Original Paper

Developing a Machine Learning–Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study

Pooja Guhan[1], PhD; Naman Awasthi[1], PhD; Kathryn McDonald[2], BS; Kristin Bussell[3], PhD; Gloria Reeves[2], MD; Dinesh Manocha[1], PhD; Aniket Bera[4], PhD

[1]Department of Computer Science, University of Maryland, College Park, MD, United States

[2]Department of Psychiatry, Child and Adolescent Division, University of Maryland, Baltimore, MD, United States

[3]School of Nursing, University of Maryland, Baltimore, MD, United States

[4]Department of Computer Science, Purdue University, West Lafayett, IN, United States

Corresponding Author:
Pooja Guhan, PhD
Department of Computer Science
University of Maryland
8125 Paint Branch Dr
College Park, MD, 20742
United States
Phone: 1 2406309133
Fax: 1 (301) 405 6707
Email: pguhan@umd.edu

Abstract

Background: Patient engagement is a critical but challenging public health priority in behavioral health care. During telehealth sessions, health care providers need to rely predominantly on verbal strategies rather than typical nonverbal cues to effectively engage patients. Hence, the typical patient engagement behaviors are now different, and health care provider training on telehealth patient engagement is unavailable or quite limited. Therefore, we explore the application of machine learning for estimating patient engagement. This can assist psychotherapists in the development of a therapeutic relationship with the patient and enhance patient engagement in the treatment of mental health conditions during tele–mental health sessions.

Objective: This study aimed to examine the ability of machine learning models to estimate patient engagement levels during a tele–mental health session and understand whether the machine learning approach could support therapeutic engagement between the client and psychotherapist.

Methods: We proposed a multimodal learning-based approach. We uniquely leveraged latent vectors corresponding to affective and cognitive features frequently used in psychology literature to understand a person's level of engagement. Given the labeled data constraints that exist in health care, we explored a semisupervised learning solution. To support the development of similar technologies for telehealth, we also plan to release a dataset called Multimodal Engagement Detection in Clinical Analysis (MEDICA). This dataset includes 1229 video clips, each lasting 3 seconds. In addition, we present experiments conducted on this dataset, along with real-world tests that demonstrate the effectiveness of our method.

Results: Our algorithm reports a 40% improvement in root mean square error over state-of-the-art methods for engagement estimation. In our real-world tests on 438 video clips from psychotherapy sessions with 20 patients, in comparison to prior methods, positive correlations were observed between psychotherapists' Working Alliance Inventory scores and our mean and median engagement level estimates. This indicates the potential of the proposed model to present patient engagement estimations that align well with the engagement measures used by psychotherapists.

Conclusions: Patient engagement has been identified as being important to improve therapeutic alliance. However, limited research has been conducted to measure this in a telehealth setting, where the therapist lacks conventional cues to make a confident assessment. The algorithm developed is an attempt to model person-oriented engagement modeling theories within machine learning frameworks to estimate the level of engagement of the patient accurately and reliably in telehealth. The results are encouraging and emphasize the value of combining psychology and machine learning to understand patient engagement. Further testing in the real-world setting is necessary to fully assess its usefulness in helping therapists gauge patient engagement during online sessions. However, the proposed approach and the creation of the new dataset, MEDICA, open avenues for future research and the development of impactful tools for telehealth.

(JMIR Form Res 2025;9:e46390) doi: 10.2196/46390

KEYWORDS

machine learning; mental health; telehealth; engagement detection; patient engagement

Introduction

Overview

The World Health Organization defines mental health as a “state of well-being” that allows a person to lead a fulfilling and productive life and contribute to society [1]. In addition, the World Health Organization estimates that one-fourth of the adult population is affected by a mental disorder [2]. However, there are only approximately 9 psychiatrists per 100,000 people in higher-income countries and only approximately 0.1 psychiatrists for every 1,000,000 people in lower-income countries [3,4]. As stress and pressure continue to increase, leading to poor mental health outcomes, the need for improved tele–mental health care has become critical. Tele–mental health care is the process of providing psychotherapy remotely, typically using Health Insurance Portability and Accountability Act (HIPAA)–compliant videoconferencing [5]. It offers an effective means of accessing mental health services and treatment, transcending geographical and cultural boundaries worldwide, and helps address the chronic shortage of psychotherapists. These services not only remove practical barriers to care, such as transportation, but also offer affordability and direct access to qualified therapists. Therefore, there has been an upward trend in the demand for such services [6]. Despite these undeniable benefits, this emerging treatment modality raises new challenges in patient engagement compared to in-person care. By “engagement,” we refer to the connection between a therapist and patient, characterized by a sense of basic trust and a willingness or interest to collaborate. This connection is essential for the therapeutic process and fosters the development of a strong therapeutic relationship.

Background

Patient engagement is a critical priority in behavioral health care as it involves establishing effective rapport between health care providers and patients. In the context of mental health, patient engagement is an indicator of a successful therapy session. This multidimensional concept [7,8] encompasses the interaction of cognitive and emotion-related components of the patient’s psychological state. However, one of the challenges in fully understanding and evaluating patient engagement lies in the difficulty of establishing a standardized measurement approach. Patient engagement is a complex and dynamic construct that defies a one-size-fits-all measurement. The diverse factors influencing patient engagement, such as individual characteristics, and health care contexts make it challenging to develop a universally accepted standard of measuring its effectiveness. Traditional quantitative metrics (appointment attendance rates, treatment adherence, and patient satisfaction) often fall short in capturing the depth and richness of patient engagement. While they can provide numerical data, they may overlook the subjective experiences of the interaction and do not necessarily reflect health care provider-patient alliance in treatment. In contrast, qualitative methods such as patient narratives and feedback provide valuable insights but lack scalability, objectivity, and uniformity across different health systems. Therefore, to address these limitations, machine learning techniques are being explored to complement and enhance these traditional methods by leveraging computational power, scalability, and ability to identify patterns, relationships, and insights that may not be otherwise apparent. Some prior works in engagement detection have focused on using a single modality, such as facial expressions [9,10], speech [11], body posture [12], gaze direction [13], or head pose [14], to detect engagement. Combining different modalities has been observed to improve engagement detection accuracy [15-17]. The authors expanded the work of Stanford PBL Lab’s eRing [18] by including information streams such as facial expressions, voice, and other biometric data. Monkaresi et al [19] proposed an approach to detect engagement levels in students during a writing task by not only making use of facial features but also features obtained from remote video-based detection of heart rate. The dataset used was generated by the authors, and they used self-reports instead of external annotation for classification purposes. Chang et al [20] made use of facial expressions as well as body posture for detecting engagement in learners. Fedotov et al [21] proposed the use of audio, facial, and body pose features to detect engagement and disengagement for an in-the-wild dataset that had been created using videos obtained from uncontrolled environments or situations. The distribution of class labels in this dataset is not balanced.

Despite the existence of a variety of such algorithms to perform engagement detection, the results obtained from these approaches (particularly single modality–based approaches) could be misleading in a telehealth setting due to factors such as camera position and resistant or guarded clients. In telehealth appointments, therapists have limited visual data available to them, as they can only view the patient’s face rather than their full body. Asking the patient to adjust their position to gain a better view of their body during telehealth appointments is a potential solution for therapists. However, this approach may lead to communication difficulties, as it can impact the patient’s proximity to the microphone, potentially affecting the quality of audio communication between the patient and the therapist. Therefore, therapists must rely on verbal communication strategies to engage patients than in-person care because they cannot use typical nonverbal cues to convey interest and be responsive to the patient (eg, a handshake at the beginning of a session, adjusting the distance between the patient and health care provider by moving a chair closer or further away, and observing a patient's response to questions while maintaining eye contact). It is also more difficult for therapists to convey attentiveness because eye contact requires the therapist to look at a camera rather than observe or look at a person. They may also have limited training on telehealth patient engagement, leading to a lack of guidance on measuring patient engagement. Therefore, this highlights the need for a system that can provide feedback on patient engagement by just using the data that are easily accessible, namely, text, audio, and face visuals. Engagement is critical for both retention of patients in care as well as effectiveness of mental health treatment. Developing such a system will make it possible to enhance the quality of tele–mental health and ultimately improve patient outcomes.

Objective

Taking all this into consideration, we propose and explore a machine learning–based approach that takes visual, audio, and text data as input and estimates the engagement level of a patient with mental health problems through a regression-based approach. We demonstrate the effectiveness of our method in tele–mental health sessions and provide a new dataset called Multimodal Engagement Detection in Clinical Analysis (MEDICA) to advance mental health research by understanding patient engagement levels during therapy sessions. MEDICA consists of 1229 short video clips from mock mental health therapy sessions used by medical schools in their psychiatry teaching curriculum. To the best of our knowledge, MEDICA is the first multimodal dataset that focuses on mental health research and consists of annotations useful for understanding the psychological state of patients with mental health problems during a therapy session.

Methods

Datasets

MEDICA Dataset

Overview

Engagement is an overloaded term, and the definition varies with the application, making it difficult and expensive to collect, annotate, and analyze such data. As a result, we find too few multimodal-based engagement detection datasets currently available for us to use. Our problem statement revolves specifically around detecting patient engagement during a tele–mental health session. In such a setting, the only information we can work with includes the patient's face and speech (audio and text). There exist datasets such as Carnegie Mellon University Multimodal Corpus of Sentimental Intensity (CMU-MOSI) [22], Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [23], and Stanford Emotional Narratives Dataset (SEND) [24] that capture such settings. However, they are not specifically for engagement detection. Given the lack of a dataset that allows researchers to use multimodal features (video, text, and audio) for engagement, we propose MEDICA, a novel dataset developed specifically to cater to engagement detection using tele–mental health session videos. To use these data to address a broader range of issues related to mental health, we also include labels pertaining to stress and emotions.

