to more biological contemporary sensor readings acquired by HRs (photoplethysmography sensors) and skin conductance (GSR sensors) [43]. This dataset could be an instantiation of a scenario reflecting the overall proliferation of depression among the population, therefore introducing numerosity in records, and diversity in terms of multiple depressive states can lead to conclusive, less elusive outcomes. Following suggestions stated in [44], we consider this developed model and associated results as a general approach, which can be improved with the collection of similar samples.

Addressing the intersubject variability again, it is noticeably high for this cohort of patients, despite the similarities in demographics and location. It can be surmised then that a relatively short, multiple monitoring duration (30 min) across a week, even in the presence of potential external stimuli, is insufficient to draw meaningful conclusions about individuals' baseline mental health state. These individual differences likely arise from a complex interplay between biological/genetic, social, and/or environmental factors [45]. To render models able to attain precision psychiatry, longitudinal continuous personal time-series models (as in this work), but of a significantly longer duration (months instead of weeks), will prove to be more promising.

Considering the findings of the recent work reported in [46], it appears that a longer duration of data collection and more dimensions to the data, in fact, improve the performance of wearable-derived features when used with machine learning. Their observational study utilized three months of continuous wearable data and medical examinations across this period to detect the onset of mental illness (measured by confirmed diagnoses due to "administration of hypnotics, anxiolytics, or antidepressants" and/or "psychiatric visits"). The score of 71.2 as quantified by the area under the receiver operating characteristic (AUROC) curve was achieved and featured importance explored in terms of monthly averages of HR, physical activity duration, and sleep rhythms revealed sleep habit disturbances to be the primary contributor for predictive modeling.

An overarching reason for the elusiveness of depression or mental health detection is the associated stigma and the reluctance of individuals to openly report symptoms. As such, the threshold for actual psychiatric consultation or diagnoses is relatively higher than diseases such as diabetes or cardiovascular disease (objective biomarkers being present). Interestingly, this mirrors the relative likelihood of noticeable physical symptoms manifestation in the individuals, in the form of decreased interest in the activity, disrupted sleeping behavior, or eating disorders. This phenomenon is observed by Saito et al. [46] and Ahmed et al. [47] as well, where only severe cases exhibit discernible physiological patterns, which can be captured by wearables.

Psychiatry and psychology fields deal with the problem of discrepancy between self-reported and objective assessments. Moukaddam et al. [48] suggest a holistic approach where researchers must assess the information provided by sensors and what it reflects objectively along with its link with the patient's self-reports. Oftentimes, the labels, especially self-reported ones, might not reflect the disease it is testing for, but rather a related, comorbid/co-occurring condition with similar pathology. As purported by Harvey et al. [49], anhedonia, or the reduced ability to feel pleasure and enjoy activities, is a core symptom of depression also present in other mental health disorders as it impairs reward processing and can be gleaned from mobility patterns via smartphone collected data. This is echoed by Moukaddam et al. [48] who state that the numerous daily fluctuations of emotions in individuals, positive and negative, along with their intensity and relation to the adequacy of the response or fidelity of continued function, can be expressed through wearables. These digital markers of behavioral change during specific contexts can consolidate objective data of value during any future clinical visits, as discordance between measured reality and patient memory of events can happen due to recall being subject to bias. This is the documented case of increased attention bias of depressive individuals to negative emotional stimuli stemming from the dysfunction of excitation and inhibition [50].

Furthermore, Teismann et al. [51] assert that states of depression and anxiety are characterized by dysfunctions of affective experience and affective quality perception (i.e., the inability to accurately describe what exactly they are feeling). This has the effect of modulating neutral responses on self-reported scales, rather than eliciting valence or arousal responses at either extremity (higher or lower). Naturally, while we can expect the chronic experiences of negative emotions to be indicative of declining emotional health [50], neutral scores warrant more information to discriminate if there has been a model error or subject-specific variation in the presence of other mental health conditions.

Zhang et al. [50] also note that higher arousal levels were observed in a depressed population, and it appeared that rising emotional intensity was harder to suppress. This could be a reason for the increased arousal evaluation metrics in Table VII. Similarly, Teismann et al. [51] bring to attention that higher anxiety scores are noticed for higher arousal, and anecdotally had a stronger association than with depression. Motivated by the continued use of HR to detect depression, stress, and other psychiatric disorders, we expected PPG-HR to have higher predictive power than empirically found in this study [52].

