MET: Multimodal Perception of Engagement for Telehealth

Pooja Guhan, 1 Naman Awasthi, 1 Kathryn McDonald, 2 Kristin Bussell, 2 Gloria Reeves, 2 Dinesh Manocha, 1 Aniket Bera, 1

1 University of Maryland, College Park, USA
2 University of Maryland, Baltimore County, USA
pguhan@umd.edu, nawasthi@umd.edu, KMcdonald@som.umaryland.edu, kristin.bussell@umaryland.edu,
greeves@som.umaryland.edu, dmanocha@umd.edu, bera@umd.edu

Abstract

We present MET, a learning-based algorithm for perceiving a human’s level of engagement from videos that give us access to only the face, speech and text. We leverage latent vectors corresponding to Affective and Cognitive features frequently used in psychology literature to understand a person’s level of engagement in a semi-supervised GAN-based framework. The method is extremely useful in the case of telehealth. We showcase the efficacy of this method from the perspective of mental health and more specifically how this can be leveraged for a better understanding of patient engagement during tele-mental health sessions. We also explore the usefulness of our framework and contrast it against existing works in being able to estimate another important mental health indicator, namely valence, and arousal. Our framework reports 40% improvements in RMSE over SOFA method in Engagement Regression and 50% improvements in RMSE over SOFA method in Valence-Arousal Regression. To tackle the scarcity of publicly available datasets in the telemental health space, we release a new dataset, MEDICA, for mental health patient engagement detection. Our dataset, MEDICA consists of 1299 videos, each 3 seconds long. To the best of our knowledge, our approach is the first method capable to model telemental health session data based on psychology-driven Affective and Cognitive features, which also accounts for data sparsity by leveraging a semi-supervised setup. To assert the usefulness of our method, we will also compare the association of the engagement values obtained from our model with the other engagement measures used by psychotherapists.

Introduction

As humans, we are able to very easily have effective communication and exhibit prosocial behavior while interacting with other humans, especially during an in-person meeting. This is because seeing one another allows us to pick up on nonverbal cues and body language. This also allows us to better understand each other, to recognize the other person’s point of view, and most importantly, their level of engagement. But we lack this connection in this digital age where video conferencing technology allows us to get face time with our colleagues even when we are not physically in the same room as them. Our inability to see beyond the person’s face takes away critical information like our estimate of their engagement levels that can allow us to establish better social interactions virtually. The disconnect we feel because of our uncertainty over the other person’s engagement levels is similar to our experiences while conversing with robots or chatbots (Oertel et al. 2020). These intelligent systems fail to have more natural human-like conversations because they lack the architectures to compute the engagement levels of the user by simply using the data from the sensors (mostly visual and audio) that they have access to. The situation is pretty much the same in the case of video conference calls as well. By engagement, we refer to the connection between the two interacting individuals that include a sense of basic trust and willingness/interest to collaborate. Apart from engagement, valence and arousal also have been considered very useful while evaluating the other person’s actions and intentions (Nummenmaa et al. 2012). Valence is basically positive affectivity, and arousal captures a person’s response to exciting information. Valence and arousal are two of the three dimensions of the Valence-Arousal-Dominance model (Mehrabian and Russell 1974). Therefore, our focus in this work is to develop a method that can help provide feedback on psychological indicators like engagement, and valence-arousal. This will not only set the stage for making human-robot or human-computer conversations more natural but will also most importantly improve human-human interactions held virtually. The latter is specifically critical because it can be extremely beneficial for a setting like that of telehealth and particularly in the case of telemental health (NAMI 2016, Llerena, Strauss, and Cohen 2012, Modecki, Murphy, and Waters 2020). Telemental health is the process of providing psychotherapy remotely, typically utilizing HIPAA-compliant video conferencing (ADA 2020). Given that it relies a lot on tech-
nology, even experienced human therapists face challenges sometimes engaging with patients due to unfamiliarity with the setup as well as other factors. For instance, in a telemental health session, the therapist has limited visual data (e.g., a therapist can only view the patient’s face rather than full body language), so the therapist has fewer non-verbal cues to guide their responses. It is also more difficult for the therapist to estimate attentiveness since eye contact required during in-person sessions is replaced with the patient looking at a camera or screen. Such challenges make it difficult for a therapist to perceive several of the patient’s mental health indicators like engagement level, valence, and arousal. Therefore, developing a system that can provide feedback on engagement, using multimodal data has the potential to improve the therapeutic outcomes while performing telemental health. Taking inspiration from existing psychology literature, we take a multi-componential approach and propose a framework that estimates these crucial indicators of social interactions in terms of the affective and cognitive states of the person.

1. **Cognitive state** involves comprehending complex concepts and issues and acquiring difficult skills. It conveys deep (rather than surface-level) processing of information whereby the person gains a critical or higher-order understanding of the subject matter and solves challenging problems. ([Posner 1980](#))

2. **Affective state** encompasses affective reactions such as excitement, boredom, curiosity, and anger. ([Li and Lerner 2013](#)). The range of affective expressions will vary based on individual demographic factors (e.g., age), cultural backgrounds/norms, and mental health symptoms.

Additionally, since the state of these psychological indicators is temporal in nature, we are interested in analyzing it across micro-level time scales ranging a few seconds. These characteristics of our approach align perfectly with the person-oriented analysis discussed by ([Sinatra, Heddy, and Lombardi 2015](#)).

**Main Contributions:** The novel components and main contributions of our work include:

1. We propose MET, a novel framework that takes into account two crucial psychological states, namely- Affective and Cognitive states. These are incorporated as the modalities in our multimodal framework.

2. We propose a novel regression-based framework that can capture psychology-inspired cues capable of perceiving the different important psychological indicators namely, engagement, valence, and arousal of a person. The input to the proposed framework would be the visual, audio, and text data available, while the output would be the desired psychological indicator (engagement/valence-arousal).

3. We show the usefulness of this method during telemental health sessions. We release a new dataset, MEDICA (Multimodal Engagement Detection In Clinical Analysis), to enhance mental health research, specifically towards understanding the engagement levels of patients attending the therapy sessions. To the best of our knowledge, there is no other multimodal dataset that caters specifically to the needs of mental health-based research. Additionally, while there are some image-based or sensory information-based datasets, there is no dataset that addresses the possibility of exploring engagement detection using visual, audio, and text modalities. MEDICA is a dataset that is a collection of around 1299 short video clips obtained from mock mental health therapy sessions conducted between an actor (who acts like a patient) and a real therapist, which is used by medical schools in their psychiatry curriculum.