To the best of our knowledge, this dataset is one of the first publicly available datasets specifically catering to multimodal research in patient engagement in mental health. Table 1 presents a comparison between MEDICA and other related datasets. Despite the rise in telehealth services and the poor patient-to-therapist ratios for individuals with mental health problems, there are no datasets that even try modeling telehealth sessions to give the community an opportunity to innovate and develop new technologies. MEDICA is a humble attempt by us to kick-start interesting research opportunities. The MEDICA dataset is unique in that it contains multimodal data from multiple sensors including speech and video. These data have been annotated with the engagement levels exhibited by each participant at different points in time.

Table 1. Comparison of the Multimodal Engagement Detection in Clinical Analysis (MEDICA) dataset with other related datasets[a].

Dataset name	Samples	Unique speakers	Modes	Emotion labels	Engagement information	Other mental health cues
RECOLA[b] [25]	3400	27	{v,a}	Yes	No	Physiological (electrocardiogram and electrothermal activity)
CMU MOSEAS[c] [26]	715	Multiple	{v,a,t}	Yes	No	—[d]
CMU-MOSI[e] [22]	2199	Multiple	{v,a,t}	Yes	No	—
CMU-MOSEI[f] [23]	3228	1000	{v,a,t}	Yes	No	—
SEND[g] [24]	193	49	{v,a,t}	Yes	No	—
DAiSEE[h] [27]	9068	112	{v}	No	Yes	—
HBCU[i] [9]	120	34	{v}	No	Yes	—
In the wild [28]	195	78	{v}	No	Yes	—
SDMATH [29]	20	20	{v,a}	Yes	Yes	—
MEDICA	1229[b]	13	{v,a,t}	Yes	Yes	Hesitation, stress, and attentiveness

[a] Modes indicate the subset of modalities present from visual (v), audio (a), and text (t).

[b] RECOLA: Remote Collaborative and Affective.

[c] CMU MOSEAS: Carnegie Mellon University Multimodal Opinion Sentiment, Emotions and Attributes.

[d] Not applicable.

[e] CMU-MOSI: Carnegie Mellon University Multimodal Corpus of Sentiment Intensity.

[f] CMU-MOSEI: Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity.

[g] SEND: Stanford Emotional Narratives Dataset.

[h] DAiSEE: Dataset for Affective States in E-Environments.

[i] HBCU: historically black college and university.

[j] Current status of the dataset. The size of the dataset will be increased.

Acquisition

MEDICA has been developed by collecting publicly available mock therapy session videos created by different psychiatry medical schools for training their students. The patients in these videos are being advised for depression, social anxiety, and posttraumatic stress disorder. We collected 13 videos, each having a duration of approximately 20 to 30 minutes. We limit the videos to the setup where both the therapist and the patient are not visible together in the same frame. In addition, we take only those videos where there is only 1 patient. Each video has a unique English-speaking patient. The therapists in these videos are actual therapists, while the patients are trained actors who have been dictated to act and react in a particular way. Figure 1 shows the snapshots taken from various video clips in the MEDICA dataset.

Figure 1. Examples from the Multimodal Engagement Detection in Clinical Analysis (MEDICA) dataset created for mental health research. This dataset has been created using publicly available videos that are usually used for training purposes by different medical schools. The faces of the individuals have been blurred to protect their identity.

Processing and Annotation

Our primary focus was to create a dataset that captures the behavior of patients with mental health problems during their sessions. Therefore, we selected only portions of the videos where only the patient was visible in the frames, which appeared at various intervals and durations throughout the recordings. We took these scattered clips and divided them into smaller clips of 3 seconds each, resulting in a dataset of size 1229. We use Moviepy and speech-recognition libraries to extract audio and text from the video clips, respectively. Each video was annotated for attentiveness, stress, and engagement, which were scored on a Likert scale ranging from –3 to 3; hesitation was a binary target variable (yes or no). Humans tend to have multiple emotions with varying intensities while expressing their thoughts and feelings. Therefore, the videos have been labeled for 8 emotions related to mental health: happy, sad, irritated, neutral, anxious, embarrassed, scared, and surprised. This will enable us to develop systems capable of understanding the various interacting emotions of the users.

Real-World Data

The videos collected to develop MEDICA were created by the psychiatry schools in controlled settings. By controlled, we mean that the therapists in these videos were real, while the patients were trained method actors who had been instructed in a very detailed manner regarding what to expect and how they should react. While these videos were approved by a team of psychotherapists as being inclusive of potential scenarios and patient reactions, the work would still be meaningless if not tested in the wild, that is, in settings where we have real patients who have not been instructed what to do or how to behave during the session with their real therapist. To accomplish this, we collaborated with 8 child psychotherapists. These psychotherapists are individuals who were trained and licensed as psychotherapists for both children and adults. The administrative staff identified patients that were scheduled for sessions with any of the participating therapists and provided the caregiver contact information. Our research assistant contacted eligible caregivers and provided information related to the purpose, potential risks and benefits, and responsibilities of study participation. They were informed of the study protocol that involved recording 1 to 2 telemental sessions to test our proposed approach. A total of 20 caregivers provided consent to participate in the study. They were provided with instructions on the setup of the equipment that we provided to ensure a clean recording. “Clean” refers to a recording executed with a camera of good quality with appropriate lighting conditions. The equipment included a smartphone with a good quality camera, a ring light with a stand to ensure that the session was recorded in a well-lit environment, and a hot spot internet connection to ensure that the session occurred smoothly without any network glitches. They were also given the assurance regarding preserving the confidentiality of the data being collected. All video recordings were labeled only by study number, with all identifiable information removed and stored in a double password–protected database. Participants were also informed that any facial images other than the caregiver who participated in the session would be deidentified (using methods such as blurring, etc). These steps were taken to ensure that only the caregiver’s level of engagement was measured and analyzed. In addition, this step protected the confidentiality of nonparticipants who did not provide consent.

On average, each of these sessions lasted approximately 20 minutes. The demographics of the caregivers who participated in our real-world experiments can be found in Multimedia Appendix 1. The entire data collection process can be divided into three parts:

1. Before each tele–mental health session of a caregiver with their therapist, a research assistant helped the caregiver with setting up the equipment to record their session. The assistant also ensured that the caregivers were comfortable using the equipment.
2. During the session, we made sure that the tele–mental health session proceeded in the same manner as a typical session. The caregivers were requested to use the equipment provided if possible, but there were no restrictions imposed by us on either the therapist or the caregiver regarding how the session should be conducted. After the presession process, the research assistant would log off. Therefore, during the session, it would be just the therapist and the caregiver having a conversation. No one else from the study would be a part of it. The only thing different about this session was that the caregiver was being recorded using the smartphone given to them. We did not record the therapist.
3. After the session, after the session was complete, a research assistant guided the participant regarding the steps to stop the recording and save the data collected.

Upon completion of the session, the therapist scored the quality of the collaborative relationship (therapeutic alliance) established between them during the session with the Working Alliance Inventory (WAI) [30,31]. It is used for both child and adult therapy and is a widely studied therapeutic alliance measure. The WAI was modeled on the theoretical work conducted by Bordin [32]. It captures 3 dimensions of the alliance: bond, task, and goals. Extensive tests showed 12 items per dimension to be the minimum length for effective representations of the inventory. A composite score is computed based on these 12 items for each of the sessions conducted. Henceforth, we refer to this score as the WAI score. The WAI scores can range from 12 to 84 [33,34]. Figure 2 shows frames from a few of the video recordings collected.

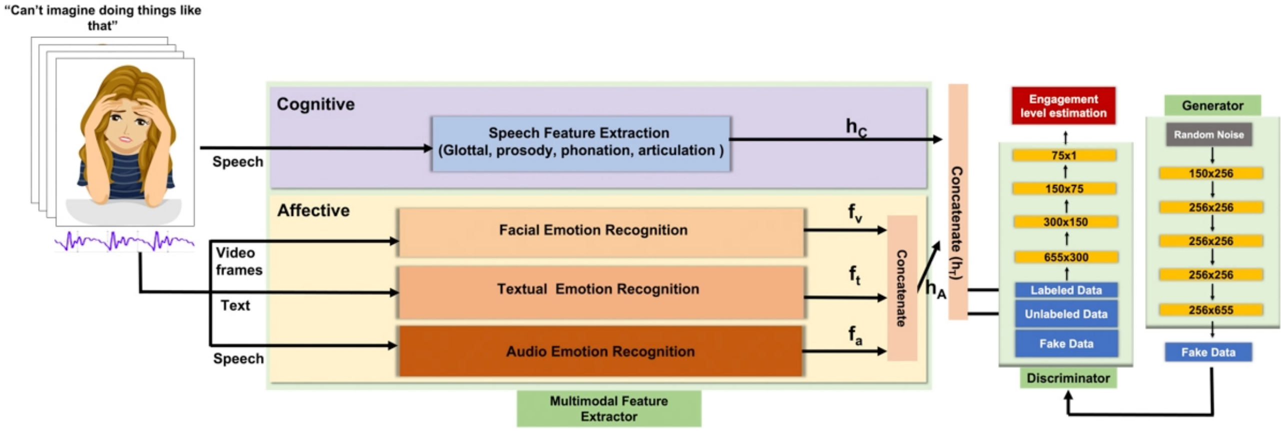
Figure 2. A few frames from the real-world videos we collected. The faces have been blurred here to protect the identity of the patients. However, the consent of the patients was taken to use their unblurred faces as input to our model.