Our study highlights the opportunities and limitations of multimodality owing to its conferring of confounding in specific point-of-care applications. The blending of multiple sensor modalities results in each person becoming their own reference, as each individual or group of individuals in the same demographic has similar routine behavior (meal times, sleep patterns, exercise, work schedules, places visited, and so on) [47]. While this can prove to be beneficial, in cases where the information quality is not as rich or diverse as expected, models can overfit one's own wearable phenotypes or symptomatology. Indeed, this appears to be the case in Table VI, where fusing two modalities boosted the detection of arousal compared to one modality, but three modalities made it worse again. As such, model fairness across different subgroups should be maintained while tailoring personalized models to each stratum of the population in future studies.

Although multimodality is traditionally expected to improve data diversity and enhance extracted insights, challenges affect multimodal networks to a greater degree than their unimodal counterparts. Multimodal models are more susceptible to many anomalies observed in data features and recording techniques. Illustrating this, Lahat et al. [4] enumerate potential sources of multimodal inference errors in relation to the data acquisition as capture resolution, data span incongruence, and alignment issues. Furthermore, data-embedded features could stymie multimodal efforts, such as data noise and origin variance. Many of the stated challenges affect both unimodal and multimodal approaches; however, their impact on the presence of multimodality is particularly amplified, and their potential solutions' effect diminished [4]. In other words, the more the sensor sources are collected, the more varied sources of errors will be. Especially with low control over the data collection procedure, accounting and addressing all sources of error, given they are discovered, become increasingly difficult with each additional modality.

As always, there is the question of anonymity and accuracy pertaining to the mechanisms to protect patients' data rights. Generally, adding context (geography/visited locations/age/gender) to wearable devices to ascertain more fine-grained user behavior can improve the specificity of models in detecting diseases. However, it is imperative for clinicians and individuals to mutually reach a consensus on the upper and lower thresholds for automated screening threshold and subsequent medical follow-ups. It is likely that additional dimensions, such as age and gender, can contextualize individual behavior to a better degree. This is observed in the results of [51], where age and gender proved to be good indicators of valence and arousal, even without any wearable data.

With the global incidence of mental health disorders rapidly surmounting, it is of utmost importance to leverage ubiquitous wearables for quantifying fluctuations in emotional intensity and responses, reduced sociability, physical activity, prosody, and the overall cognitive function to promote holistic healthy lifestyles [53]. We address the common issue that the majority of predictive models are trained on data obtained exclusively in a Western cultural context, by establishing benchmarks of valence, arousal, and depression modeling trained on a population cohort based in China. One of the advantages of wearable sensors is that the monitoring can be conducted at any place, any time, and increases the regularity of user-specific screening in a passive way [54].

B. Limitations

To glean a better understanding of the statistical properties of the DAPPER dataset, we use statistical methods, t-test, and analysis of variance (ANOVA), to inspect data homogeneity within and across categories. The t-test method measures the significant difference between two entries or signals through the comparison of the means. On the other hand, ANOVA compares the means to calculate significant differences between groups of signals [55].

The t-test is applied by contrasting random samples of PPG, GSR, and ACCEL signals across our binary categories, valence, arousal, and depression. Being conversely capable of analyzing groups of data signals, ANOVA is run to measure significant differences across and simultaneously within categories with the same group (i.e., low valence and high valence, low arousal and high arousal, and moderately depressed and severely depressed).