We compare the performance of our proposed framework against prior methods on MEDICA and RECOLA. Using MEDICA, we try to understand the usefulness of our approach for engagement detection. We report RMSE of 0.10 on engagement detection task. We also tested our model on real-world data and report our findings. We train our framework separately on the RECOLA dataset for estimating Valence-Arousal in video clips. For this task, we report RMSE of 0.064 on valence estimation and 0.062 on arousal estimation. The fact that we are able to accomplish state-of-the-art results for both engagement and valence-arousal tasks using the same network (with slight variation in the output layer depending on the task) sets our method apart from other existing architectures.

**Related Work**

In this section, we give a brief overview of previous works on unimodal and multimodal engagement detection, semi-supervised learning, and also valence-arousal estimation.

**Unimodal / Multimodal Engagement Detection**

Prior works in engagement detection include unimodal as well as multimodal ([Inoue et al. 2019](#), [Monkaresi et al. 2016](#), [Navarathna et al. 2017](#), [Mittal et al. 2020b](#), c) based approaches. Some recent works have focused on detecting only affective cues ([Bhattacharya et al. 2019](#), [2020](#), Banerjee, Bhattacharya, and Bern [2020](#), Mittal et al. 2020[a]).

Facial expressions ([Whitehill et al. 2014](#), [Murshed et al. 2019](#)), speech ([Yu, Aoki, and Woodru 2004](#)), body posture ([Sanghvi et al. 2011](#)), gaze direction ([Nakano and Ishii 2010](#)) and head pose ([Sharma et al. 2019](#)) have been used as single modalities for detecting engagement. Combining different modalities has been observed to improve engagement detection accuracy ([Psaltis et al. 2017](#), [Grafsgaard et al. 2013](#), [Asian et al. 2014](#), [Frank et al. 2016](#)). [Monkaresi et al. 2016](#) proposed a multimodal framework to detect the level of engagement of participants during project meetings by leveraging features from facial expressions, voice and biometric data. [Monkaresi et al. 2016](#) 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. ([Chang et al. 2018](#), [Fedotov et al. 2018](#)) explored the use of body postures for detecting engagement.

**Semi-supervised Learning and GANs**

Recently, semi-supervised learning (SSL) has gained much importance as it has enabled us to deploy machine learning systems in real-life applications despite a lack of labeled data. It is the ability to improve classification in situations where we have few labeled data samples, and a lot of unlabeled data that has led it to be widely adopted in various applications like image search ([Lu et al. 2005](#)), speech analysis ([Yu et al. 2010](#)), ([Liu and Kirchhoff 2013](#)), natural language processing and emotion recognition ([Zhang et al. 2016](#)). There have also been some exploration in SSL to do engagement detection using facial ([Nezami, Richards, and Hamey 2017](#)) and affective ([Alyuz et al. 2016](#)).

Over the past few years, SSL in general has gradually been
combined with neural networks (Kipf and Welling 2016; Dalal 2019). (Kingma et al. 2014) proposed deep generative models for semi-supervised learning, based on variational autoencoders. Most state-of-the-art SSL methods using Generative Adversarial Nets (GANs) (Goodfellow et al. 2014) use the discriminator of the GAN as the classifier. They have been explored in the context of emotion recognition (Liang, Chen, and Jin 2019) but not engagement detection.

Valence-Arousal

Use in Healthcare: Valence and arousal values have other potential applications beyond patient engagement. This data may be used to improve early detection of patient safety issues (Llerena, Strauss, and Cohen 2012), including medication side effects (e.g., sedation), illicit substance use (Coffey et al. 2002), Garland and Howard 2021 [Liu and Bailey 2019], as well as risk of violence (self-harm, aggression) (Modecki, Murphy, and Waters 2020). The valence and arousal assessment can also be used in the assessment of important patient relationships, since behavioral health sessions may involve a parent, spouse, or other family members. For example, high levels of parental “expressed emotion” (e.g., criticism) may be associated with negative child behavioral outcomes for youth with autism spectrum disorder (Romero-Gonzalez, Chandler, and Simonoff 2018). Finally, objectively measured arousal and valence may be used as a tool to support diagnostic evaluations for disorders associated with altered arousal (e.g., ADHD, PTSD) (Bishop 2020) as well as conditions that have alterations in emotional valence (e.g., mood disorders).

Affective analysis: Due to the complexity of human interactions and the affective features they exhibit, it is difficult to encapsulate all variations into a set of discrete labels. Therefore, there has been increasing research interest in estimating the two important psychological dimensions - valence and arousal. While there exist some works that use valence-arousal values to make some kind of predictions (Zhao et al. 2020), there also have been works that describe methods to predict the values of valence and arousal individually (Ringeval et al. 2015; Parthasarathy and Busso 2016). These methods however ignore the inherent dependency that exists between the two dimensions (Oliveira et al. 2006) and therefore, a lot of crucial information could go missing. (Parthasarathy and Busso 2017) captures this dependency and exploits it by proposing a multi-tasking framework to predict the values of valence, arousal, and dominance. Multimodal approaches (Guo et al. 2019; Zhang et al. 2020) using visual-audio data or audio-linguistic data have also been proposed to predict the values of valence and arousal.

Our Approach

We provide an overview of our proposed framework in the following section.

Notation and Overview

We present an overview of the semi-supervised GAN multimodal engagement detection model in Fig. 2. Given an input of a video, audio, and text corresponding to a subject, the first objective is to extract useful psychology-derived features and then predict the different psychological indicators of the person under consideration. Affective state $h_A$ needs all three modes, i.e., video frames, audio, and text, while cognitive state $h_C$ extracts useful information from the patient’s speech. The concatenated vector $h_T = (h_C, h_A)$ is fed to the GAN network to perform semi-supervised learning-based regression.

Cognitive State Mode

The cognitive state is usually measured and evaluated using neuropsychological exams that are 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 take into account various essential perceptual indicators. Recently, there has been a lot of work around using speech as a potential biomarker for detecting cognitive decline (Fiore 2017; Themistocleous, Eckerstrom, and Kokkinakis 2018). 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 non-intrusive, inexpensive, and can potentially be real-time. Four major speech-based features namely - glottal($f_{g}$) (Ambrosini et al. 2019), phonation($f_{ph}$), articulation($f_{ar}$) and prosody($f_{pr}$) (Rektorova et al. 2016); have been found to be extremely useful to check for signs of cognitive impairment and are also being used a lot currently to detect early signs of extreme cognitive impairment conditions such as Parkinson’s and Alzheimer’s (Belalcázár-Bolaños et al. 2016; Arias-Vergara, Vásquez-Correa, and Orozco-Arroyave 2017). Therefore, the feature obtained from this model, $h_c$ can be written as: $h_C = concat(f_g, f_{ph}, f_{ar}, f_{pr})$.