While efforts were taken to ensure clean video recordings, we still could not use 7 out of the 20 videos collected for any type of analysis. The reasons for it ranged from the video being extremely shaky due to the continuous movement of some participants during their session to not having the permissions to process and analyze videos where the child spoke for the entire session instead of the caregiver. The remaining videos were processed in the same way as MEDICA. At the end of data processing, we obtained approximately 438 clips, each lasting 3 seconds. Unlike MEDICA where we had annotations for every 3 seconds, we had only a single WAI score for the entire session. Therefore, all the 3-second clips obtained from a session were annotated with the WAI score obtained for the session. Any analysis conducted on these data treated all the 438 clips independently.

Proposed Model Design

Figure 3 presents an overview of our engagement estimation model. It has 2 key components: *multimodal feature extractor* and *semisupervised learning network*. We discuss each of them in detail in the subsequent sections.

Figure 3. Overall block diagram of the proposed architecture. The features obtained from the 2 modalities, cognitive and affective, are combined to form hT (real) of shape 655×1. The generator takes in random noise (Gaussian) and tries to generate a fake feature map that looks similar to hfakeT that looks similar to hT (real). The discriminator tries to distinguish between hT and hfakeT. The yellow bars in the discriminator and generator refer to linear layers with activation function as leaky rectified linear activation unit. The shape of the linear layers used has been written on the bars. The number of linear layers used can be changed depending upon the dataset size and task at hand.


"Can't imagine doing things like that"
Speech
Video frames
Text
Speech
Cognitive
Speech Feature Extraction (Glottal, prosody, phonation, articulation)
h_C
Affective
Facial Emotion Recognition
Textual Emotion Recognition
Audio Emotion Recognition
f_v
f_t
f_a
Concatenate
h_A
Concatenate (h_T)
Multimodal Feature Extractor
Engagement level estimation
75x1
150x75
300x150
655x300
Labeled Data
Unlabeled Data
Fake Data
Discriminator
Generator
Random Noise
150x256
256x256
256x256
256x256
256x655
Fake Data

Multimodal Feature Extractor

Overview

Multimodal feature extractor in general refers to a system or algorithm that extracts and combines features from multiple modalities, such as audio, video, text, and other types of data, to provide a more comprehensive representation of input data. The goal of multimodal feature extraction is to capture complementary information from different sources, which can lead to better performance in various tasks. In our model, the multimodal feature extractor is used to comprehend the psychological state of the patient with mental health problems. We take inspiration from literature in psychology and psychiatry to build an extractor that can enable the learning network to recognize, understand, and evaluate engagement as close as possible to that of a psychotherapist. Therefore, we proposed a multicomponent approach to build the feature extractor and introduce 2 modules, namely, cognitive and affective states, to capture the different cues used by psychotherapists to assess their patients. As mentioned earlier, the visual data available are limited. Therefore, apart from the visuals, we also need to make use of verbal information, that is, understanding the speaking style of the person (audio) and what the person said (text).

Therefore, given this information corresponding to a participant, the multimodal feature extractor outputs a feature vector h_T. h_T is obtained after concatenating the feature vectors h_C and h_A obtained from the cognitive and affective state modalities (components of the multimodal feature extractor), respectively. We now discuss the cognitive and affective state modalities in more detail in the subsequent sections.

Cognitive State Modality

The cognitive state involves comprehending complex concepts and issues and acquiring difficult skills. It conveys deep (rather than surface level) processing of information, where the person gains a critical or higher-order understanding of the subject matter and solves challenging problems.

Psychotherapists usually measure and evaluate the cognitive state of the person using the mental status examination, typically conducted via in-person interviews or self-evaluations, to gauge memory, thinking, and the extent of understanding of the topic of discussion. There has been a lot of work around determining biomarkers for detecting signs of a person's cognitive state. However, these methods are either offline or fail to consider various essential perceptual indicators. Recently, there has been a lot of work around using speech as a potential biomarker for detecting cognitive decline. For instance, stress negatively affects the cognitive functions of a person, and this can be easily detected using speech signals. Moreover, speech-based methods are attractive because they are nonintrusive, inexpensive, and can be done in real time. Four distinct features have proven to be highly effective in identifying signs of cognitive impairment and are increasingly utilized to detect conditions like Alzheimer's and Parkinson's diseases. These features include glottal (f_g), prosody (f_{pr}), phonation (f_{ph}), and articulation measures (f_{ar}), each offering unique insights into speech characteristics associated with cognitive decline.

Glottal features are acoustic measures used to characterize speech under stress. The glottal is the space between the vocal cords in the larynx, and the opening and closing of the vocal cords create sound during speech. During periods of stress, there is an aberration in the amount of tension applied in the opening (abduction) and closing (adduction) of the vocal cords. Prosody features characterize the speaker's intonation and speaking styles. Under this feature, we analyze variables such as timing, pitch, and loudness during the production of speech. Phonation refers to the production of vocal sounds by the vibration of the vocal cords. In individuals with cognitive decline, phonation is often characterized by bowing and inadequate closure of the vocal cords. These changes can lead to instability and irregularity in the vibration of the vocal cords, which in turn affects the production of speech. They are analyzed in terms of features related to perturbation measures, such as jitter (temporal perturbations of the fundamental frequency), shimmer (temporal perturbation of the amplitude of the signal), amplitude perturbation quotient, and pitch perturbation quotient. Apart from these, the degree of unvoiced is also included. Articulation features refer to the movements of the lips, tongue, and jaw during speech production. Reduced amplitude and velocity of these movements can impact speech intelligibility. The analysis of articulation is primarily based on the first 2 vocal formants F1 and F2, which correspond to the frequency of the sound waves produced by the vocal tract.

For a given audio input, we extract these features using *librosa* [35] and *praat* [36] libraries. Therefore, we define features corresponding to cognitive state as the concatenation of these 4 audio features. Therefore, cognitive state features

h_c = concat (f_g, f_{pr}, f_{ph}, f_{ar}) **(1)**

Affective State Modality

The affective state encompasses emotional reactions such as excitement, boredom, curiosity, and anger. The range of affective expressions will vary based on individual demographic factors (eg, age), cultural backgrounds or norms, and mental health symptoms.

To understand the affective state, we check if there exists any inconsistency between the emotions perceived and the statement the person made. Balomenos et al [37] and Porter and ten Brinke [38] suggest that when different modalities are modeled and projected onto a common space, they should point to similar affective cues; otherwise, the incongruity suggests distraction, deception, etc. In other words, if E1, E2, and E3 represent the emotions perceived individually from what the patient said (text), the way they said it or sounded (audio), and how they looked or expressed (visuals), respectively, then the patient would be considered engaged if E1, E2, and E3 are similar; otherwise, they would be considered disengaged. Therefore, we adopt pretrained emotion recognition models to extract affective features corresponding to audio, visuals, and text from each video sample separately.

Audio features (f_a) are extracted using mel-frequency cepstral coefficients (MFCC) from audio clips available in the dataset. MFCC is a commonly used audio feature extraction technique in signal processing and machine learning. It is based on the concept of the mel scale, which is a nonlinear transformation of frequency that is more perceptually relevant to human hearing. By applying MFCC to audio signals, we can extract features that capture important aspects of the sound, such as pitch, timbre, and spectral shape. These features are often used in speech and music analysis as well as in various machine learning applications, including emotion recognition and speaker identification. In our work, the extracted MFCC features from audio are then passed as input to a multilayer perceptron network trained for emotion recognition in speech, specifically using the data available in the Crowd-source Emotional Multimodal Actors Dataset (CREMA-D) to extract affective features from the audio data. Through this method, we obtained a feature vector f_a, for each audio clip, which provides valuable insights into the emotional characteristics of the speech.

Visual features (f_v) correspond to the affective features extracted from video frames to understand the emotions expressed through facial cues of the patient. We used a deep learning architecture called Visual Geometry Group-B (VCG-B), which has been pretrained as described in the study by Arriaga et al [39], to extract the affective features. We modified the output dimensions of the second last layer of the network to produce a feature vector f_v of length 100.

Text (f_t), characterizing the emotions, sentiment, and other affective cues present in the text, is obtained using a Bidirectional Encoder Representations from Transformers(BERT)–based model, which has been pretrained on the GoEmotions dataset [40]. The Bidirectional Encoder Representations from Transformers model works by processing the text input through a multilayer neural network that is trained to predict the context of each word based on its surrounding words. This allows the model to capture the complex relationships between words and their meanings, which are essential for understanding the affective context of the text.

Therefore, we represent the affective state of the patient as a concatenation of f_a, f_v, and f_t. Hence, affective state features

h_A= concat (f_a, f_v, f_t) **(2)**

Therefore, the overall output of the multimodal feature extractor is a feature vector h_T*which* can be defined as

h_T = concat (h_C, h_A)

The learning network discussed in the next section makes use of this.