TABLE VIII
ANOVA Results Across WESAD and DAPPER

Classes	T-Statistic	P-value
WESAD		
Class 0	0.242	0.999
Class 1	0.385	0.999
Class PAN	0.283	0.999
DAPPER - Depression		
Class 0	691.071	0.000
Class 1	787.614	0.000
Class PAN	741.913	0.000
DAPPER - Arousal		
Class 0	704.105	0.000
Class 1	761.943	0.000
Class PAN	741.913	0.000
DAPPER - Valence		
Class 0	704.105	0.000
Class 1	761.943	0.000
Class PAN	741.910	0.000

TABLE IX
Paired *t*-Test Results Across WESAD and DAPPER

Modality	T-Statistic	P-value
WESAD		
ECG	0.751	0.453
ACCEL	-0.169	0.865
EMG	-0.791	0.429
DAPPER - Depression		
PPG-HR	-40.437	4.33e-297
GSR	-706.522	0.000
ACCEL	-13.469	2.19e-40
DAPPER - Arousal		
PPG-HR	-4.846	1.3106e-06
GSR	140.197	0.000
ACCEL	2.898	0.003
DAPPER - Valence		
PPG-HR	25.360	1.075e-130
GSR	12.033	9.951e-33
ACCEL	-2.989	0.002

Both statistical tests' results can be summarized with two values: F-statistic and p-value. F-statistic is the ratio of the variances of two populations, while p-value signifies the strength of evidence against the null hypothesis, being in this case that the variance of two populations under examination is equal. More concretely, higher p-values mark cases with low evidence to reject the null hypothesis, suggesting that observed results are more likely to be due to chance. In contrast, lower p-value scores suggest strong evidence for the statistical uniqueness of the results. We expect signals belonging to the same categorical groups to exhibit less statistical difference than that of signals belonging to opposing groups. Establishing a baseline, we import WESAD, a rigorously used and validated time-series physiological–psychological biomarker-oriented dataset [56]. WESAD contains similar biosignals to DAPPER, proving it easy to compare results from one dataset to another. It contains electrocardiogram (ECG), electromyography (EMG), and ACCEL signals recorded using a respiBAN chest strap, which is subsequently synchronized with user state-trait anxiety inventory (STAI) and positive and negative affect schedule (PANAS) questionnaires. By combining the subjective survey results, sections of recorded signals are labeled as either transient (0), baseline (1), stress (2), amusement (3), or meditation (4).

In an attempt to make WESAD similar in comparison in format to DAPPER, signals are split into smaller chunks (length = 214 583 sampling points) after empirical examination of median label length, as observed in Fig. 11. Subsequently, the aforementioned four labels had to be made into two to reasonably conform with our current datasets' formatting. Consequently, we run both t-test and ANOVA on WESAD's ECG data, given it most conformed to DAPPER's formatting. The results from the prior tests are tabulated in Tables VIII and IX, respectively. In contrast with DAPPERs, WESAD's results are more in line with expectations: data belonging to the same class are more uniform than data from opposite classes. Note that WESAD does likely elicit higher information gain due to its 700-Hz recording frequency versus the preprocessed 1-Hz signals provided through DAPPER. Since one of the goals of our work was to study the utility of low-frequency and low-cost sensors, this does not pose a problem in this data exploration aspect.

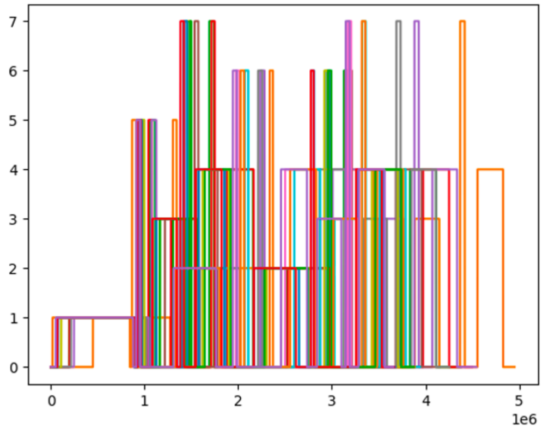
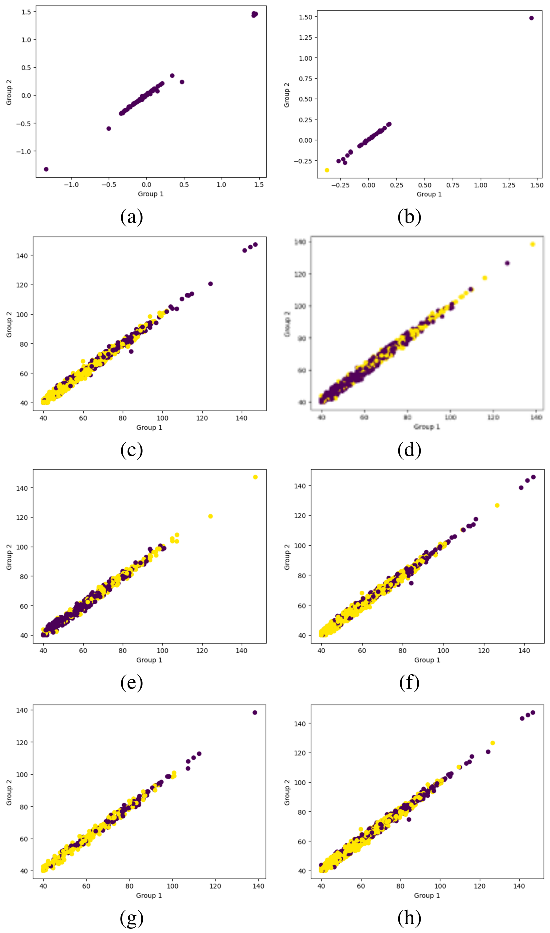