Affective State Mode

In order to understand the affective state, we aim to check if there exists any inconsistency between the emotions perceived through what the person said, the tone with which the person expressed it, and the facial expressions that the person made. Often when a person is disengaged, the emotions perceived through the person’s facial expressions may not match the emotions perceived from the statement the person made. (Balomenios et al. 2004; Porter and Ten Brinke 2008) suggests 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. Therefore, motivated by this, we adopt pre-trained emotion recognition models to extract affective features from each video sample separately. Let $f_1, f_a, f_v$ correspond to the affective features obtained from the text (features from the caption of the video), audio (features of audio in the video), and video (video frame features) respectively. Therefore, $h_A = concat(f_1, f_a, f_v)$. 

![Figure 2: Overview: Here we present an overview of MET. We present a novel semi-supervised multi-modal GAN framework to detect engagement levels for telemental health based on psychology literature. TASK in the figure can be engagement detection or valence-arousal estimation](image-url)
GAN Framework

Different machine learning techniques can be explored to solve the engagement/valence-arousal problem. However, a large amount of high-quality labeled data is needed to train a robust model. For this reason, we propose a semi-supervised learning-based solution. We also wish to have a framework that can capture different possible variations in the variables that define engagement/valence-arousal. Therefore, the inclusion of semi-supervised GANs in the framework not only allows us to work with very few labeled data points but also in the process of trying to generate fake samples closer to the real sample distribution, the GAN generator manages to capture the data manifold well (Kumar, Sattigeri, and Fletcher 2017). This improves our model’s generalizability and makes it more robust compared to the previously defined semi-supervised learning approaches.

Datasets

MEDICA (Multimodal Engagement Detection In Clinical Analysis) Dataset

Engagement is a fairly overloaded term, and the definition varies with the application, making it hard 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 engagement in a person during a video call conversation. In such a setting, when two people are in a video call, the only data available to us readily is the person’s face and speech. There exists datasets like CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Zadeh et al. 2018), SEND (Ong et al. 2019) that capture such settings. However, they are not specifically for engagement detection. Given the lack of a dataset that can allow 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 telemedical health session videos. To use this data to address a broader range of issues related to mental health, we also include labels pertaining to stress and emotions. According to the author’s knowledge, this dataset is one of the first publicly available datasets that caters specifically to multimodal research in patient engagement in mental health. Table 1 presents a comparison between MEDICA and other related datasets.

1. Data Acquisition: The authors download publicly available mock sessions of mental health therapy videos. These videos are primarily used by clinical therapists to teach their students how to address patients’ grievances. The patients in the videos are being advised for depression, social anxiety, and PTSD. We have collected 13 videos, each having a duration of around 20 minutes - 30 mins. These videos involved conversations between a therapist and the patient, with each of them occurring in the frames either alone (i.e., only patient or only therapist) or together (rarely). Additionally, we also take only those videos that had the therapist dealing with only one patient and are conversing in English.

2. Data Processing and Annotation: Since our focus solely was to create a dataset that depicted the behavior of mental health patients during their sessions, we considered only parts of the videos where we had only the patient visible in the frames. The duration for which this happened was scattered across the video for different durations. We took these scattered clips and divided them into smaller clips of 3 seconds each. This results in a dataset of size 1229. We use Moviepy and speech-recognition libraries to extract audio and text from the video clips. Each video was annotated for attentiveness, stress, emotions displayed, engagement levels, and hesitation. Attentiveness, stress, and engagement were scored on a Likert scale of [-3,3], whereas hesitation is 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 multiple emotions. This is to motivate and provide the ability to the system to understand the various interacting emotions of the users. The data has also been annotated for 8 emotions related to mental health, namely - happy, sad, irritated, neutral, anxious, embarrassed, scared, and surprised. The annotation was carried out by a group of 20 psychotherapists. In order to avoid ambiguity in the interpretation of what “engagement” meant, annotators were provided with the definition we had agreed upon based on our conversations with the psychotherapists from our collaborating medical school and psychiatry literature for reference and clarity.

| Dataset Name | Samples | Unique speakers | Modes | Emotion | Engagement | Other mental health cues |
|--------------|---------|----------------|-------|---------|------------|-------------------------|
| MEDICA       | 1229    | 13             | [v,a,t]| ✓       | ✓          | hesitation, stress, attentiveness, physiological (electrocardiogram, and electrodermal activity) |
| RECOLA       | 3400    | 19             | [v,a] | ✓       | ✓          |                         |
| CMU-MOSEAS   | 715     | multiple       | [v,a,t]| ✓       | ✓          |                         |
| DASEE        | 9068    | 112            | v     | ✓       | ✓          |                         |
| HBUC        | 120     | 24             | v     | ✓       | ✓          |                         |
| in-the-wild | 195     | 78             | v     | ✓       | ✓          |                         |
| SIMATH       | 20      | 20             | [v,a,t]| ✓       | ✓          |                         |
|              |         |                |       |         |            |                         |

Table 1: Comparison of the MEDICA dataset with other related datasets. Modes indicates the subset of modalities present from (v)isual, (a)udio, (t)ext. *: Current status of the dataset. The size of the dataset will be increased.

Real-World Data

After having tested our method with proxy datasets like MEDICA, we were interested in testing our method in the real world. Therefore, we collaborated with around 8 psychotherapists to achieve this.

1. Data Collection and Processing: 20 patients voluntarily agreed to be a part of this research after a psychotherapist explained to them its purpose, what they can expect, their responsibilities, potential benefits, and risks they can face in the process. They were also given the assurance of confidentiality of the data being collected. Please refer to the supplementary material to get the demographic details of the participants. It was made clear that during the video storage process, we will “de-identify” any facial images besides the patient who is recorded so that they will not be recognizable on the recorded video. Efforts were made to limit their personal information, including the experiment evaluations and medical records, to people who have a need to review the information. On average, each of these sessions lasted around 20 mins. Now, we wish to test MET on the real-world data. Therefore, similar to MEDICA, the parts of the video where only the patient is visible and seen talking was identified. These
were then divided into 3-second clips. Audio and text were extracted similar to MEDICA.

2. **Scoring Mechanism(Annotation):** After a psychotherapy session is complete, the therapists score the collaborative relationship (therapeutic alliance) that was established between them and their clients during the session. The quality of this therapeutic alliance is measured using the working alliance inventory (WAI). WAI was modeled on Bordin’s (Bordin 1979) theoretical work. It captures 3 dimensions of the alliance (Bond, Tasks, and Goals). Extensive tests showcased 12 items per dimension to be the minimum length for the effective representation of the inventory. A composite score is computed based on these 12 items for each of the sessions conducted. Henceforth, this score will be referred to as the WAI score. We have collected the WAI scores reported by psychotherapists for each of these sessions. These scores are representative of the evaluation of the patient’s level of engagement.

Both MEDICA and the real-world data collection processes have been IRB approved by both the medical school as well as the computer science department.