Learning Network

In this section, we discuss the second part of our proposed model, that is, the semisupervised learning network. Different machine learning techniques can be explored to solve the task of predicting the level of engagement. However, obtaining a large amount of high-quality labeled data to train a robust model for predicting patient engagement is inevitably laborious and requires expert medical knowledge. Considering that unlabeled data are relatively easy to collect, we propose a semisupervised learning–based solution. Semisupervised learning enables us to deploy machine learning systems in real-life applications (eg, image search [41], speech analysis [42,43], and natural language processing) where we have few labeled data samples and a lot of unlabeled data. Some prior works have also explored semisupervised learning for engagement detection in nonmedical domains. One of the earliest studies in this area is by Alyuz et al [44], where they consider the development of an engagement detection system, more specifically emotional or affective engagement of students, in a semisupervised fashion to personalize systems such as intelligent tutoring systems according to their needs. Nezami et al [45] conducted experiments to detect user engagement using a facial feature–based semisupervised model. Most state-of-the-art semisupervised learning methods use generative adversarial nets (GANs) [46].

GANs are a class of machine learning models and typically have 2 neural networks competing with each other to generate more accurate predictions. These 2 neural networks are referred to as the generator and the discriminator. The generator's goal is to artificially manufacture outputs (ie, fake data) that could easily be mistaken as real data. The goal of the discriminator is to identify the real from the artificially generated data. In trying to generate high-quality outputs, the generator learns to capture the different possible variations in the input variables, and therefore, the data manifold well. This is extremely helpful when we may not be able to access data containing a wide variety of similar engagement-related cues visible across different patients. In general, the training process of GANs involves a dynamic interplay and competition between the generator and discriminator networks. Both the generator and discriminator networks improve their abilities iteratively. The generator aims to generate more realistic samples, while the discriminator aims to become better at distinguishing between real and generated (artificial) data. The eventual goal of the training process is for the generator and discriminator to reach Nash equilibrium [47]. Once the equilibrium is achieved, no player (generator or discriminator) has any incentive to deviate toward a more optimal position, that is, no player has a better strategy given the choices of the other player.

Taking all this into consideration, we propose using a multimodal semisupervised GAN as our learning network for regressing on the levels of patient engagement. This type of machine learning model combines multiple modalities (in this case, cognitive and affective modalities) to generate realistic samples (or features) that can fool the discriminator. The term *semisupervised* refers to the network's ability to use a limited amount of labeled data alongside a larger amount of unlabeled data to enhance its accuracy. Such a network will generalize better and be more robust compared to previously defined semisupervised learning approaches. It is different than the semisupervised GAN framework SR-GAN proposed by Olmschenk et al [48] in 2 main ways. First, our network is multimodal, whereas SR-GAN is not. Second, unlike the SR-GAN generator, which focuses on generating realistic images, our GAN's generator is designed to create realistic feature maps that mimic the multimodal feature vector h_T, derived from the fusion of cognitive and affective state modalities (as discussed in the previous section). The generator takes Gaussian noise as input and produces a synthetic feature vector h_{fakeT}, aiming to deceive the discriminator into classifying it as h_{fakeT}. The discriminator's role, in turn, is to distinguish h_T as a fake representation and correctly identify h_T as the genuine feature vector.

Overall, our multimodal semisupervised GAN can therefore be divided into 3 parts: multimodal feature extractor, generator, and discriminator.

For the purpose of training, we make use of the following four standard loss functions, namely, labeled loss (L_{lab}), unlabeled loss (L_{un}), fake loss (L_{fake}), and generator loss (L_{gen}):

1. L_{lab}—this loss is similar to typical supervised learning regression losses, where the goal is to assess the network's ability to estimate the engagement level accurately, as close as possible to the ground truth (ie, the actual engagement level annotated for the data sample). Therefore, we compute the mean squared error of model output ($\hat{y}_t$, ie, predicted level of engagement) with ground truth (y_t, ie, actual level of engagement).

$$L_{lab} = MSE\ (y_t - \hat{y}_t) \quad \textbf{(3)}$$

2. L_{un}—in semisupervised GANs, the labeled dataset contains a small portion of data with labels, while most of the data are unlabeled. The goal of the L_{un} loss function is to improve the performance of the generator by minimizing the distance between the feature spaces of the labeled and unlabeled datasets. This helps to ensure that the generator produces feature maps that are consistent with both labeled and unlabeled data, thereby improving the generalization performance of the model. By minimizing the distance between the feature spaces of labeled and unlabeled datasets, the model can effectively learn from both labeled and unlabeled data, which is particularly useful in cases where labeled data is scarce.

$$L_{un} = ||\, E_{x\sim p_{labeled}}\, h_T - E_{x\sim p_{unlabeled}} h_T\, ||_2{}^2 \quad \textbf{(4)}$$

3. L_{fake}—it is a loss function used to update the discriminator network. The objective of this loss is to maximize the distance between the feature space of unlabeled dataset and fake features generated by the generator. This enables the discriminator to distinguish the real and fake data features better.

$$L_{fake} = -||\log\,(\,|\,E_{x\sim p_{fake}}\, h_{fakeT} - E_{x\sim p_{unlabeled}}\, h_T\,|\, +\, 1)||_1 \quad \textbf{(5)}$$

4. L_{gen}—the generator network's objective is to produce feature vectors h_T (fake) that can be passed off as real feature vectors h_T: $\mathrm{L}_{\mathrm{gen}}$ updates the generator network to do just that. $\mathrm{L}_{\mathrm{gen}}$ encourages the generator to learn the underlying distribution of unlabeled data features (real) and generate fake features that match the feature statistics of the real features.

$$L_{gen} = ||\, E_{x\sim p_{fake}}\, h_{fakeT} - E_{x\sim p_{unlabeled}}\, h_T\, ||_2{}^2 \quad \textbf{(6)}$$

Similar to SR-GAN, we also make use of a gradient penalty (P) to keep the gradient of the discriminator in check, which helps convergence. The gradient penalty is calculated with respect to a randomly chosen point on the convex manifold connecting the unlabeled feature vector h_T to the fake feature vector h_T. We compute it as

$$\mathrm{P}{=}E_{x\sim p_{interpolate}}\left[\max\left(\left(||\nabla_{\hat{x}}\,(h_T)||^2{}_2 - 1\right), 0\right)\right] \quad \textbf{(7)}$$

Here, $p_{interpolate}$ examples are generated by $\alpha p_{unlabeled}$ + $(1 - \alpha) p_{fake}$ for $\alpha \sim U\ (0,1)$.

$U\ (0,1)$ represents a uniform distribution over the range from 0 to 1.

The overall loss function used for training the network is:

$$\mathrm{L} = \mathrm{L}_{\mathrm{lab}} + \mathrm{L}_{\mathrm{un}} + \mathrm{L}_{\mathrm{fake}} + \mathrm{L}_{\mathrm{gen}} + \lambda \mathrm{P} \quad \textbf{(8)}$$

Overall Pipeline

In this section, we present an overview of how the various components of our proposed approach interact to estimate the engagement level of a patient with mental health problems during a tele–mental health session. We provide insights into the working mechanism of our model and shed light on the underlying processes involved in engagement level estimation in a multimodal setting.

As depicted in Figure 3, the proposed model involves data preprocessing, which includes extracting video frames, audio signal, and text data from the input video clip. To capture engagement dynamics over small time scales, we analyze engagement at the microlevel, considering video clips of a few seconds duration. This aligns with the person-oriented analysis suggested by Sinatra et al [49]. Specifically, for the MEDICA dataset and real-world data, we consider video clips of 3 seconds duration. All samples, labeled and unlabeled, from the dataset are treated as *real* data for the semisupervised GAN. The multimodal feature extractor (consisting of the cognitive and affective state modalities) is used to obtain the h_T features from the preprocessed visual, audio, and text data. h_T is considered real. In parallel, the generator network creates h_{fakeT} (ie, fake h_T) features using gaussian noise as input. The labeled and unlabeled h_T features obtained from the multimodal feature extractor and the h_{fakeT} features obtained from the generator are then input to the discriminator network. In addition to distinguishing between real (h_T) and fake (h_{fakeT}) features, the discriminator network outputs the engagement level of the patient with mental health problems as a continuous value.

The subsequent sections outline the different studies being conducted to verify our approach.

Study 1: Testing Our Proposed Approach on MEDICA

The purpose of the first study is to test the applicability of our approach to estimate the levels of engagement exhibited by the participants present in the videos. The test was conducted on the MEDICA dataset. In addition, we compared the performance of the proposed model against other similar frameworks to understand its effectiveness. To test the models, we partitioned the dataset in a ratio of 70:10:20 for training, testing, and validation. Motivated by recent works in clinical psychotherapy [50-52], we used the standard evaluation metric of root mean square error (RMSE) to evaluate our approach. Smaller the RMSE value, better is the performance of the model for the engagement estimation task.

Study 2: Ablation Studies

The proposed model design involves the introduction of 2 modalities, namely, affective and cognitive. In this study, we explored the contribution of these 2 modalities for the purpose of engagement detection. We aimed to test the effectiveness of the model with and without these modalities. Therefore, we ran the overall pipeline described earlier by removing either one of the modalities corresponding to affective and cognitive states and reported our findings. The tests were conducted on the MEDICA dataset. Similar to our previous study, we performed the evaluation using RMSE.