Fig. 11. Plotting label transition frequency of all user data. The horizontal axis denotes the number of sampling points in a signal, and the vertical axis denotes the label. Note that labels 5–7 were equivalent to 0, as described by Schmidt et al. [56].

To further our understanding of DAPPER's statistical attributes and highlight any irregularities that might arise due to labeling errors, we perform clustering both of our and WESAD data. The results of the KNN clustering algorithm are shown in Fig. 12. One can observe that the KNN clustered points further corroborate the same results as shown by the t-test and ANOVA tests: data from the same category are far more dissimilar to one another in DAPPER than in the WESAD baseline.

KNN clusters reinforce the primary problem incident in the DAPPER cohort; faulty sensors or edge case scenarios where all patients share the defining characteristic (all depressed to some degree) can lead to poor separability of patient subgroups. This repeated observation of the result indicates an underlying incompatibility with DAPPER and our original intended application, mobile sensing. Without further information about the exact acquisition conditions, we can only make assumptions about the manner in which the data were collected and if there was a systemic error or methodological error.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 12. Clustering data using KNN ($K = 2$) on ECG and PPG data from WESAD and DAPPER, respectively. Each signal is plotted in 2-D using the first two captured values. Clustering results are shown using the point's color. WESAD groups are more homogeneous than DAPPER. (a) WESAD Class 0. (b) WESAD Class 1. (c) DAPPER Class 0 (depression). (d) DAPPER class 1 (depression). (e) DAPPER Class 0 (arousal). (f) DAPPER Class 1 (arousal). (g) DAPPER Class 0 (valence). (h) DAPPER Class 1 (valence).

This brings to light the larger issue of third-party datasets saturating the domain of medical intervention or emotion recognition in naturalistic settings without sufficient sanity checks. There is a lack of standard guidelines to quantify characteristics of datasets and quality of collection procedures owing to variability in devices, equity of demographic representation, and context of subjective questionnaire logging [57]. Recently, initiatives such as the mobilize-D procedure [58] provide a systematic approach for data standardization with considerations of heterogeneity in acquisition. Particularly, in the case of emotion recognition, the latent complexity of human emotion in the expression of multiple emotions and its interplay with physiological responses renders it difficult to produce actionable outputs [59]. Finally, there is always the tradeoff between the privacy of personal data and the accuracy of developed models that need to be considered when new datasets are introduced [60].

V. Conclusion

This work proposed an ensemble deep learning model with an attention mechanism for the purpose of binary classification of arousal, valence, and depression states with sensor data derived from wearable devices to be utilized in naturalistic settings. The signals were acquired from the DAPPER dataset and belonged to individuals in the 30-min time period prior to being subject to an ESM sampling questionnaire. The raw values were transformed into different image representations, and we empirically found that CWT and STFT were able to achieve the best sensitivity scores of 58.75%:45.59%, 62.34%:43.41%, and 49.43%:51.70% for predicting depression, valence, and arousal with a mixture of unimodality and bimodality with CWTs and STFT: ACCEL-CWT, GSR-CWT & ACCEL-STFT, and ACCEL-STFT. We come to the conclusion that low-cost sensor readings from the DAPPER dataset may not be sufficient to capture the complexities of emotional and mental state, and likely have utility in being an auxiliary modality alongside more individual characteristics, such as age and gender. To the best of our knowledge, this is one of the first experiments in the domain of emotional state variations measured by affective states among a predominantly depressed population and serves as a benchmark for developing algorithms using the DAPPER dataset. Future work can address self-supervision and multilabel extensions to this study and implement models capable of detecting fine-grained emotions (tense, angry, happy, and so on) from arousal and valence measures.

References

[1] T. Ota, "Chaotic use of depression-related medical terms: How should it be settled?" *Seishin Shinkeigaku Zasshi*, vol. 115, no. 3, pp. 261–266, 2013.

[2] C. Kraus, A. Mkrtchian, B. Kadriu, A. C. Nugent, C. A. Zarate, and J. W. Evans, "Evaluating global brain connectivity as an imaging marker for depression: Influence of preprocessing strategies and placebo-controlled ketamine treatment," *Neuropsychopharmacology*, vol. 45, no. 6, pp. 982–989, May 2020.

[3] R. Hartmann, F. M. Schmidt, C. Sander, and U. Hegerl, "Heart rate variability as indicator of clinical state in depression," *Frontiers Psychiatry*, vol. 9, p. 735, Jan. 2019.

[4] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[5] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in *Mobile Computing, Applications, and Services* (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), M. Gris and G. Yang, Eds. Berlin, Germany: Springer, 2012, pp. 211–230.

[6] S. Mitsue and T. Yamamoto, "Relationship between depression and movement quality in normal young adults," *J. Phys. Therapy Sci.*, vol. 31, no. 10, pp. 819–822, 2019.

[7] 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, Jan. 2015, Art. no. e0115894.

[8] K. S. Young, C. E. Parsons, A. Stein, and M. L. Kringelbach, "Motion and emotion: Depression reduces psychomotor performance and alters affective movements in caregiving interactions," *Frontiers Behav. Neurosci.*, vol. 9, p. 26, Feb. 2015.

[9] E. Harmon-Jones, P. A. Gable, and T. F. Price, "Does negative affect always narrow and positive affect always broaden the mind? Considering the influence of motivational intensity on cognitive scope," *Current Directions Psychol. Sci.*, vol. 22, no. 4, pp. 301–307, Aug. 2013.

[10] L.-C. Yu et al., "Building Chinese affective resources in valence-arousal dimensions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 540–545.

[11] G. Castellano, S. D. Villalba, and A. Camurri, in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, in Lecture Notes in Computer Science. Lisbon, Portugal: IEEE, 2007, pp. 71–82. [Online]. Available: https://ieeexplore.ieee.org/xpl/conhome/1002992/all-proceedings

[12] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

[13] W. Li and P. Pasquier, "Automatic affect classification of human motion capture sequences in the valence-arousal model," in *Proc. 3rd Int. Symp. Movement Comput.*, Jul. 2016, pp. 1–8.

[14] W. Han, X. Feng, M. Zhang, K. Peng, and D. Zhang, "Mood states and everyday creativity: Employing an experience sampling method and a day reconstruction method," *Frontiers Psychol.*, vol. 10, pp. 1–12, Jul. 2019.

[15] X. Shui, M. Zhang, Z. Li, X. Hu, F. Wang, and D. Zhang, "A dataset of daily ambulatory psychological and physiological recording for emotion research," *Sci. Data*, vol. 8, no. 1, p. 161, Jun. 2021.

[16] M. S. Zitouni, C. Y. Park, U. Lee, L. Hadjileontiadis, and A. Khandoker, "Arousal-valence classification from peripheral physiological signals using long short-term memory networks," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 686–689.

[17] E. J. Choi and D. K. Kim, "Arousal and valence classification model based on long short-term memory and DEAP data for mental healthcare management," *Healthcare Informat. Res.*, vol. 24, no. 4, pp. 309–316, Oct. 2018.

[18] S. Dockray, N. Grant, A. A. Stone, D. Kahneman, J. Wardle, and A. Steptoe, "A comparison of affect ratings obtained with ecological momentary assessment and the day reconstruction method," *Social Indicators Res.*, vol. 99, no. 2, pp. 269–283, Nov. 2010.

[19] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005.

[20] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," 2015, *arXiv:1506.00327*.

[21] A. Sarkar, S. K. S. Hossain, and R. Sarkar, "Human activity recognition from sensor data using spatial attention-aided CNN with genetic algorithm," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 5165–5191, 2023, doi: 10.1007/s00521-022-07911-0.

[22] A. S. Guinea, M. Sarabchian, and M. Mühlhäuser, "Improving wearable-based activity recognition using image representations," *Sensors*, vol. 22, no. 5, p. 1840, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8914937/

[23] A. Nedorubova, A. Kadyrova, and A. Khlyupin, "Human activity recognition using continuous wavelet transform and convolutional neural networks," 2021, *arXiv:2106.12666*.

[24] J. C. Burgess, "Applications of digital signal processing, edited by Alan V. Oppenheim," *J. Acoust. Soc. Amer.*, vol. 65, no. 5, p. 1354, May 1979.

[25] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153–183, Jan. 2009.

[26] M. K. Kıymık, İ. Güler, A. Dizibüyük, and M. Akın, "Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application," *Comput. Biol. Med.*, vol. 35, no. 7, pp. 603–616, Oct. 2005.

[27] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London, A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[28] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," 2018, *arXiv:1804.02391*.

[29] Y. Kong and T. Yu, "A deep neural network model using random forest to extract feature representation for gene expression data classification," *Sci. Rep.*, vol. 8, no. 1, p. 16477, Nov. 2018.

[30] S. A. Nasrat, U. Lee, M. S. Zitouni, A. H. Khandoker, S. Kang, and H. F. Jelinek, "Emotion recognition in the wild from long-term heart rate recording using wearable sensor and deep learning ensemble classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1676–1678.

[31] J. Ramesh, N. Keeran, A. Sagahyroon, and F. Aloul, "Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning," *Healthcare*, vol. 9, no. 11, p. 1450, Oct. 2021.

[32] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

[33] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021.

[34] C.-A. Wang, T. Baird, J. Huang, J. D. Coutinho, D. C. Brien, and D. P. Munoz, "Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task," *Frontiers Neurol.*, vol. 9, pp. 1–12, Dec. 2018.

[35] W. Sato, T. Kochiyama, and S. Yoshikawa, "Physiological correlates of subjective emotional valence and arousal dynamics while viewing films," *Biol. Psychol.*, vol. 157, Nov. 2020, Art. no. 107974.

[36] S. Ghiasi, A. Greco, R. Barbieri, E. P. Scilingo, and G. Valenza, "Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge," *Sci. Rep.*, vol. 10, no. 1, p. 5406, Mar. 2020.

[37] K. Weimann and T. O. F. Conrad, "Transfer learning for ECG classification," *Sci. Rep.*, vol. 11, no. 1, p. 5251, Mar. 2021.

[38] M. M. Leeflang, A. W. Rutjes, J. B. Reitsma, L. Hooft, and P. M. Bossuyt, "Variation of a test's sensitivity and specificity with disease prevalence," *CMAJ, Can. Med. Assoc. J.*, vol. 185, no. 11, pp. E537–E544, Aug. 2013.

[39] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, "The relation between valence and arousal in subjective experience," *Psychol. Bull.*, vol. 139, no. 4, pp. 917–940, 2013.

[40] A. H. Y. Chu, R. M. van Dam, S. J. H. Biddle, C. S. Tan, D. Koh, and F. Müller-Riemenschneider, "Self-reported domain-specific and accelerometer-based physical activity and sedentary behaviour in relation to psychological distress among an urban Asian population," *Int. J. Behav. Nutrition Phys. Activity*, vol. 15, no. 1, p. 36, Apr. 2018.

[41] P. Prociow, K. Wac, and J. Crowe, "Mobile psychiatry: Towards improving the care for bipolar disorder," *Int. J. Mental Health Syst.*, vol. 6, no. 1, p. 5, 2012.

[42] I. Habes et al., "Pattern classification of valence in depression," *NeuroImage, Clin.*, vol. 2, pp. 675–683, Jan. 2013.

[43] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, Jan. 2020.

[44] A. Kołakowska, W. Szwoch, and M. Szwoch, "A review of emotion recognition methods based on data acquired via smartphone sensors," *Sensors*, vol. 20, no. 21, p. 6367, Nov. 2020.

[45] P. Greenland and S. Hassan, "Precision preventive medicine-ready for prime time?" *JAMA Internal Med.*, vol. 179, no. 5, pp. 605–606, May 2019.

[46] T. Saito, H. Suzuki, and A. Kishi, "Predictive modeling of mental illness onset using wearable devices and medical examination data: Machine learning approach," *Frontiers Digit. Health*, vol. 4, pp. 1–13, Apr. 2022.

[47] A. Ahmed, J. Ramesh, S. Ganguly, R. Aburukba, A. Sagahyroon, and F. Aloul, "Investigating the feasibility of assessing depression severity and valence-arousal with wearable sensors using discrete wavelet transforms and machine learning," *Information*, vol. 13, no. 9, p. 406, Aug. 2022.

[48] N. Moukaddam, A. Sano, R. Salas, Z. Hammal, and A. Sabharwal, "Turning data into better mental health: Past, present, and future," *Frontiers Digit. Health*, vol. 4, pp. 1–16, Aug. 2022.

[49] P. D. Harvey, A. Khan, and R. S. E. Keefe, "Using the positive and negative syndrome scale (PANSS) to define different domains of negative symptoms: Prediction of everyday functioning by impairments in emotional expression and emotional experience," *Innov. Clin. Neurosci.*, vol. 14, nos. 11–12, pp. 18–22, Dec. 2017.

[50] L. Zhang, H. Fan, S. Wang, and H. Li, "The effect of emotional arousal on inhibition of return among youth with depressive tendency," *Frontiers Psychol.*, vol. 10, pp. 1–13, Jul. 2019.

[51] H. Teismann, J. Kissler, and K. Berger, "Investigating the roles of age, sex, depression, and anxiety for valence and arousal ratings of words: A population-based study," *BMC Psychol.*, vol. 8, no. 1, p. 118, Nov. 2020.

[52] M. Sheikh, M. Qassem, and P. A. Kyriacou, "Wearable, environmental, and smartphone-based passive sensing for mental health monitoring," *Frontiers Digit. Health*, vol. 3, Apr. 2021, Art. no. 662811.

[53] B. A. Hickey et al., "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review," *Sensors*, vol. 21, no. 10, p. 3461, May 2021.

[54] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, and J. Tørresen, "Mental health monitoring with multimodal sensing and machine learning: A survey," *Pervas. Mobile Comput.*, vol. 51, pp. 1–26, Dec. 2018.

[55] Z. Yu, M. Guindani, S. F. Grieco, L. Chen, T. C. Holmes, and X. Xu, "Beyond t test and ANOVA: Applications of mixed-effects models for more rigorous statistical analysis in neuroscience research," *Neuron*, vol. 110, no. 1, pp. 21–35, Jan. 2022.

[56] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 400–408.

[57] S. Canali, V. Schiaffonati, and A. Aliverti, "Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness," *PLOS Digit. Health*, vol. 1, no. 10, Oct. 2022, Art. no. e0000104.

[58] L. Palmerini et al., "Mobility recorded by wearable devices and gold standards: The Mobilise-D procedure for data standardization," *Sci. Data*, vol. 10, no. 1, p. 38, Jan. 2023.

[59] S. Saganowski, B. Perz, A. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Trans. Affect. Comput.*, early access, May 20, 2022, doi: 10.1109/TAFFC.2022.3176135.

[60] The Lancet Digital Health, "Wearable health data privacy," *Lancet Digit. Health*, vol. 5, no. 4, p. e174, Apr. 2023.

Abdullah Ahmed received the B.Sc. degree in computer engineering from the American University of Sharjah, Sharjah, United Arab Emirates, in 2018. He is currently pursuing the M.Sc. degree in electrical and computer engineering with the University of Massachusetts Amherst, Amherst, MA, USA.

Since 2022, he has been a Research Assistant with the Center for Human Health and Performance, University of Massachusetts Amherst, studying a variety of human-centered computing applications and systems with a focus on the relationship between physiology and affect.

Jayroop Ramesh (Student Member, IEEE) received the B.Sc. (cum laude) and M.Sc. degrees in computer engineering from the American University of Sharjah, Sharjah, United Arab Emirates, in 2019 and 2022, respectively. He will pursue the Ph.D. degree in computer science with the University of Oxford, Oxford, U.K.

He is a Research Assistant at the American University of Sharjah, United Arab Emirates. As a graduate research assistant, he has authored multiple academic publications in international journals and conferences focusing on health informatics, biophysical signal processing, machine learning, and cloud optimization.

Sandipan Ganguly received the B.Sc. degree in computer science from University College London, London, U.K. in 2018, and the M.Sc. degree in intelligent systems from King's College London, London, in 2019.

During his studies, his areas of interest were using Natural Language Processing (NLP) to gauge and find patterns in user sentiment to further explore the impact of public opinions, emotions, and arguments from online blogs and forums on different industries, such as the Film Box Office and Social Media Engagement Statistics, University of Khartoum, Sudan. He is currently a Software Engineer with Welldoc, an artificial intelligence (AI)-driven digital healthcare platform that coaches patients on multiple chronic conditions and comorbidities, and also a Independent Researcher experienced in using artificial intelligence (AI)-driven digital approaches for both academia and corporate entities.

Raafat Aburukba (Member, IEEE) received the bachelor's degree in computer science and software engineering and the master's and Ph.D. degrees in computer engineering from the University of Western Ontario, London, ON, Canada, in 2002, 2005, and 2013, respectively.

Prior to joining the American University of Sharjah (AUS), Sharjah, United Arab Emirates, he was a Faculty Member with the Department of Computer Science and Software Engineering, Pennsylvania State University, Erie, PA, USA. He is currently an Assistant Professor with the Department of Computer Science and Engineering, AUS. His research interests are on cloud computing middleware and applications, edge computing, cooperation and coordination in distributed systems, privacy in distributed systems, economic-based models, and approaches for decentralized scheduling and application to cloud computing and smart spaces.

Assim Sagahyroon (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Khartoum, in 1981, the M.Sc. degree in electrical engineering from Northwestern University, Evanston, IL, USA, in 1984, and the Ph.D. degree from The University of Arizona, Tucson, AZ, USA, in 1989.

From 1993 to 1999, he has been a member of the Department of Computer Science and Engineering, Northern Arizona University, Flagstaff, AZ, USA. In 1999, he joined the Department of Math and Computer Science, California State University, Long Beach, CA, USA. In 2003, he joined the Department of Computer Science and Engineering, American University of Sharjah, Sharjah, United Arab Emirates, where he served as the Department Head for seven years. He is currently a Professor of Computer Science and Engineering and the Associate Dean of Undergraduate Affairs with the American University of Sharjah. He has many publications in international conferences and journals. His research interests include innovative applications of emerging technology in the medical field, power consumption of portable electronics, and field programmable gate arrays (FPGA)-based design.

Dr. Sagahyroon was an Invited Technical Reviewer of the National Science Foundation Programs and served on technical program committees of many international conferences.

Fadi Aloul (Senior Member, IEEE) received the B.S. (summa cum laude) degree in electrical engineering from Lawrence Technological University, Southfield, MI, USA, in 1997, and the M.S. and Ph.D. degrees in computer science and engineering from the University of Michigan, Ann Arbor, MI, USA, in 1999 and 2003, respectively.

He is currently a Professor of Computer Science and Engineering and the Dean of Engineering with the American University of Sharjah (AUS), Sharjah, United Arab Emirates. He is also the Director of the HP Institute, American University of Sharjah. He has more than 130 publications in international journals and conferences, in addition to one U.S. patent. His current research interests include cyber security, mobile applications, and design optimization.

Dr. Aloul received a number of awards, including the Global Engineering Deans Council (GEDC) Airbus Engineering Diversity Award, the Sheikh Khalifa Award for Higher Education, the AUS Excellence in Teaching Award, the Abdul Hameed Shoman Award for Young Arab Researchers, and the Sheikh Rashid's Award for Outstanding Scientific Achievement. He is a regularly invited speaker and a panelist across a number of international conferences related to cyber security, technology, innovation, and education. He is a Certified Information Systems Security Professional (CISSP).