**RECOLA Dataset**

RECOLA (Ringeval et al. 2013) (License: EULA) is a multimodal data corpus consisting of video recordings of spontaneous interaction happening between two people in French. The dataset has 23 videos in total, and each video shows one person who is visible conversing with another person (not visible in the video) in French. The audio of the person not visible in the video is inaudible. Each video has been annotated by six people (three males and three females).

Annotation Processing: The valence-arousal values were provided by annotators for every 0.04 second of the videos. Voice utterances do not have meaning in a short span like 0.04 seconds, and video frame information does not change in such a short span. Therefore, we remodeled the dataset by dividing the videos into clippings of 3 sec each and extracting the corresponding frames, audio, and text. We rescale the valence and arousal values to lie between 0 and 1 instead of -1 and 1 for simplicity. As there were six annotators, we weighed each annotator’s reported valence/arousal equally and averaged their six values to arrive at the final valence and final arousal values corresponding to every interval of 3 sec of the video. All the videos in the dataset have a duration of 5 mins, and valence, and arousal annotations have been provided every 0.04 second of the video. The net valence and arousal value for this duration is taken to be the mean of the valence and arousal values of the 75 sample points available in the original dataset for every second.

**Architecture and Implementation Details**

**Cognitive State Mode:**

In this mode, for the given input audio, we extract the glottal, prosody, articulation and phonation based features using librosa (McFee et al. 2015) and praat (Paszke et al. 2017) libraries.

Glottal features ($f_{gl}$) help in characterizing speech under stress (Cummings and Clements 1990). During periods of stress, there is an alternation in the amount of tension applied in the opening (abduction) and closing (Adduction) of the vocal cords (Moore et al. 2003).

Prosody features ($f_{pr}$) characterize the speaker’s intonation and speaking styles. Under this, we analyze variables like timing, intonation, and loudness during the production of speech.

Phonation ($f_{ph}$) in people with cognitive decline is characterized by bowing and inadequate closure of vocal cords, which produce problems in stability and periodicity of the vibration. They are analyzed in terms of features related to perturbation measures such as jitter (temporal perturbation of the fundamental frequency), shimmer (temporal perturbation of the amplitude of the signal), and amplitude perturbation quotient (APQ), and pitch perturbation quotient (PPQ).

Apart from these, the degree of unvoiced is also included. Articulation ($f_{ar}$) related issues in people with cognitive decline are mainly related to reduced amplitude and velocity of lip, tongue, and jaw movements. The analysis is based primarily on the computation of the first two vocal formants $F_1$ and $F_2$.

**Affective State Mode:**

In this mode, we extract affective features from audio, video and text data input.

1. Audio ($f_a$): Mel-frequency Cepstrum (MFCC) features were extracted from the audio clips available in RECOLA and MEDICA. The affective features were extracted using an MLP network that has been trained for emotion recognition in speech using the available in the RAVDESS (Livingstone and Russo 2018) and CREMA-D (Cao et al. 2014) datasets. A 150 length feature vector was obtained corresponding to each audio clip.

2. Video ($f_v$): The VGG-B architecture used in (Arriaga, Valdenegro-Toro, and Plöger 2017) was used to extract affective features from the video frames. The output dimensions of the second last layer were modified to give a feature vector of length 100.

3. Text ($f_t$): We extract affect features from the text using a bert-based model pre-trained network on GoEmotions dataset for MEDICA. For RECOLA dataset, we extract affect features from text using CamBERT model.

**Semi-supervised learning using GANs**

We use a multimodal semi-supervised GAN-based network architecture for regressing the values of an engagement or valence-arousal corresponding to each feature tuple $h_{pr}$. The network is similar to the semi-supervision framework SR-GAN proposed by (Olnischken, Zhu, and Tang 2018). But we develop a generator to model the feature maps generated by Cognitive and Affective state modules ($h_{pr}$). Additionally, the discriminator can output one (for an engagement task) or two (for a valence-arousal task) continuous values between 0 and 1.

Our model training pipeline is described in Fig. 2. The 5 losses used to train these components are:

1. **Labeled Loss:** Mean squared error of model output($\hat{y}_t$) with ground truth($y_t$).

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

2. **Unlabeled Loss:** Minimize the distance between the unlabeled and labeled dataset’s feature space.

$$L_{un} = \|E_{x \sim \text{p}^{\text{unlabeled}}} f(x) - E_{x \sim \text{p}^{\text{labeled}}} f(x) \|^2$$

3. **Fake loss:** Maximize the distance between unlabeled dataset’s features with respect to fake images.

$$L_{fake} = -||\log(E_{x \sim \text{p}^{\text{fake}}} f(x) - E_{x \sim \text{p}^{\text{unlabeled}}} f(x)) + 1)||_1$$
Generator Loss: Minimize the distance between the feature space of fake and unlabeled data.

\[ L_{gen} = \| \mathbb{E}_{x \sim p_{fake}} f(x) - \mathbb{E}_{x \sim p_{unlabeled}} f(x) \|_2^2 \] (4)

Gradient penalty: As described in (Olmschenk, Zhu, and Tang 2018), gradient penalty is used to keep the gradient of the discriminator in check which helps in convergence. The gradient penalty is calculated with respect to a randomly chosen point on the convex manifold connecting the unlabeled samples to the fake samples.

## Experiments and Results

Motivated by recent works in clinical psychotherapy (Békés et al. 2021), we use the standard evaluation metric of RMSE to evaluate all our approaches. The semi-supervised models were trained on NVIDIA GeForce GTX 1080ti GPU with batch size 512, learning rate 0.0001 for 50 epochs. We present two sets of experiments to understand the usefulness of our proposed method.

### Experiment-1: Engagement Detection

The first experiment demonstrates the ability of our model to predict score for the state of engagement exhibited by the person in the video. This experiment was performed on the MEDICA dataset.

#### Training Details

As our proposed methodology leverages a semi-supervised approach, we extract labeled samples from MEDICA and unlabeled samples from MOSEI dataset. After preprocessing, we extract 12854 unlabeled data points from MOSEI. We split the 1299 labeled data from MEDICA into 70:10:20 for training, validation, and testing respectively. Therefore, the split of labeled training data to unlabeled training data points is 909:12854. We compare our model with the following SOTA methods for engagement detection:

1. **Kaur, Amanjot, et al. (LBP-TOP)**: use a deep multiple instance learning-based framework for detecting engagement in students. They extract LBP-TOP features from the facial video segments and perform a linear regression using a DNN to estimate the engagement scores.

2. **Nezami, Omid Mohamad, et al. (SVM)**: perform a semi-supervised engagement detection using a semi-supervised support vector machine.

| Method         | RMSE for Engagement |
|----------------|----------------------|
| LBP-TOP        | 0.96                 |
| S3VM           | 0.17                 |
| MET            | 0.10                 |

Table 2: Comparisons on MEDICA Dataset

We use the publicly available implementation for LBP-TOP (Kaur et al. 2018) and train the entire model on MEDICA. S3VM (Nezami, Richards, and Hamey 2017) does not have a publicly available implementation. We reproduce the method to the best of our understanding.