Study 3: Analysis on Real-World Data

Evaluating the model only on MEDICA is not sufficient to prove the usability of the model in the real-world setting. Therefore, this study aimed to test our approach on data collected from real-world scenarios. The study involved testing the proposed model that has been trained on MEDICA using real-world data. We also compared the performance of our proposed model against other baselines explored in study 1 on the real-world data. Apart from testing our model's usability for the real world, we also tested its capability to perform on unseen real-world data, which is different from the data (MEDICA) it has been trained on. Contrary to previous experiments where we compared the performance on engagement estimation values, in this test, we will study the correlation of the model-estimated engagement values with the WAI score. This will help us to also test whether the scores estimated by our model are clinically useful for therapists.

Ethical Considerations

Both the MEDICA dataset and the real-world data collection processes have been approved by the institutional review boards (#HP-00092966) at both the University of Maryland College Park's computer science department and University of Maryland Baltimore medical school.

Participation of the caregivers while collecting the real-world data was completely voluntary. They could withdraw anytime without any negative consequences. They were informed about the purpose of the study and the analysis that would be conducted using the data collected.

The therapist and the participating caregiver were given the option of allowing us to store videotaped recordings for development of future studies or to request to have all videotape sessions destroyed after a 2-year period. The video recordings were stored in our laboratory databases and subject to strict University of Maryland privacy protocol, including encryption and password protection. While the research is ongoing, only select project personnel with internal clearance can have access to the data. In addition, we ensured that the data are untampered using standard cryptographic hash functions. During the video storage process, we deidentified any facial images beside the health care provider or caregiver who are recorded (eg, sibling interrupts session) so that they are not recognizable on the recorded video. All data related to enrollment, demographics, and questionnaires will be stored in an electronic database at the University of Maryland and will be double password protected with only select research members having access. The participating caregiver was given an individual ID number, and their data were deidentified to ensure it could not be linked back to them.

While the participating caregiver may not directly benefit from the study, they received 1-month free internet hot spot and a Fitbit, which were provided as part of the real-world study equipment. They were compensated with a US $20 gift card after returning the equipment.

Results

The semisupervised models were trained on an NVIDIA GeForce GTX 1080 Ti GPU with a batch size of 512 and a learning rate of 0.0001 for 50 epochs.

Study 1: Testing Our Proposed Approach on MEDICA

The purpose of the first study is to demonstrate the ability of our model to estimate the level of engagement exhibited by the patient in the video. This study was performed on the MEDICA dataset. As our proposed methodology leverages a semisupervised approach, we extracted labeled samples from MEDICA and unlabeled samples from the CMU-MOSEI dataset. After preprocessing, we extracted 12,854 unlabeled data points from CMU-MOSEI. We split the 1229 labeled data points from MEDICA into 70:10:20 for training, validation, and testing, respectively. Therefore, the split of the labeled training data to unlabeled training data points was 860:12,854. In addition, the ground truth engagement levels corresponding to the training data points in the MEDICA dataset were normalized to fall within a range of 0 and 1. We compared our model with the following state-of-the-art methods for engagement detection:

1. Kaur et al [28] used a deep multiple instance learning–based framework for detecting engagement in students. They extracted local binary pattern on three orthogonal planes (LBP-TOP) features from the facial video segments and performed linear regression using a deep neural network to estimate the engagement scores.

2. Nezami et al [45] performed a semisupervised engagement detection using a semisupervised support vector machine.

In addition to being state-of-the-art, these methods can be used in a telehealth setting like ours. We used the publicly available implementation proposed by Kaur et al [28] and trained the entire model on MEDICA. Nezami et al [45] do not have a publicly available implementation. We reproduced the method to the best of our understanding.

Table 2 summarizes the RMSE values obtained for methods described by Kaur et al [28] and Nezami et al [45], including ours. We observed an improvement of at least 40%. Our approach is one of the first methods of engagement estimation built on the principles of psychotherapy. The modules used, specifically cognitive and affective states, help the overall model to effectively mimic the ways a psychotherapist perceives the patient’s level of engagement. Similar to psychotherapists, these modules also look for specific engagement-related cues exhibited by the patient in the video.

Table 2. Comparison of RMSE values obtained from various methods for estimating levels of engagement on the Multimodal Engagement Detection in Clinical Analysis dataset.

Method	RMSE[a] for engagement
Kaur et al [28]	0.96
Nezami et al [45]	0.17
Our approach	0.10

[a]RMSE: root mean square error.

Study 2: Ablation Studies

To show the importance of the different components (affective and cognitive) used in our approach, we run our method on MEDICA by removing either one of the modules corresponding to affective or cognitive states and report our findings. Table 3 summarizes the results obtained from the ablation experiments. We observe that the ablated model (ie, only using affective [A] or cognitive [C] modules) does not perform as well as the model that includes both modules. To understand and verify the contribution of these modules further, we leveraged the other labels (stress, hesitation, and attention) available in MEDICA and performed regression tasks using our proposed architecture on all of them. We observed that mode C performed better when predicting stress and hesitation values. Mode A performed better in estimating a patient’s level of attentiveness. These results agree with our understanding of cognitive state and affective state. Therefore, the combination of affective and cognitive state modes helps in efficiently predicting the engagement level of the patient.

Table 3. Ablation experiments on the Multimodal Engagement Detection in Clinical Analysis dataset. We ran our proposed model using only 1 modality at a time (either affective or cognitive) and compared the performance with that of using both modalities together.

Modality	RMSE[a] for engagement	RMSE for stress	RMSE for hesitation	RMSE for attention
Affective	0.24	0.15	0.146	0.07
Cognitive	0.3	0.13	0.16	0.08
Affective + cognitive (our approach)	0.10	0.12	0.14	0.1

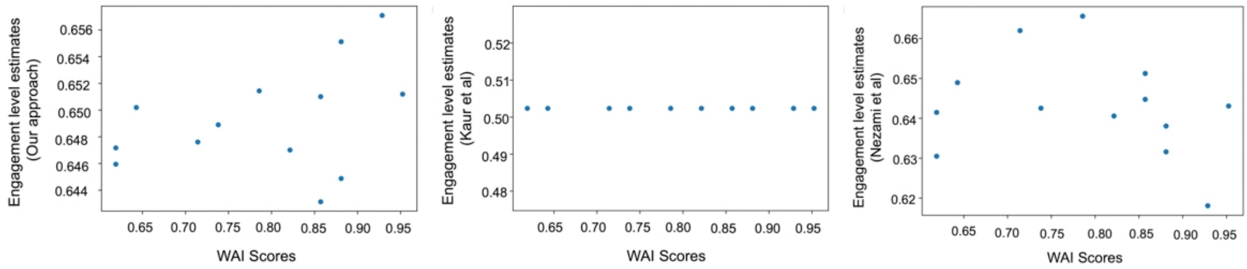
[a]RMSE: root mean square error.

Study 3: Analysis on Real-World Data

Our model that has been trained for estimating engagement levels using the MEDICA dataset was tested on the processed real-world data. WAI scoring is based on certain observations the therapist makes during the session with the patient. The score obtained from our model was different from the WAI score, but we claim that similar to WAI, our estimates also captured the engagement levels of the patient well. If this is indeed the case, then both WAI and our estimates should be correlated. As discussed earlier, a single WAI score is reported by the therapist (health care provider) for the entire session. Unlike WAI, which reports a single score per session, our method, along with those proposed by Kaur et al [28] and Nezami et al [45], performs microanalysis, generating engagement estimates at multiple time points within a session.

To enable a fair comparison, we calculated the mean of the engagement estimates across each session for all methods. The mean engagement estimates across sessions were as follows: Kaur et al [28] had a mean of 0.50 (SD: 7.82e-17), Nezami et al [45] had a mean of 0.64 (SD: 0.017), and our approach had a mean of 0.65 (SD: 0.006). We subsequently examined each of their correlations with the corresponding WAI scores across sessions. Given that WAI scores and engagement estimates are measured on different scales, WAI scores were normalized to a 0-1 range for consistency with the engagement values from the methods. We found that Kaur et al [28] obtained a correlation score of –0.03, Nezami et al [45] obtained a score of –0.24 and our approach obtained a much better score of 0.38. We plot the engagement level estimations obtained for each session against its corresponding WAI score, as shown in Figure 4. The plot corresponding to our approach shows a noticeably better alignment with the WAI score trends compared to other methods [28,45].

Figure 4. Scatterplot between the engagement level estimations obtained from each method for each session: (A) the plot between engagement level estimations obtained from our approach and the Working Alliance Inventory (WAI) scores, (B) the plot between the engagement level estimations obtained by Kaur et al [28] for each session and the WAI scores, (C) the plot between the engagement level estimations for each session obtained by Nezami et al [45] and the WAI scores.



Additionally, instead of just taking the mean, we also took the median of the engagement level estimates available at different instances of the sessions. The average median of the engagement estimates across sessions for different methods were as follows: Kaur et al [28] had an average median of 0.50 (IQR 0.0), Nezami et al [45] had an average median of 0.64 (IQR 0.02), and our approach had an average median of 0.65 (IQR 0.007).

We then examined the median of the engagement estimates for each session, obtained using different methods, and assessed their correlation with the WAI scores. In this case, Kaur et al [28] obtained a correlation score of 0, Nezami et al [45] obtained a score of –0.18 and our approach obtained a relatively high score of 0.40. These findings reinforce that our model captures WAI patterns more effectively than prior methods. Its stronger positive correlation with the WAI scores indicates its capability to better estimate engagement levels during therapy sessions.

The conceptual model of our proposed approach is also supported by the theoretical work by Bordin [32]. According to this theory, the therapist-provider alliance is driven by 3 factors: bond, agreement on goals, and agreement on tasks; these factors align well with the features identified in this work. While bond would correspond with the affective state, goals and task agreement correspond with the cognitive state. The merit of the approach by Bordin [32] is that it has been used for child and adult therapy, and it is one of the more widely studied therapeutic alliance measures. Therefore, it is no surprise that our approach can work well to provide an estimate of engagement levels in a tele–mental health session.

Discussion

Principal Findings

In this study, we developed a machine learning–based model that integrates features related to cognitive and affective psychological states to predict patient engagement levels during tele–mental health sessions. The input to our proposed algorithm is the visual, audio, and text data available, while the output is the engagement level of the patient. Our investigation demonstrated the significant promise of our proposed method, achieving an average improvement of 40% in RMSE for predicting engagement levels on the newly introduced MEDICA dataset compared to prior works. In real-world settings, our method provided patient engagement estimates that closely aligned with the trends observed in WAI scores assigned by therapists based on their sessions [30,31]. This study highlights the potential of our multimodal engagement estimation algorithm to enhance the quality of telehealth services and ultimately improve patient outcomes.

Comparison With Prior Work

Patient engagement is recognized as a pivotal determinant of a successful therapy session, particularly in the realm of mental health care [53]. Current psychiatric literature [54-56] has extensively examined a range of strategies and methodologies designed to foster [57] and sustain patient engagement during therapy sessions. These strategies are crucial for achieving optimal therapeutic outcomes. There has been a growing interest in applying artificial intelligence (AI) to mental health [58-62], with innovative approaches being explored to enhance patient care. However, despite the proliferation of AI-based methods in mental health applications, there remains a noticeable gap in research specifically focused on understanding and quantifying engagement. This gap is even more evident in the context of tele–mental health [63]. The transition from traditional in-person therapy to digital platforms has introduced unique challenges, as conventional indicators of engagement, such as body language [64], eye contact [65], and physical presence, maybe less discernible or entirely absent in web-based settings.

Our research addresses this critical gap by pioneering an approach that estimates patient engagement specifically in the tele–mental health context. This work is the first of its kind to target engagement estimation in this setting, offering a novel perspective that has largely been overlooked in previous studies. While the broader concept of engagement has been explored by AI in various domains, existing methods [28,45,66] have not been adapted or tested within the unique demands of mental health care, particularly in a remote or web-based environment. From an algorithmic standpoint, previous works have explored both unimodal and multimodal learning-based networks to understand engagement. Unimodal approaches focus on individual modalities, such as speech, body posture, biometric data (eg, heart rate variability and skin conductance) [67,68], gaze direction [69], and head pose [70]. Multimodal approaches [71] integrate various signals, leveraging the complementary strengths of each modality to provide a more holistic understanding of engagement. In contrast to these works, our research focuses on using the types of information that are readily available during a tele–mental health session, specifically audio, video, and text data. This approach is designed to be practical and easily implementable in real-world telehealth settings. This also ensures that the engagement estimation process is minimally disruptive to the therapeutic experience during the telehealth session. We validated our approach using our newly introduced dataset, MEDICA. In addition, we tested it on data collected from real-world therapy sessions. Our method achieved state-of-the-art performance in both cases.

Strengths and Implications

In this work, we explored the multidimensional and temporally dynamic nature of patient engagement [72]. Our method estimates engagement at regular short intervals. These characteristics of our approach align perfectly with the person-oriented analysis discussed by Sinatra et al [49]. We released a new dataset, MEDICA, to enhance mental health research, specifically toward understanding the engagement levels of patients attending therapy sessions. To the best of our knowledge, there is no other multimodal dataset that caters specifically to the needs of the mental health–based research. In addition, while there are some image-based [9,27,28] or sensor information–based datasets, there is no dataset that addresses the possibility of exploring engagement detection using visual, audio, and text modalities. We verified the usefulness of our machine learning method using 3 studies. Study 1 demonstrates the ability of our model to estimate the engagement level exhibited by the patient in video. It was done to understand the usefulness of the proposed model in comparison to other existing engagement estimation methods such as those by Nezami et al [45] and Kaur et al [28]. Our approach reported an RMSE [50-52] of 0.1, which on an average is 40% less than the RMSE reported by other methods (Table 1). Our study is one of the first to estimate engagement specifically for patients with mental health problems in a tele–mental health setting. The modalities used, specifically cognitive and affective states, propel the overall algorithm to mimic the way a psychotherapist perceives the patient's level of engagement. As part of study 2, we were interested to understand the contribution of the 2 modalities (cognitive and affective states). The experiments conducted as part of this study are also termed as *ablation* as it involves running the network pipeline using only 1 of the 2 modalities available at a time. We leveraged the annotations in MEDICA on stress, hesitation, and attention to dig deeper regarding the contribution of the cognitive and affective state modalities. We ran the same ablation experiments on not only patient engagement but also stress, hesitation, and attention estimation. While the combination of affective and cognitive states gives better results (RMSE of 0.1), it was interesting to note that the observations obtained supported the theoretical discussion about the 2 states. Cognitive state modality was able to understand stress (RMSE of 0.13) better. Affective state that was built on the concept of capturing distraction using emotion inconsistency between visual, audio, and text data helped relate better with the understanding of attention (RMSE of 0.07). These results agree with our understanding of cognitive state and affective state discussed earlier.

WAI is a popular measure of patient engagement used by many mental health care providers. Therefore, in study 3, we wished to understand and check if the values being estimated by our method related well with the trends observed in WAI. We did so by computing the correlation scores between WAI and the values estimated by our proposed model. In addition, to test the effectiveness of our approach, we also compared the correlation scores obtained between WAI and patient engagement values estimated by other existing methods. Positive values are preferred because they indicate similarity in distribution behavior. Negative values would indicate that the trends observed in WAI and the engagement values predicted are different and opposite. The results (Table 2) revealed that our model was able to capture the trends in WAI much better than other engagement estimation methods, reinforcing its clinical relevance. The correlation strength observed between WAI and our method's estimated engagement values is better than other approaches, but it also indicates the need for further investigation. The trend is positive, suggesting that we have successfully modeled engagement to some extent. Further analysis will provide opportunities to enhance and strengthen the relationship between WAI and the engagement values estimated by machine learning.

Limitations and Future Work

One of the primary limitations of the proposed algorithm is that engagement predictions may not be optimal in case of occlusions or missing modalities, data corruption due to low internet bandwidth, and integration of wearable devices with our model. We recognize that this situation is likely to occur in a tele–mental health session and plan to incorporate solutions for it in our future work. In addition, we aim to explore ways to make the predictions more explainable, allowing psychotherapists to receive evidence-based suggestions to support their final decisions.

Conclusions

Telehealth behavioral services, delivered via videoconferencing systems, have emerged as a cost-effective, dependable, and secure option for mental health treatment, particularly in recent times. However, gauging patient engagement, a critical mental health care standard, remains challenging in telehealth due to the absence of nonverbal cues and sensory data such as heart rate. To address this, we propose a novel multimodal semisupervised GAN that leverages affective and cognitive features from psychology to estimate engagement levels using visuals, audio, and text from video calls. This approach can significantly enhance social interactions and assist psychotherapists during tele–mental health sessions. Engagement is typically assessed through patient reports, which are susceptible to response bias, and current measures such as "show rate" and "patient satisfaction" do not accurately reflect the health care provider–patient alliance. Our model demonstrates its effectiveness on both our newly introduced dataset called MEDICA as well as real-world data. Given the lack of systematic engagement measurement and the limited training for health care providers on telehealth engagement, our proposed system offers a promising solution to these challenges, promoting better patient-health care provider interactions and making telehealth more effective.

Acknowledgments

This research was fully supported by the MPower grant.

Data Availability

The datasets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Demographic information on the real-world experiment participants.
[DOCX File , 14 KB-Multimedia Appendix 1]

References

1. WHO highlights urgent need to transform mental health and mental health care. World Health Organization. Jun 17, 2022. URL: https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care [accessed 2024-11-12]
2. The world health report 2001: mental health; new understanding, new hope. World Health Organization. 2001. URL: https://iris.who.int/handle/10665/42390 [accessed 2024-11-12]
3. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. Dec 15, 2012;380(9859):2197-2223. [doi: 10.1016/S0140-6736(12)61689-4] [Medline: 23245608]
4. Oladeji BD, Gureje O. Brain drain: a challenge to global mental health. BJPsych Int. Aug 2016;13(3):61-63. [FREE Full text] [doi: 10.1192/s2056474000001240] [Medline: 29093905]
5. Telemental health/providers. Anxiety & Depression Association of America. URL: https://adaa.org/find-help/treatment-help/telemental-health-providers [accessed 2024-11-12]
6. Greenwood H, Krzyzaniak N, Peiris R, Clark J, Scott AM, Cardona M, et al. Telehealth versus face-to-face psychotherapy for less common mental health conditions: systematic review and meta-analysis of randomized controlled trials. JMIR Ment Health. Mar 11, 2022;9(3):e31780. [FREE Full text] [doi: 10.2196/31780] [Medline: 35275081]
7. Kelders SM, van Zyl LE, Ludden GD. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. Front Psychol. 2020;11:926. [FREE Full text] [doi: 10.3389/fpsyg.2020.00926] [Medline: 32536888]
8. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med. Nov 2016;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
9. Whitehill J, Serpell Z, Lin YC, Foster A, Movellan JR. The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans Affective Comput. Jan 1, 2014;5(1):86-98. [doi: 10.1109/taffc.2014.2316163]
10. Murshed M, Dewan MA, Lin F, Wen D. Engagement detection in e-learning environments using convolutional neural networks. In: Proceedings of the IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). 2019. Presented at: DASC/PiCom/CBDCom/CyberSciTech 2019; August 5-8, 2019; Fukuoka, Japan. [doi: 10.1109/dasc/picom/cbdcom/cyberscitech.2019.00028]
11. Yu C, Aoki PM, Woodruff A. Detecting user engagement in everyday conversations. arXiv. Preprint posted online on October 13, 2004. [FREE Full text] [doi: 10.21437/interspeech.2004-327]