### Results for Engagement Detection

Table 2 summarises the RMSE values obtained for all the methods described above. This is because different psychotherapy papers (Békés et al. 2021; Salinas-Oñate et al. 2020; Tseng et al. 2020) that have been published recently have used RMSE for their analysis. We observe an improvement of at least 40%. Our approach is one of the first to do engagement detection specifically for mental health patients in a telemental session setting. The modules used, specifically cognitive and affective states, help the overall framework to effectively mimic the way a psychotherapist perceives the patient’s level of engagement. Similar to a psychotherapist, these modules too look for specific engagement-related cues exhibited by the patient in the video.

### Experiment-2: Valence, Arousal Estimation

The second experiment demonstrates MET’s capabilities in estimating valence-arousal values of a person in a telemental health video. These experiments were performed on the RECOLA dataset.

#### Training Details

We split RECOLA dataset into train and test set with the ratio 90:10. We ensure that the two sets are mutually exclusive with respect to the participants appearing in each set. This prevents data leaks in our training/testing. The training set is further divided into labeled and unlabeled sets with a ratio of 40:60. We compare our model with the following SOTA methods for valence arousal estimation:

1. **Deng Didan et al. (Deng, Chen, and Shi 2020)** proposed a multitask CNN-RNN model for jointly learning three tasks, namely- Facial AU, expression classification and estimating valence, and arousal values. We use their publicly available implementation and train the entire model on RECOLA.

2. **Nguyen, Dung, et al. (Nguyen et al. 2021)** proposed a network consisting of a two-stream autoencoder and an LSTM for integrating visual and audio signals for estimating valence-arousal values. The authors report the results of their framework on RECOLA.

3. **Lee, Jiyoung, et al. (Lee et al. 2020)** extract color, depth and thermal information from videos and pass this as a multi-modal input to a spatiotemporal attention network to predict valence-arousal values. The authors report results of their framework on RECOLA.

#### Results for Valence-Arousal Estimation

Table 3 summarises the results obtained. We notice that our framework outperforms the SOTA methods discussed by almost 50%. Contrary to other methods, MET understands and effectively models the different facets of arousal and valence. While the affective state module relates to the kind of emotions experienced (extreme or normal or something in between), this aligns with the empirical and theoretical literature (Shuman, Sander, and Scherer 2013) that the affective module enables the system in understanding the degree of the emotion being experienced (extreme or normal or something in between). This is especially helpful in telemental health settings where the patient’s level of engagement is understood as a multifaceted nature of valence and arousal.

| Method         | Valence RMSE | Arousal RMSE |
|----------------|--------------|--------------|
| Nguyen, Dung, et al. (Nguyen et al. 2021) | 0.187 | 0.14        |
| Deng Didan et al. (Deng, Chen, and Shi 2020) | 0.114 | 0.089 |
| Lee, Jiyoung, et al. (Lee et al. 2020) | 0.192 | 0.084 |
| MET            | 0.0062       | 0.0062       |

Table 3: Comparison on RECOLA Dataset

#### Ablation Studies

To motivate the importance of the different components (Affective and Cognitive) used in our approach, we run MET on MEDICA and RECOLA by removing either one
of the modes corresponding to affective state or cognitive state and report our findings.

**Ablation study on Engagement Detection**

We summarize the ablation results performed in Table [2] We can observe that the ablated frameworks (i.e only using mode A or C) do not perform as well as MET. Therefore, in order to understand and verify the contribution of these modes further, we leveraged the other labels (stress, hesitation, and attention) available in MEDICA and performed regression tasks using MET on all of them. We observed that mode C performs better at 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.

**Ablation Study on Valence-Arousal Estimation**

It is interesting to observe from Table [5] that while valence has a greater dependency on mode A, arousal seems to be slightly more dependent on mode C. Both modes A and C individually and combined outperform the SOTA methods in Arousal detection. The combination of A and C help drastically decreasing the RMSE for valence.

**Analysis on Real World Data**

MET trained for engagement task 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 is different than WAI, but we claim that like WAI, our estimates also capture the engagement levels of the patient well. Therefore, if this is indeed the case, then both WAI and our estimates should be correlated. Now, the therapist reports one score for the entire session. But since our model performs microanalysis, we have engagement level estimates available in many instances during the session. Therefore, to make our comparison meaningful, we took the median of the estimates obtained from MET for each session. We then observed the correlation between WAI and MET median scores of the sessions. Fig. [3]1 shows the results obtained. Instead of the median, we also took the mean of the engagement level estimates available at different instances of the sessions and checked for their correlation with the WAI scores. Fig. [3]2 reports our findings for it. In both these figures, we have plotted the correlation matrices to understand the extent to which the engagement level estimates provided by our model correlate with the WAI scores. It is expected that the WAI scores will have the highest correlation with themselves. We can clearly observe this from figures [3]1 and [3]2. The high correlations between the MET scores and WAI scores support our claim about our model’s capability to effectively represent the WAI scores.

The conceptual model of MET is also supported by Bordin’s 1979 theoretical work ([Bordin 1979]). According to this theory, the therapist-provider alliance is driven by three factors - bond, agreement on goals, and agreement on tasks that fit nicely with the features identified in this work. While bond would correspond with affective, goals and task agreement correspond with cognitive. The merit of Bordin’s approach is that it has been used for child therapy and adults, and it is one of the more widely studied therapeutic alliance measures. Therefore, it is no surprise that our framework is able to work well to provide an estimate of engagement levels in a telemental health session.

**Conclusions, Limitations and Future Work**

We propose MET, a novel multimodal semi-supervised GAN framework that leverages affective and cognitive features from the psychology literature to estimate useful psychological state indicators like engagement and valence-arousal of a person. The method makes it possible to use the modalities easily available during a video call, namely, visuals, audio, and text to understand the audience, their reactions, and actions better. This can in turn help us have better social interactions. To the best of our knowledge, we are the first ones to do so. MET can be an incredible asset for therapists during telemental health sessions. The lack of non-verbal cues and sensory data like heart rate makes it very difficult for them to make an accurate assessment of engagement (a critical mental health indicator). The lack of datasets has always been a big challenge to use AI to solve this and other mental-health-related tasks. Therefore, to promote better research opportunities, we release a new dataset for engagement detection in mental health patients called MEDICA. We show our model’s usefulness on this as well as real-world data. As part of future work, we hope to build this dataset further to accommodate other related tasks apart from looking into possible kinds of variations arising due to cultural and geographical differences among patients and, therefore, making it more inclusive. Our work has some limitations and may not work well in case of occlusions, missing modality, and data corruptions due to low internet bandwidth. We plan to address this as part of future work. We would also like to explore making the predictions more explainable to enable psychotherapists to receive evidence-guided suggestions to make their final decisions.
References

ADAA. 2020. Telemental Health.

Alyuz, N.; Okur, E.; Oktay, E.; Genc, U.; Aslan, S.; Mete, S. E.; Arnrich, B.; and Esme, A. A. 2016. Semi-supervised model personalization for improved detection of learner's emotional engagement. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, 100–107.

Ambrosini, E.; Caielli, M.; Milis, M.; Loizou, C.; Azzolino, D.; Dumanti, S.; Bertagnoli, L.; Cesari, M.; Moccia, S.; Cid, M.; et al. 2019. Automatic speech analysis to early detect functional cognitive decline in elderly population. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 212–216. IEEE.

Arias-Vergara, T.; Vásquez-Correa, J. C.; and Orozco-Arroyave, J. R. 2017. Parkinson’s disease and aging: analysis of their effect in phonation and articulation of speech. Cognitive Computation, 9(6): 731–748.

Arriaga, O.; Valdenegro-Toro, M.; and Plöger, P. 2017. @articleadair2017attitude, title=Attitude-Scenario-Emotion (ASE) sentiments are superficial, author=Adair, Heather and Carruthers, Peter, journal=Behavioral and Brain Sciences, volume=40, year=2017, publisher=Cambridge University Press. arXiv preprint arXiv:1710.07557.

Aslan, S.; Cataltepe, Z.; Diner, I.; Dundar, O.; Esme, A. A.; Ferens, R.; Kamhi, G.; Oktay, E.; Soysal, C.; and Yener, M. 2014. Learner engagement measurement and classification in 1:1 learning. In 2014 13th International Conference on Machine Learning and Applications, 545–552. IEEE.

Balomenos, T.; Raouzaiaou, I.; Ioannou, S.; Drosopoulos, A.; Kar pouzis, K.; and Kollias, S. 2004. Emotion analysis in man-machine interaction systems. In International Workshop on Machine Learning for Multimodal Interaction, 318–328. Springer.

Banerjee, A.; Bhattacharya, U.; and Bera, A. 2020. Learning Unseen Emotions from Gestures via Semantically-Conditioned Zero-Shot Perception with Adversarial Autoencoders. arXiv preprint arXiv:2009.08906.

Békés, V.; Aafjes-van Doorn, K.; Zilcha-Mano, S.; Prout, T.; and Hoffman, L. 2021. Psychotherapists’ acceptance of telepsychotherapy during the COVID-19 pandemic: A machine learning approach. Clinical Psychology & Psychotherapy, 28(6): 1403–1415.

Belalázar-Bolaños, E. A.; Orozco-Arroyave, J. R.; Vargas-Bonilla, J. F.; Haderlein, T.; and Nöth, E. 2016. Glottal flow patterns analyses for Parkinson’s disease detection: acoustic and nonlinear approaches. In International Conference on Text, Speech, and Dialogue, 400–407. Springer.

Bhattacharya, U.; Rewkowski, N.; Guhan, P.; Williams, N. L.; Mittal, T.; Bera, A.; and Manocha, D. 2020. Generating Emotive Gaits for Virtual Agents Using Affect-Based Autoregression. ISMAR.

Bhattacharya, U.; Roncal, C.; Mittal, T.; Chandra, R.; Bera, A.; and Manocha, D. 2019. Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping.

Bishop, L. R. S. 2020. Fear of Feeling: How Fear of Experiencing Emotions Relates to Post-Trauma Functioning Among Female Survivors of Interpersonal Violence. Ph.D. thesis, Clark University.

Bordin, E. S. 1979. The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, research & practice, 16(3): 252.

Cao, H.; Cooper, D. G.; Keutmann, M. K.; Gur, R. C.; Nenkova, A.; and Verma, R. 2014. CREMA-D: Crowdsourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4): 377–390.

Chang, C.; Zhang, C.; Chen, L.; and Liu, Y. 2018. An ensemble model using face and body tracking for engagement detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, 616–622.

Coffey, S. F.; Saladin, M. E.; Drobes, D. J.; Brady, K. T.; Danksy, B. S.; and Kilpatrick, D. G. 2002. Trauma and substance cue reactivity in individuals with comorbid posttraumatic stress disorder and cocaine or alcohol dependence. Drug and alcohol dependence, 65(2): 115–127.

Cummings, K. E.; and Clements, M. A. 1990. Analysis of glottal waveforms across stress styles. In International Conference on Acoustics, Speech, and Signal Processing, 369–372. IEEE.

Da, S. 2019. A Generative Model for Semi-Supervised Learning.

Deng, D.; Chen, Z.; and Shi, B. E. 2020. Multitask emotion recognition with incomplete labels. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 592–599. IEEE.

Fedotov, D.; Perepelkina, O.; Kazimirova, E.; Konstantinova, M.; and Minker, W. 2018. 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, 1–9.

Fiore, K. 2017. Medpage.

Frank, M.; Tofighi, G.; Gu, H.; and Fruchter, R. 2016. Engagement detection in meetings. arXiv preprint arXiv:1608.08711.

Garland, E. L.; and Howard, M. O. 2021. Prescription opioid misusers exhibit blunted parasympathetic regulation during inhibitory control challenge. Psychopharmacology, 1–10.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances in neural information processing systems, 2672–2680.

Gould, D.; and Udry, E. 1994. Psychological skills for enhancing performance: Arousal regulation strategies. Medicine & Science in Sports & Exercise.

Grafsgaard, J. F.; Wiggins, J. B.; Boyer, K. E.; Wiebe, E. N.; and Lester, J. C. 2013. Embodied affect in tutorial dialogue: student gesture and posture. In International Conference on Artificial Intelligence in Education, 1–10. Springer.

Guo, J.; Song, B.; Zhang, P.; Ma, M.; Luo, W.; et al. 2019. Affective video content analysis based on multimodal data fusion in heterogeneous networks. Information Fusion, 51: 224–232.

Inoue, K.; Lala, D.; Takanashi, K.; and Kawahara, T. 2019. Latent character model for engagement recognition based on multimodal behaviors. In 9th International Workshop on Spoken Dialogue System Technology, 119–130. Springer.

Kaur, A.; Mustafa, A.; Mehta, L.; and Dhall, A. 2018. Prediction and localization of student engagement in the wild. In 2018 Digital Image Computing: Techniques and Applications (DICTA), 1–8. IEEE.
Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In Advances in neural information processing systems, 3581–3589.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Kumar, A.; Sattigeri, P.; and Fletcher, T. 2017. Semi-supervised learning with gans: Manifold invariance with improved inference. In Advances in Neural Information Processing Systems, 5534–5544.

Lee, J.; Kim, S.; Kim, S.; and Sohn, K. 2020. Multi-Modal recurrent attention networks for facial expression recognition. IEEE Transactions on Image Processing, 29: 6977–6991.

Li, Y.; and Lerner, R. M. 2013. Interrelations of behavioral, emotional, and cognitive school engagement in high school students. Journal of Youth and Adolescence, 42(1): 20–32.

Liang, J.; Chen, S.; and Jin, Q. 2019. Semi-supervised Multimodal Emotion Recognition with Improved Wasserstein GANs. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 695–703. IEEE.

Liu, J.; and Bailey, R. L. 2019. Effects of substance cues in negative public service announcements on cognitive processing. Health communication, 34(9): 964–974.

Liu, Y.; and Kirchhoff, K. 2013. Graph-based semi-supervised learning for phone and segment classification. In INTERSPEECH, 1840–1843.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5): e0196991.

Llerena, K.; Strauss, G. P.; and Cohen, A. S. 2012. Looking at the other side of the coin: a meta-analysis of self-reported emotional arousal in people with schizophrenia. Schizophrenia research, 142(1-3): 65–70.

Lu, K.; Zhao, J.; Xia, M.; and Zeng, J. 2005. Semi-supervised learning for image retrieval using support vector machines. In International Symposium on Neural Networks, 677–681. Springer.

Ma, J.; and Fruchter, R. 2015. eRing: Body Motion Engagement Detection and Feedback in Global Teams. In SAVI Symposium on New ways to teach and learn for student engagement, Stanford University.

MacHardy, Z. M.; Syharath, K.; and Dewan, P. 2012. Engagement analysis through computer vision. In 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 535–539. IEEE.

Madan, C. R.; Scott, S. M.; and Kensinger, E. A. 2019. Positive emotion enhances association-memory. Emotion, 19(4): 733.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, volume 8, 18–25.

Mehrabian, A.; and Russell, J. A. 1974. An approach to environmental psychology. the MIT Press.

Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020a. Emotions Don’t Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. In Proceedings of the 28th ACM International Conference on Multimedia, 2823–2832.

Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020b. MSER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. In AAAI, 1359–1367.

Mittal, T.; Guhan, P.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020c. EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege’s Principle. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14234–14243.

Modecki, K. L.; Murphy, L. K.; and Waters, A. M. 2020. Exposure to violence and neglect images differentially influences fear learning and extinction. Biological psychology, 151: 107832.

Monkaresi, H.; Bosch, N.; Calvo, R. A.; and D’Mello, S. K. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Transactions on Affective Computing, 8(1): 15–28.

Moore, E.; et al. 2003. Evaluating objective feature statistics of speech as indicators of vocal affect and depression. Ph.D. thesis, Georgia Institute of Technology.

Mursheed, M.; Dewan, M. A. A.; Lin, F.; and Wen, D. 2019. Engagement Detection in e-Learning Environments using Convolutional Neural Networks. In 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 80–86. IEEE.

Nakano, Y. I.; and Ishii, R. 2010. Estimating user’s engagement from eye-gaze behaviors in human-agent conversations. In Proceedings of the 15th international conference on Intelligent user interfaces, 139–148.

NAMI. 2016. Engagement: A New Standard for Mental Health Care.

Navarathna, R.; Carr, P.; Lucey, P.; and Matthews, I. 2017. Estimating audience engagement to predict movie ratings. IEEE Transactions on Affective Computing, 10(1): 48–59.

Nezami, O. M.; Richards, D.; and Hamey, L. 2017. Semi-supervised detection of student engagement. In 21st Pacific-Asia Conference on Information Systems, PACIS 2017, 2–8. AIS Electronic Library (AISeL).

Nguyen, D.; Nguyen, D. T.; Zeng, R.; Nguyen, T. T.; Tran, S.; Nguyen, T. K.; Sridharan, S.; and Fookes, C. 2021. Deep Auto-Encoders with Sequential Learning for Multimodal Dimensional Emotion Recognition. IEEE Transactions on Multimedia.

Nummenmaa, L.; Glerene, E.; Viinikainen, M.; Jääskeläinen, I. P.; Hari, R.; and Sams, M. 2012. Emotions promote social interaction by synchronizing brain activity across individuals. Proceedings of the National Academy of Sciences, 109(24): 9599–9604.

Oertel, C.; Castellano, G.; Chetouani, M.; Nasir, J.; Obaid, M.; Pelachaud, C.; and Peters, C. 2020. Engagement in human-agent interaction: An overview. Frontiers in Robotics and AI, 7: 92.

Oliveira, A. M.; Teixeira, M. P.; Fonseca, I. B.; and Oliveira, M. 2006. JOINT MODEL-PARAMETER VALIDATION
OF SELF-ESTIMATES OF VALENCE AND AROUSAL: PROBING A DIFFERENTIAL-WEIGHTING MODEL OF AFFECTIVE INTENSITY. Proceedings of Fechner Day, 22: 245–250.

Olmschenk, G.; Zhu, Z.; and Tang, H. 2018. Generalizing semi-supervised generative adversarial networks to regression. CoRR, abs/1811.11269.

Ong, D.; Wu, Z.; Tan, Z.-X.; Reddan, M.; Kahhale, I.; Mattak, A.; and Zakr, J. 2019. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. IEEE Transactions on Affective Computing.

Parthasarathy, S.; and Busso, C. 2016. Defining Emotionally Salient Regions Using Qualitative Agreement Method. In Interspeech, 3598–3602.

Parthasarathy, S.; and Busso, C. 2017. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In Interspeech, volume 2017, 1103–1107.

Parthasarathy, S.; and Busso, C. 2019. Semi-supervised speech emotion recognition with ladder networks. arXiv preprint arXiv:1905.02921.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Porter, S.; and Ten Brinke, L. 2008. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. Psychological science, 19(5): 508–514.

Posner, M. I. 1980. Orienting of attention. Quarterly journal of experimental psychology, 32(1): 3–25.

Psaltis, A.; Apostolakis, K. C.; Dimitropoulos, K.; and Daras, P. 2017. Multimodal student engagement recognition in prosocial games. IEEE Transactions on Games, 10(3): 292–303.

Randhavane, T.; Bhattacharya, U.; Kapsakis, K.; Gray, K.; Bera, A.; and Manocha, D. 2019. Identifying emotions from walking using affective and deep features. arXiv preprint arXiv:1906.11884.

Rektorova, I.; Mekyska, J.; Janousova, E.; Kostalova, M.; Eliasova, I.; Mrackova, M.; Berankova, D.; Necasova, T.; Smekal, Z.; and Marecek, R. 2016. Speech prosody impairment predicts cognitive decline in Parkinson’s disease. Parkinsonism & related disorders, 29: 90–95.

Ringeval, F.; Eyben, F.; Kroupi, E.; Yuce, A.; Thiran, J.-P.; Ebrahimi, T.; Lalanne, D.; and Schuller, B. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters, 66: 22–30.

Ringeval, F.; Sonderegger, A.; Sauer, J.; and Lalanne, D. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), 1–8. IEEE.

Romero-Gonzalez, M.; Chandler, S.; and Simonoff, E. 2018. The relationship of parental expressed emotion to co-occurring psychopathology in individuals with autism spectrum disorder: A systematic review. Research in developmental disabilities, 72: 152–165.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In Advances in neural information processing systems, 2234–2242.

Salinas-Oñate, N.; Baeza-Rivera, M. J.; Ortiz, M.; and Betancourt, H. 2020. Development and validation of a scale for measuring cultural beliefs about psychotherapy patients in southern Chile. Psicologia: reflexão e crítica, 33.

Saloni Dattani, H. R.; and Roser, M. 2021. Mental Health.

Sanhvi, J.; Castellano, G.; Leite, I.; Pereira, A.; McOwan, P. W.; and Paiva, A. 2011. 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, 305–312.

Sharma, P.; Joshi, S.; Gautam, S.; Filipe, V.; and Reis, M. J. 2019. Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning. arXiv preprint arXiv:1909.12913.

Shi, W.; Gong, Y.; Ding, C.; MaXiaoyu. Tao, Z.; and Zheng, N. 2018. Transductive semi-supervised deep learning using min-max features. In Proceedings of the European Conference on Computer Vision (ECCV), 299–315.

Shuman, V.; Sander, D.; and Scherer, K. R. 2013. Levels of valence. Frontiers in psychology, 4: 261.

Sinatra, G. M.; Hedly, B. C.; and Lombardi, D. 2015. The challenges of defining and measuring student engagement in science.

Themistocles, C.; Eckerstrom, M.; and Kokkinakis, D. 2018. Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. Frontiers in neurology, 9: 975.

Tseng, V. W.-S.; Sano, A.; Ben-Zeev, D.; Brian, R.; Campbell, A. T.; Hauser, M.; Kane, J. M.; Scherer, E. A.; Wang, R.; Wang, W.; et al. 2020. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. Scientific reports, 10(1): 1–17.

Whitehill, J.; Serpell, Z.; Lin, Y.-C.; Foster, A.; and Movellan, J. R. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing, 5(1): 86–98.

WHO. 2020. WHO urges more investments, services for mental health.

Wu, Y.; Mu, G.; Qin, C.; Miao, Q.; Ma, W.; and Zhang, X. 2020. Semi-Supervised Hyperspectral Image Classification via Spatial-Regulated Self-Training. Remote Sensing, 12(1): 159.

Yu, C.; Aoki, P. M.; and Woodruff, A. 2004. Detecting user engagement in everyday conversations. arXiv preprint cs/0410027.

Yu, D.; Varadarajan, B.; Deng, L.; and Acero, A. 2010. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. Computer Speech & Language, 24(3): 433–444.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. 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 (Volume 1: Long Papers), 2236–2246.
Telemental Health

The World Health Organization defines mental health as “a state of well-being” that allows a person to lead a ful-filling and productive life and contribute to society (WHO 2020). It has been estimated that around 15.5% of the population suffers from mental illness globally, and these numbers are rising continuously. There is, however, a worldwide shortage of mental health providers. This, combined with issues related to affordability and reachability, has resulted in more than 50% of mental health patients remaining untreated. The mental health landscape became even bleaker during the COVID-19 pandemic. However, the rapid expansion of telemental health services, especially during the pandemic, has increased access to clinical care options and introduced the opportunity to use artificial intelligence (AI) based strategies to improve the quality of human-delivered mental health services. Availability of only visual, audio and text type of data during a telemental health session makes it difficult for a therapist to perceive the patient’s mental health indicators like engagement, engagement-arousal levels. Engagement is considered one of the key standards for mental health care. It is critical for both retention in care as well as the accuracy of the diagnosis. Valence and arousal values have also been shown to be extremely useful for a therapist to understand and improve early detection of patient safety issues (Llerena, Strauss, and Coifert 2012), including medication side effects (e.g., sedation), illicit substance use (Coffey et al. 2002), medication adherence (Garland and Howard 2021), Liu and Bailey (2019), as well as the risk of violence (self-harm aggression) (Modelli, Murphy, and Waters 2020). Therefore, developing a system that can provide feedback on engagement, and valence-arousal, using multimodal data, has the potential to improve the therapeutic outcomes while performing telemental health. The components of psychological states used by us in MET also happen to be the categories of the different cues used by a mental health therapist to assess someone’s engagement, valence-arousal levels. One of the biggest motivations for us to incorporate these states (cognitive and affective) as modules of the framework is not only to lay the foundations for a trustworthy analysis from the perspective of a clinician but also open a possibility for the user to understand the reasons behind a specific assessment.

Some More Related Work

Unimodal/Multimodal Engagement Detection (Frank et al. 2016) proposed a multimodal framework to detect the level of engagement of participants during project meetings in a work environment. The authors expanded the work of Stanford’s PBL Labs called eRing (Ma and Fruchter 2015) by including information streams such as facial expressions, voice, and other biometric data. (Monkaresi et al. 2016) 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 annotations for classification purposes. (Chang et al. 2018) make use of facial expressions as well as body posture for detecting engagement in learners. (Fedotov et al. 2018) proposes the use of audio, facial, and body features to detect engagement and disengagement for an imbalanced in-the-wild dataset.

Semi-Supervised Learning

(Zhang et al. 2016) proposed a novel multimodal SSL architecture to detect emotions on RECOLA dataset using audio and video-based modalities. The authors also describe a method to handle mislabeled data. In order to perform emotion recognition in speech, (Parthasarathy and Busso 2019) proposes training ladder networks in a semi-supervised fashion. There also has been some exploration in SSL to do engagement detection. One of the earliest works in this direction includes the works of (Alyuz et al. 2016) where they consider the development of an engagement detection system, more specifically emotional or affective engagement of the student in a semi-supervised fashion to personalize systems like Intelligent Tutoring Systems according to their needs. (Nezami, Richards, and Hamey 2017) conduct experiments to detect user engagement using facial feature-based semi-supervised model.

Semi-Supervised Learning with GANs

Early semi-supervised learning methods include self-training (Wu et al. 2020), transductive learning (Shi et al. 2018), graph-based models, and other learning methods. Most state-of-the-art semi-supervised learning methods using Generative Adversarial Nets (GANs) (Goodfellow et al. 2014) use the discriminator of the GAN as the classifier. The earliest works in the application of GANs for semi-supervised learning include (Salimans et al. 2016) which presented a variety of new architectural features and training procedures such as feature matching and minibatch discrimination techniques to encourage convergence of GANs. (Liang, Chen, and Jin 2019) proposed an SSL based Wasserstein GAN to perform multimodal emotion recognition using separate generators and discriminators for each of the modalities being explored namely-audio and visual.