12. Sanghvi J, Castellano G, Leite I, Pereira A, McOwan PW, Paiva A. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: Proceedings of the 6th International Conference on Human-Robot Interaction. 2011. Presented at: HRI '11; March 6-9, 2011; Lausanne, Switzerland. [doi: 10.1145/1957656.1957781]
13. Nakano YI, Ishii R. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: Proceedings of the 15th International Conference on Intelligent User Interfaces. 2010. Presented at: IUI '10; February 7-10, 2010; Hong Kong, China. [doi: 10.1145/1719970.1719990]
14. Sharma P, Joshi S, Gautam S, Maharjan S, Khanal SR, Reis MC, et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. arXiv. Preprint posted online on September 18, 2019. [FREE Full text] [doi: 10.1007/978-3-031-22918-3_5]
15. Psaltis A, Apostolakis KC, Dimitropoulos K, Daras P. Multimodal student engagement recognition in prosocial games. IEEE Trans Games. Sep 2018;10(3):292-303. [doi: 10.1109/tciaig.2017.2743341]
16. Grafsgaard JF, Wiggins JB, Boyer KE, Wiebe EN, Lester LC. Embodied affect in tutorial dialogue: student gesture and posture. In: Proceedings of the 16th International Conference on Artificial Intelligence in Education. 2013. Presented at: AIED 2013; July 9-13, 2013; Memphis, TN. [doi: 10.1007/978-3-642-39112-5_1]
17. Aslan S, Cataltepe Z, Dundar O, Esme AA, Ferens R. Learner engagement measurement and classification in 1:1 learning. In: Proceedings of the 13th International Conference on Machine Learning and Applications. 2014. Presented at: ICMLA 2014; December 3-6, 2014; Detroit, MI. [doi: 10.1109/icmla.2014.111]
18. Ma J, Fruchter R. eRing: Body Motion Engagement Detection and Feedback in Global Teams. Stanford, CA. Stanford University; Apr 2015.
19. Monkaresi H, Bosch N, Calvo RA, D'Mello SK. Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans Affective Comput. Jan 1, 2017;8(1):15-28. [doi: 10.1109/taffc.2016.2515084]
20. Chang C, Zhang C, Chen L, Liu Y. An ensemble model using face and body tracking for engagement detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018. Presented at: ICMI '18; October 16-20, 2018; Boulder, CO. [doi: 10.1145/3242969.3264986]
21. Fedotov D, Perepelkina O, Kazimirova E, Konstantinova M, Minker W. Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data. In: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data. 2018. Presented at: MCPMD '18; October 16, 2018; Boulder, CO. [doi: 10.1145/3279810.3279842]
22. Zadeh A, Zellers R, Pincus E, Morency LP. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv. Preprint posted online on June 20, 2016. [FREE Full text]
23. Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. Presented at: ACL 2018; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1208]
24. Ong DC, Wu Z, Zhi-Xuan T, Reddan M, Kahhale I, Mattek A, et al. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. IEEE Trans Affect Comput. Jul 1, 2021;12(3):579-594. [FREE Full text] [doi: 10.1109/taffc.2019.2955949] [Medline: 34484569]
25. Ringeval F, Sonderegger A, Sauer J, Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2013. Presented at: FG 2013; April 22-26, 2013; Shanghai, China. [doi: 10.1109/fg.2013.6553805]
26. Zadeh A, Cao YS, Hessner S, Liang PP, Poria S, Morency LP. CMU-MOSEAS: a multimodal language dataset for Spanish, Portuguese, German and French. Proc Conf Empir Methods Nat Lang Process. Nov 2020;2020:1801-1812. [FREE Full text] [doi: 10.18653/v1/2020.emnlp-main.141] [Medline: 33969362]
27. Gupta A, D'Cunha A, Awasthi K, Balasubramanian V. DAiSEE: towards user engagement recognition in the wild. arXiv. Preprint posted online on September 7, 2016. [FREE Full text]
28. Kaur A, Mustafa A, Mehta L, Dhall A. Prediction and localization of student engagement in the wild. arXiv. Preprint posted online on April 3, 2018. [FREE Full text] [doi: 10.1109/dicta.2018.8615851]
29. Sathayanarayana S, Satzoda RK, Carini A, Lee M, Salamanca L, Reilly J. Towards automated understanding of student-tutor interactions using visual deictic gestures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014. Presented at: CVPRW 2014; June 23-28, 2014; Columbus, OH. [doi: 10.1109/cvprw.2014.77]
30. Berger T, Boettcher J, Caspar F. Internet-based guided self-help for several anxiety disorders: a randomized controlled trial comparing a tailored with a standardized disorder-specific approach. Psychotherapy (Chic). Jun 2014;51(2):207-219. [doi: 10.1037/a0032527] [Medline: 24041199]
31. Gómez Penedo JM, Berger T, Grosse Holtforth M, Krieger T, Schröder J, Hohagen F, et al. The Working Alliance Inventory for guided Internet interventions (WAI-I). J Clin Psychol. Jun 2020;76(6):973-986. [doi: 10.1002/jclp.22823] [Medline: 31240727]
32. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychother Theory Res Pract. 1979;16(3):252-260. [doi: 10.1037/h0085885]
33. Herrero R, Vara MD, Miragall M, Botella C, García-Palacios A, Riper H, et al. Working alliance inventory for online interventions-short form (WAI-TECH-SF): the role of the therapeutic alliance between patient and online program in

therapeutic outcomes. Int J Environ Res Public Health. Aug 25, 2020;17(17):6169. [FREE Full text] [doi: 10.3390/ijerph17176169] [Medline: 32854381]

34. Shinno S, Matsumoto K, Hamatani S, Inaba Y, Ozawa Y, Kawasaki Y, et al. Feasibility of guided internet-based cognitive behavioral therapy for panic disorder and social anxiety disorder in Japan: pilot single-arm trial. JMIR Form Res. Feb 29, 2024;8:e53659. [FREE Full text] [doi: 10.2196/53659] [Medline: 38421717]
35. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference. 2015. Presented at: SciPy 2015; July 6-12, 2015; Austin, TX. [doi: 10.25080/majora-7b98e3ed-003]
36. Magdin M, Sulka T, Tomanová J, Vozár M. Voice analysis using PRAAT software and classification of user emotional state. Int J Interact Multimed Artif Intell. 2019;5(6):33-42. [doi: 10.9781/ijimai.2019.03.004]
37. Balomenos T, Raouzaiou A, Ioannou S, Drosopoulos A, Karpouzis K, Kollias S. Emotion analysis in man-machine interaction systems. In: Proceedings of the First International Workshop on Machine Learning for Multimodal Interaction. 2004. Presented at: MLMI 2004; June 21-23, 2004; Martigny, Switzerland. [doi: 10.1007/978-3-540-30568-2_27]
38. Porter S, ten Brinke L. Reading between the lies: identifying concealed and falsified emotions in universal facial expressions. Psychol Sci. May 2008;19(5):508-514. [doi: 10.1111/j.1467-9280.2008.02116.x] [Medline: 18466413]
39. Arriaga O, Valdenegro-Toro M, Plöger P. Real-time convolutional neural networks for emotion and gender classification. arXiv. Preprint posted online on October 20, 2017
40. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. arXiv. Preprint posted online on May 1, 2020. [FREE Full text] [doi: 10.18653/v1/2020.acl-main.372]
41. Lu K, Zhao J, Xia M, Zeng J. Semi-supervised learning for image retrieval using support vector machines. In: Proceedings of the Second International Symposium on Neural Networks. 2005. Presented at: ISNN 2005; May 30-June 1, 2005; Chongqing, China. [doi: 10.1007/11427391_108]
42. Yu D, Varadarajan B, Deng L, Acero A. Active learning and semi-supervised learning for speech recognition: a unified framework using the global entropy reduction maximization criterion. Comput Speech Lang. Jul 2010;24(3):433-444. [doi: 10.1016/j.csl.2009.03.004]
43. Liu Y, Kirchhoff K. Graph-based semi-supervised learning for phone and segment classification. In: Proceedings of the INTERSPEECH 2013. 2013. Presented at: INTERSPEECH 2013; August 25-29, 2013; Lyon, France. URL: https://www.isca-archive.org/interspeech_2013/liu13e_interspeech.html# [doi: 10.21437/interspeech.2013-453]
44. Alyuz N, Okur E, Oktay E, Genc U, Aslan S, Mete SE, et al. Semi-supervised model personalization for improved detection of learner's emotional engagement. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016. Presented at: ICMI '16; November 12-16, 2016; Tokyo, Japan. [doi: 10.1145/2993148.2993166]
45. Nezami OM, Richards D, Hamey L. Semi-supervised detection of student engagement. In: Proceedings of the PACIS 2017. 2017. Presented at: PACIS 2017; July 16-20, 2017; Langkawi, Malaysia.
46. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. 2014. Presented at: NIPS'14; December 8-13, 2014; Montreal, QC.
47. Saeed AQ, Sheikh Abdullah SN, Che-Hamzah J, Abdul Ghani AT. Accuracy of using generative adversarial networks for glaucoma detection: systematic review and bibliometric analysis. J Med Internet Res. Sep 21, 2021;23(9):e27414. [FREE Full text] [doi: 10.2196/27414] [Medline: 34236992]
48. Olmschenk G, Zhu Z, Tang H. Generalizing semi-supervised generative adversarial networks to regression using feature contrasting. arXiv. Preprint posted online on on November 27, 2018. [FREE Full text]
49. Sinatra GM, Heddy BC, Lombardi D. The challenges of defining and measuring student engagement in science. Educ Psychol. Feb 27, 2015;50(1):1-13. [doi: 10.1080/00461520.2014.1002924]
50. Békés V, Aafjes-van Doorn K, Zilcha-Mano S, Prout T, Hoffman L. Psychotherapists' acceptance of telepsychotherapy during the COVID-19 pandemic: a machine learning approach. Clin Psychol Psychother. Nov 2021;28(6):1403-1415. [FREE Full text] [doi: 10.1002/cpp.2682] [Medline: 34723404]
51. Vail AK, Girard J, Bylsma L, Cohn J, Fournier J, Swartz H. Goals, tasks, and bonds: toward the computational assessment of therapist versus client perception of working alliance. In: Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition. 2021. Presented at: FG 2021; December 15-18, 2021; Jodhpur, India. [doi: 10.1109/fg52635.2021.9667021]
52. Tracey TJ, Kokotovic AM. Factor structure of the Working Alliance Inventory. Psychol Assess J Consult Clin Psychol. Sep 1989;1(3):207-210. [doi: 10.1037/1040-3590.1.3.207]
53. Ewalds Mulliez AP, Pomey MP, Bordeleau J, Desbiens F, Pelletier JF. A voice for the patients: evaluation of the implementation of a strategic organizational committee for patient engagement in mental health. PLoS One. Oct 24, 2018;13(10):e0205173. [FREE Full text] [doi: 10.1371/journal.pone.0205173] [Medline: 30356239]
54. Sagen JS, Smedslund G, Simonsen AE, Habberstad A, Kjeken I, Dagfinrud H, et al. Patient engagement in the development and delivery of healthcare services: a systematic scoping review. BMJ Open Qual. Jun 27, 2023;12(2):e002309. [FREE Full text] [doi: 10.1136/bmjoq-2023-002309] [Medline: 37369560]

55. Khosravi M, Azar G. Factors influencing patient engagement in mental health chatbots: a thematic analysis of findings from a systematic review of reviews. Digit Health. Apr 22, 2024;10:20552076241247983. [FREE Full text] [doi: 10.1177/20552076241247983] [Medline: 38655378]
56. Remmert JE, Destefano L, Chinman M, Oslin DW, Mavandadi S. Development of a peer-delivered primary care intervention to improve veteran mental health treatment engagement. Psychol Serv. Aug 01, 2024. [doi: 10.1037/ser0000883] [Medline: 39088007]
57. Jørgensen K, Lerbæk B, Frederiksen J, Watson E, Karlsson B. Patient participation in mental health care - perspectives of healthcare professionals and patients: a scoping review. Issues Ment Health Nurs. Aug 20, 2024;45(8):794-810. [doi: 10.1080/01612840.2024.2358931] [Medline: 38900284]
58. Elyoseph Z, Levkovich I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. JMIR Ment Health. Mar 18, 2024;11:e53043. [FREE Full text] [doi: 10.2196/53043] [Medline: 38533615]
59. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health. Jan 09, 2024;12(Suppl 1):e002583. [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
60. Thieme A, Hanratty M, Lyons M, Palacios J, Marques RF, Morrison C, et al. Designing human-centered AI for mental health: developing clinically relevant applications for online CBT treatment. ACM Trans Comput Hum Interact. Mar 17, 2023;30(2):1-50. [doi: 10.1145/3564752]
61. Rollwage M, Habicht J, Juechems K, Carrington B, Viswanathan S, Stylianou M, et al. Using conversational AI to facilitate mental health assessments and improve clinical efficiency within psychotherapy services: real-world observational study. JMIR AI. Dec 13, 2023;2:e44358. [FREE Full text] [doi: 10.2196/44358] [Medline: 38875569]
62. Roy K, Khandelwal V, Goswami R, Dolbir N, Malekar J, Sheth A. Demo alleviate: demonstrating artificial intelligence enabled virtual assistance for telehealth: the mental health case. Proc AAAI Conf Artif Intell. Jun 26, 2023;37(13):16479-16481. [doi: 10.1609/aaai.v37i13.27085]
63. Siegel A, Zuo Y, Moghaddamcharkari N, McIntyre RS, Rosenblat JD. Barriers, benefits and interventions for improving the delivery of telemental health services during the coronavirus disease 2019 pandemic: a systematic review. Curr Opin Psychiatry. Jul 01, 2021;34(4):434-443. [FREE Full text] [doi: 10.1097/YCO.0000000000000714] [Medline: 33928918]
64. Foley GN, Gentile JP. Nonverbal communication in psychotherapy. Psychiatry (Edgmont). Jun 2010;7(6):38-44. [FREE Full text] [Medline: 20622944]
65. Mitzkovitz C, Dowd SM, Cothran T, Musil S. The eyes have it: psychotherapy in the era of masks. J Clin Psychol Med Settings. Dec 03, 2022;29(4):886-897. [FREE Full text] [doi: 10.1007/s10880-022-09856-x] [Medline: 35118604]
66. Wang Y, DeVito Dabbs A, Thomas TH, Campbell G, Donovan H. Measuring engagement in provider-guided digital health interventions with a conceptual and analytical framework using nurse WRITE as an exemplar: exploratory study with an iterative approach. JMIR Form Res. Jul 22, 2024;8:e57529. [FREE Full text] [doi: 10.2196/57529] [Medline: 39037757]
67. Frageiro A, Debroize RP, Bannier E, Cury C. Eye-tracking and skin conductance to monitor task engagement during neurofeedback sessions. In: Proceedings of the 9th Graz Brain Computer Interface Conference. 2024. Presented at: BCI 2024; September 9-12, 2024; Graz, Austria.
68. Ventura RB, Porfiri M. Galvanic skin response as a measure of engagement during play in virtual reality. In: Proceedings of the ASME 2020 Dynamic Systems and Control Conference. 2020. Presented at: ASME 2020; October 5-7, 2020; Virtual Event. [doi: 10.1115/DSCC2020-3177]
69. Varma M, Washington P, Chrisman B, Kline A, Leblanc E, Paskov K, et al. Identification of social engagement indicators associated with autism spectrum disorder using a game-based mobile app: comparative study of gaze fixation and visual scanning methods. J Med Internet Res. Feb 15, 2022;24(2):e31830. [FREE Full text] [doi: 10.2196/31830] [Medline: 35166683]
70. Goldberg P, Sümer Ö, Stürmer K, Wagner W, Göllner R, Gerjets P, et al. Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction. Educ Psychol Rev. Dec 18, 2019;33:27-49. [doi: 10.1007/s10648-019-09514-z]
71. Booth BM, Bosch N, D'Mello SK. Engagement detection and its applications in learning: a tutorial and selective review. Proc IEEE. Oct 2023;111(10):1398-1422. [doi: 10.1109/jproc.2023.3309560]
72. Bernard A, de Ossorno Garcia S, Salhi L, John A, DelPozo-Banos M. Patterns of engagement in a digital mental health service during COVID-19: a cohort study for children and young people. Front Psychiatry. Jul 27, 2023;14:1143272. [FREE Full text] [doi: 10.3389/fpsyt.2023.1143272] [Medline: 37575580]

Abbreviations

AI: artificial intelligence
CMU-MOSEI: Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity
CMU-MOSI: Carnegie Mellon University Multimodal Corpus of Sentimental Intensity
GAN: generative adversarial net

HIPAA: Health Insurance Portability and Accountability Act
MEDICA: Multimodal Engagement Detection in Clinical Analysis
MFCC: mel-frequency cepstral coefficients
RMSE: root mean square error
WAI: Working Alliance Inventory
CREMA-D: Crowd-source Emotional Multimodal Actors Dataset
VGG-B: Visual Geometry Group-B
BERT: Bidirectional Encoder Representations from Transformers
LBP-TOP: local binary pattern on three orthogonal planes

Edited by A Mavragani; submitted 09.02.23; peer-reviewed by R Marshall, T Huang; comments to author 13.04.23; revised version received 30.06.23; accepted 03.09.24; published 20.01.25

Please cite as:
Guhan P, Awasthi N, McDonald K, Bussell K, Reeves G, Manocha D, Bera A
Developing a Machine Learning–Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study
JMIR Form Res 2025;9:e46390
URL: https://formative.jmir.org/2025/1/e46390
doi: 10.2196/46390
PMID: