Multimodal Framework for Therapeutic Consultations

Martin Ivanov[1,2], Alice Rueda[1,2] *Member, IEEE*, Venkat Bhat[2,3*], Sridhar Krishnan[1*] *Senior Member, IEEE*

[1]Signal Analysis Research (SAR) Group, Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, Canada M5B 2K3

[2] Interventional Psychiatry Program, St. Michael's Hospital, Toronto, Canada M5B 1W8

[3] Department of Psychiatry, University of Toronto, Toronto, Canada M5S 1A1

Abstract—Therapeutic engagement between client and clinician is a key indicator in determining treatment outcomes for clients with mental health disorders. Quantifying this type of engagement provides an opportunity for the development of an engagement quantification framework for therapeutic efficacy, based on a number of data streams including, body movement and synchronicity, speech, and gestures to determine an individual's level of engagement. In this paper, we present a subset of such a framework through the quantification of engagement based on Facial Affect Recognition, Head Motion, and Natural Language Processing. We propose the use of semantic analysis, emotion dynamics and transitions, and head motion to describe a participant's attention over the consultation. For emotion dynamics and transitions we employ seven standard categorical emotions; for head motion we use acute and chronic head movement; and for semantic analysis we employ Robustly Optimized BERT Pretraining Approach. These features derive two engagement levels: low and high. We performed experiments on the AnnoMI dataset, which contains 133 therapeutic consultation videos for low and high quality motivational interviews, and compared the resulting engagement to the level of motivational interviewing. We achieved an 89.1% average accuracy for the Clinician model and an 81.1% average accuracy for the Client model using Gradient Boost as a classifier.

Index Terms—engagement quantification, virtual therapy, facial affect recognition, therapeutic alliance, machine learning, computer vision, mental health

1 INTRODUCTION

In recent years a rapid adoption of telehealth services due to the pandemic has resulted in a transition from "in-person" to virtual care [1]. This shift has brought into question the efficacy of therapeutic treatment conducted virtually, especially in mental health care where participants have experienced the inability to generate meaningful engagement [2], [3]. In cases of depression, which is a classical mental disorder, pharmacotherapy and psychotherapy literature suggest a critical role for the therapeutic relationship in treatment outcomes [4]. Pharmacotherapy and psychotherapy literature suggest the critical role for the therapeutic relationship in treatment outcomes [4]. The therapeutic relationship is an important factor in fostering therapeutic engagement, as clients often attribute positive outcomes to the personal attributes of their therapist/clinician [4].

Therapeutic engagement is defined as "the extent to which the client actively participates in the treatment on offer" [5]. Medical literature from "in-person care" suggests that therapeutic engagement is a key factor in determining successful treatment outcomes from clinical care [5]. This is especially true for individuals suffering from severe, long-term mental health problems [6]. However, it is unclear how much telehealth services have affected the level of engagement in therapeutics. Virtual health places a higher level of responsibility on the client in terms of preparation for visits [7], but there exists no objective engagement measurement [6] and no consensus exists regarding the factors that determine engagement [5]. Having an accurate objective measure of engagement would allow for monitoring of this vital component and potential improvement in therapeutic outcomes.

Virtual health has unique advantages for collecting digital data during routine clinical care, which could enable the development of an Automated Engagement Scorecard (AES) system. An engagement scorecard is a tool for quantifying the engagement level based on a set of defined metrics. These metrics consist of quantification of social interaction such as eye contact, participation in a conversation, or attention during a conversation. AES is the automation of engagement quantification using computer vision and machine learning (ML) algorithms on three data streams (eye movement, speech, and gestures) to determine an individual's level of engagement. In this work, we describe a subset of the AES framework for engagement quantification based on a multimodal method through facial affect recognition (FAR) and natural language processing (NLP) using the AnnoMI datasets. FAR is the task of utilizing facial data to efficiently identify and locate human faces regardless of position, scale, orientation, pose or illumination, and to determine the facial expression/emotions [8]. In this way

The authors would like to thank Natural Sciences and Engineering Research Council of Canada (NSERC), New Frontiers in Research Fund, and LeaCros for the funding.

* *Senior co-authors*

Manuscript received TBD; revised TBD

affects and emotions can be thought of as interchangeable terms. There are two standard approaches to FAR, classification through facial landmarks or through facial action units (AU), we focus on the latter. AUs are the key components of Facial Action Coding Systems which are used to derive facial expressions. AUs correspond to specific groups of muscle that define facial motions such as cheek raiser (AU6) or lip corner puller (AU12) [9].

Engagement quantification is a multifaceted research area extensively documented across various domains, including education [10], [11], [12], [13], [14], [15], business [16], [17], social robotics [18], [19], gaming [20], healthcare [21], [22], [23], digital health [24], [25], and most recently digital therapeutics [26], [27], [28], [29], [30]. Engagement quantification is heavily grounded in advancements within affective computing, which recent bibliometric analyses of over 33 000 studies (1997–2025) show is evolving along three main axes. First, the creation and standardization of large-scale, multimodal datasets—including text, speech, facial video, and physiological signals such as EEG, ECG, and EDA has enabled the development of data-hungry deep learning models with finer-grained emotion labels. Second, researchers are moving beyond unimodal CNN or SVM-based pipelines toward hybrid, knowledge-driven architectures that fuse behavioral cues (e.g., facial expressions, language) with physiological measures via attention-based transformers, adversarial networks, or self-supervised encoders informed by psychological theory and knowledge graphs. Third, there is a clear shift toward real-world deployment: affective systems are being embedded in clinical settings for depression monitoring, autism screening, and pain assessment, as well as in human–robot dialogue, brain–computer interfaces, and virtual reality environments that require context-aware, empathic interactions [31], [32], [33]. Behera *et al.* used facial and head data to derive AU based facial expression, hand-over-face gestures, and chronic and acute head motion in order to classify engagement in educational tasks [34]. They determined that head movements, as well as hand-over-face gestures increased with complexity of educational material [34].Bhardwaj *et al.* utilized facial landmarks, eye tracking, emotion detection and deep neural networks to quantify engagement in 950 students using video from the "Google Meet" platform and achieved 93.6% accuracy [35]. Islam and Bae introduced FacePsy, an open-source mobile sensing system that captures facial expressions and head gestures via smartphones in natural environments to detect depression. By opportunistically analyzing key affective features—including eye-open states, smile expressions, and specific Action Units—the system achieved an AUROC of 81% for depressive episode detection and predicted PHQ-9 scores with a mean absolute error of 3.08 [36]. Guhan *et al.* developed a machine learning-based framework called Multimodal Perception of Engagement for Telehealth (MET) to estimate patient engagement in telehealth sessions. Utilizing affective and cognitive features from psychology literature, they trained a semi-supervised GAN model on the MEDICA dataset. The results showed a 40% improvement in RMSE for engagement estimation compared to state-of-the-art methods, with positive correlations observed between MET's estimates and psychotherapists' working alliance inventory scores. This suggests that the MET framework can provide accurate engagement feedback to therapists, enhancing diagnosis and treatment during telehealth sessions [27].Rodriguez *et al.* utilized a gradient-boosted forest model to predict user engagement in various activities within a digital diabetes prevention program (dDPP), developing daily and weekly models based on short- and long-term activity data. The daily model achieved over 90% predictive accuracy, surpassing the weekly model in fitting the research plan due to its ability to predict daily changes in user activities [37]. A closely related field to engagement quantification is empathy prediction, a highly studied field of research seeking to quantify the quality of social interactions through assessing text, audiovisual, audio and physiological signals. This field seeks to measure the levels of empathy expression from all participants in a conversation in order to indicate how well those participants are understanding each other [29]. Zhu *et al.* developed the MEDIC dataset, the first multimodal dataset for empathy in psychotherapy, incorporating visual, audio, and textual data. Using labels for client expression of experience and counselor emotional and cognitive reactions, they ensured high interrater reliability through Fleiss' Kappa and intraclass correlation [38]. Experiments with models like Tensor Fusion Network (TFN) and Sentimental Words Aware Fusion Network (SWAFN) demonstrated that multimodal fusion significantly enhanced empathy prediction, achieving F1 scores above 69% [38]. The SWAFN model with an LSTM classifier achieved the highest accuracy of 86.4% for predicting client expressions [38].*et al.* developed a multimodal framework to assess the teaching quality of one-to-many online instruction videos across five dimensions: clarity, classroom interaction, technical management, empathy, and time management [39]. The system combines mid-level behavior descriptors (e.g., facial expression recognition, speaker diarization) with high-level interpretable features for classification and regression tasks [39]. Their approach achieved an accuracy of 90.9% accuracy for empathy prediction with decision tree classification models [39].

Although much work has been done on engagement quantification, to the best of our knowledge, only one other framework for engagement quantification exists in therapeutic consultations, that being MET. However, the MET framework presents a number of limitations that are addressed by the AES framework. The MET framework, lacks explainability due to its heavy use of semi-supervised DNN systems for both visual and text based classifiers. In this way the qualities that led to engagement estimation are unknown, and for this reason its capabilities are questionable and unreliable for medical practitioners who need explainability for decisions. The MET framework is also designed primarily to analyze 3 second clips for engagement estimation from the MEDICA dataset, which while suitable for real-time applications does not capture the overall quality of the therapeutic session

Our study makes two key contributions to the domain of engagement quantification. First, we develop a multimodal explainable therapeutic engagement quantification framework that analyzes entire consultations and serves as a decision support tool for clinicians treating clients with mental health disorders. This framework reflects the

therapeutic outcomes of these consultations; for instance, a high level of therapeutic engagement should indicate high-quality Motivational Interviewing (MI). We base this on the assumption that high-quality MI correlate with high engagement because effective interviewers successfully persuade clients to adhere to recommended treatments for addressing detrimental habits or conditions.

Second, we introduce a comprehensive video preprocessing and multimodal analysis framework for engagement and MI quantification. The preprocessing step is crucial for filtering out non-contributing or noisy frames, allowing us to extract information from all relevant frames rather than sampling at fixed or random intervals. Our overall AES framework is designed for heterogeneous video recordings and employs scene change detection, scene harmonization, and frame quality analysis during preprocessing. It also incorporates facial detection, FAR-based emotion classification, and NLP-based semantic analysis.

The rest of this work is organized as following: Section 2 explains the AnnoMI Dataset and presents the methodologies used in AES. Section 3 discusses the validation for choices in the AES model in detail, as well as presenting the results of the AES pipeline from scene change detection to classification. Section 4 presents drawn conclusions from the results, as well as future objectives and directions for the AES model.

2 METHODS

The proposed method is a multimodal classification system using video stream and transcript of MI videos. Feature extraction for MI quality is done through frame-to-frame analysis for FAR and utterance-to-utterance analysis for NLP. FAR is conducted through the use of the open source AU extraction toolboxes and a custom scene change detection algorithm is designed to extract client/clinician frames and determine frequency, transition rates, frequency of change in affects, and the chronic and acute motion of the head throughout the videos. NLP is conducted through semantic analysis to determine the utterance types for clinician behaviours and client talk types and report those as frequency of occurrence throughout video consultation. Classification revolves around a decision fusion of NLP and FAR, whereby the features determined from the first stage are used to classify the final decision of high quality or low quality MI. The overview of this model framework can be seen in Figure 2, each part of the process is explained in the subsections that follow.

2.1 Datasets

To the authors' knowledge, no datasets currently exist for quantifying engagement in therapeutics. To validate our methodology, we used the AnnoMI dataset [40], an expert-labeled NLP dataset for MI derived from previously unused video data from predominantly American institutions. Although the dataset is based on consultations for physical health issues rather than direct therapeutic engagement, it serves as a practical stand-in because MI quality closely parallels the therapeutic alliance, a substitute measure for engagement, which is the focus of this work. AnnoMI is a multi-label NLP dataset comprised of 133 videos, and 9699 utterances collected from youtube and vimeo, and annotated by ten experts in MI. It is annotated for the therapeutic behaviours such as reflection, question, input, and other, and client talk types such as change, neutral and sustain. It also has sub-type labels for clinician behaviours such as simple or complex reflections, open or closer questions, and utterance subtypes such as information, advice, giving options, and negotiation/goal-setting. The data is also labeled with goals of MI including reducing alcohol consumption, smoking cessation, weight loss, taking medicine/following medical procedure, more exercise/increasing activity, reducing drug use, reducing recidivism, and other. Furthermore, that dataset has labels for high and low quality MI which is the focus of this work [40].

The AnnoMI videos were also manually reviewed to assess their quality and suitability for this study's objectives. Out of the original 133 videos, only 72 had camera angles appropriate for facial affect recognition, thus reducing the sample size. Among these 72 usable videos, 17 were classified as low-quality MI and 55 as high-quality MI. Fortunately, the 17 low-quality MI videos were still deemed adequate for data augmentation due to the sufficiently large sample size.

2.2 Video Pre-processing

The AES framework is designed largely to provide an objective measure for the virtual one-on-one therapy sessions. It is expected there will not be more than one person in a video frame and the light condition can be suboptimal. The AnnoMI dataset was not originally designed for the purposes of FAR, however its collection of publicly available example videos for MI techniques provides an excellent environment for developing and testing robust systems for FAR based therapeutic engagement quantification in an in-wild setting. The dataset includes videos that vary in resolution, lighting conditions, types of camera used to record the videos, and participant positioning. The videos feature a diverse range of individuals and background scenes. This variability presents a challenge to the FAR systems, which depend heavily on factors such as lighting, camera angle relative to facial features, environmental obstructions, and frame resolutions. Additionally, these videos have not been analyzed or labeled for quality.

To address these problems we constructed a system for scene change detection, scene harmonization, and frame quality analysis which labeled sections of the video as non-usable poor quality or high quality frames, with the high quality frames being further segmented into client or clinician sections. Figure 1 presents the pipeline for this system, specifically down the frame-by-frame analysis of the raw video files.

2.2.1 Histogram Equalization

To improve image contrast and mitigate lighting issues for the facial affect recognition, we employed the python OpenCV Contrast Limited Adaptive Histogram Equalization (CLAHE) [41] technique for localized histogram equalization. CLAHE divides the image into distinct regions and applying histogram equalization to each region independently. This results in better handling of shadows and

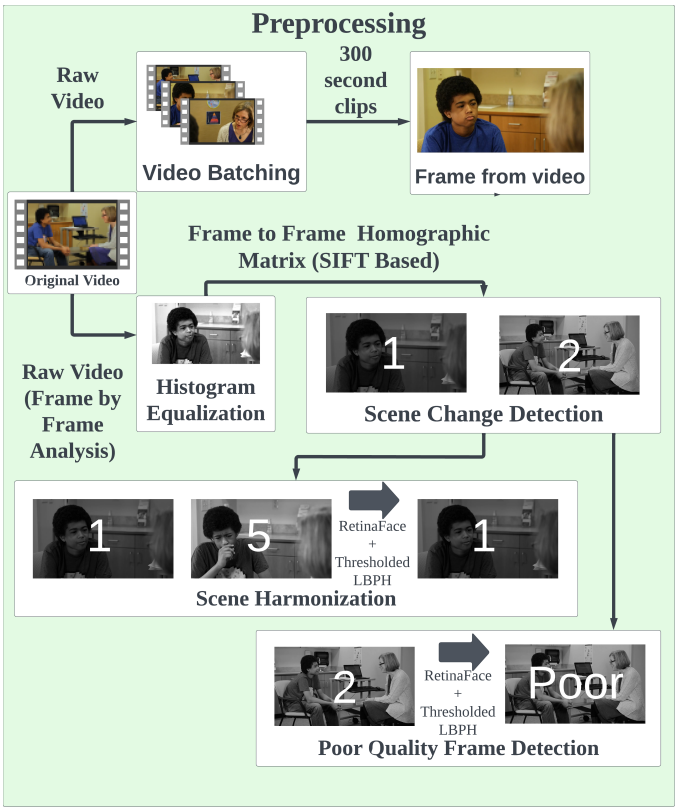

Fig. 1: Scene Change and Harmonization Pipeline

highlights as CLAHE mitigates shadows and overexposed regions of an image, and enhances details in darker regions of each image due to increased contrast, while at the same time limiting noise amplification as is the case in global histogram equalization. This equalization improves image key point detection.

2.2.2 Scene Change Detection

Scene change detection is essential for identifying low-quality frames and segmenting videos into client and clinician interactions. This process includes detecting black frames and fade in/out transitions between scenes. We utilized a technique based on homographic matrices [42], favored in robotics for their efficient computation and sensitivity to image changes and camera angle shifts. The 3x3 homographic matrix **H** is an image homography transformation that expresses the transforming from one frame to the next. To estimate the matrix **H**, each frame first underwent histogram equalization to enhance key point detection. The detected key points are used in the analyzed consecutive frames to identify and match image keypoints using Oriented Fast and Rotated BREIF (ORB) [43] and Scale Invariant Feature Transform (SIFT) [44] detectors and Brute Force Matcher. Matches were sorted by Hamming distance, and the top 90% were used to calculate the homographic matrix, a 3x3 matrix, indicative of scene consistency and represented by the following:

$$H = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix}$$

where the main diagonal h_{00}, h_{11}, and h_{22} represent the similarity of the two images to each other, and the off diagonal values of h_{02}, and h_{12} represent the suspected x and y shift from frame to frame. The closer a homographic matrix is to the identity matrix, the more similar in camera angle of the two frames. This metric is vital for assessing whether the camera has moved or another camera has been used in the consecutive frame, which potentially shifting to a wrong angle of an individual. We evaluate this by calculating the absolute difference between the sum of the main diagonal of the homographic matrix and the sum of the main diagonal of the identity matrix. A threshold is then applied to determine scene changes. Unique scene changes are recorded to aid in later client/clinician identification and to minimize redundant frame storage and computational load during scene harmonization.

2.2.3 Scene Harmonization

Homographic matrix method is highly sensitive to changes in frames. It was found that camera zoom changes or significant shifts in an individual's posture would be flagged as scene changes despite a refined threshold value. These additional frames were harmonized into one scene label and discarded since this camera angle already existed. To achieve this we implemented a deep-learning scene harmonization regime based on facial detectors and facial recognition to unflag the falsely detected scene change frames. Employing Retinaface [45] as a deep learning neural network facial detection system and Local Binary Pattern Histogram (LBPH) [46] as a facial recognition system, we were able to employ a relatively low computation facial recognition system that harmonizes scene change labels with the same individual while maintaining those with different individuals. Retinaface is employed to detect faces in detected scene change frames and to create bounding boxes around them. It also helps identify scene change frames with multiple faces, which are considered to be of poor quality scenes. LBPH uses the detected face boxes to determine if the faces in the detected scene change frames are the same or different. LBPH works using the following set of equations:

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} 2^p s(i_p - i_c) \tag{1}$$

where (x_c, y_c) are the central pixels, i_c and i_p are the intensities of the neighbor pixel, and s is the sign function defined as:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases} \tag{2}$$

and is defined for a specific point (x_c, y_c). Additionally the position of the neighbor $(x_p, y_p), p \in P$ can be calculated by:

$$x_p = x_c + R\cos\left(\frac{2\pi p}{P}\right) \tag{3}$$

$$y_p = y_c - R\sin\left(\frac{2\pi p}{P}\right) \tag{4}$$

where R is the radius of the circle and P is the number of sample points. This derivative of LBPH is referred to as

circular LBP [47] as its objective is to align an arbitrary number of neighbors on a circle with a variable radius, and generate local histograms. These localized histograms are concatenated together to create the LBPH used for facial recognition. The LBPH is a spatially enhanced feature vector that highlights facial features and computationally simplifies facial recognition. Circular LBP mask coordinates that do not correspond to the image coordinate are interpolated using bilinear interpolation defined by the following equation:

$$f(x,y) \approx \begin{bmatrix} 1-x & x \end{bmatrix} \begin{bmatrix} f(0,0) & f(0,1) \\ f(1,0) & f(1,1) \end{bmatrix} \begin{bmatrix} 1-y \\ y \end{bmatrix} \quad (5)$$

where $f(x,y)$ is the interpolated value of point (x,y) within the grid, x and y are the horizontal and vertical position in the interpolation grid, respectively. Additionally, $f(i,j)$ are values of the interpolation grid, and $\begin{bmatrix} 1-x & x \end{bmatrix}$ and $\begin{bmatrix} 1-y \\ y \end{bmatrix}$ are weights determining the interpolation along the horizontal and vertical axis, respectively. The decision on facial similarity was determined through a confidence score threshold value based on the outputted confidence of OpenCV's facial recognizer system, which utilized circular LBP as the recognizer. The confidence score measures the distance/difference between two given faces. Higher values correspond to different faces, while lower values correspond to similar/same faces. Through experimentation we determined the optimal confidence threshold to be less than 75.

2.2.4 Poor Quality Frame Detection

Our primary goal was to develop a virtual therapy system using video frames similar to platforms such as Zoom, Google Meet, or FaceTime, where typically one individual is visible and speaking directly to the camera. The system incorporates a comprehensive preprocessing capability to handle noisy poor quality frames, dynamic frame change, and invalid frames such as black, white, fading in or out, frames containing text or graphics, frames with no face present, side shots, or frames that feature more than one individual. After histogram equalization and during scene change detection, we complete the following tasks:

1) We identified black or white frames, using image contrast thresholding on a gray scale of 0 to 255.
   a) We defined black images as frames with an average pixel value ranging from 0 to 10.
   b) We defined white images as frames with an average pixel value ranging from 240 to 255 on a gray scale.
2) We identified fade in/out frames by utilizing a 50% image contrast threshold value.
   a) We labeled a frame as poor quality and discarded it if more then 50% of its pixels were dark.
   b) We defined dark pixels as those with a value of less than 50 on the gray scale.
3) We identified faceless frames and discarded them as poor quality. Frames without faces or with graphics and text are flagged using OpenCV's YUNET [48] detector.

During scene harmonization the Retinaface facial detector further assesses the remaining frames to detect side shots and frames with multiple faces, marking them as poor quality. The processed frames are then categorized, enhancing the dataset with labels for quality and identifying

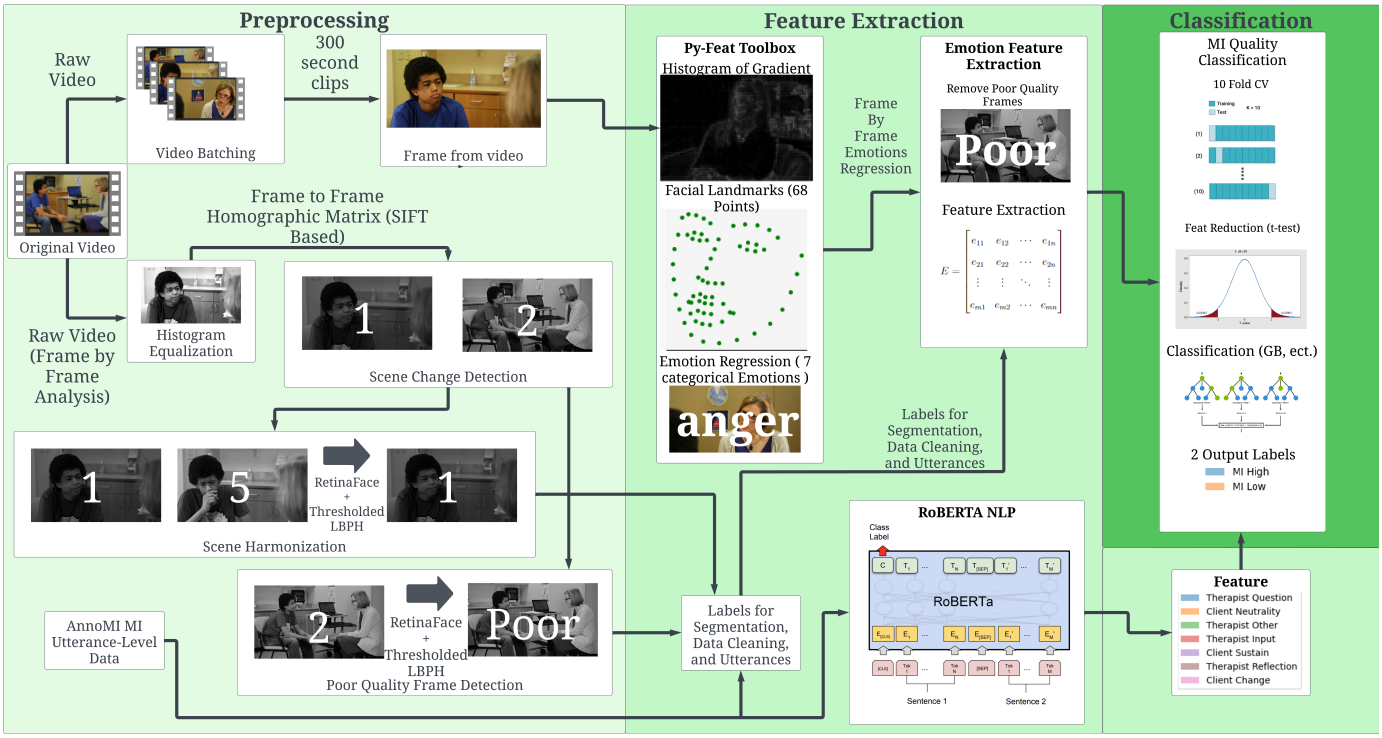

Fig. 2: AES Block Diagram using FAR and NLP modalities

clinician/client interactions, based on the original AnnoMI labels and timestamps. This organization helps in data cleaning and ensures that the video sections used meet the expectations of head-on video calls for effective virtual therapy.

To remove poor-quality frames, we employed a combination of scene change detection, harmonization, and poor-quality frame detection. First, scene detection generated a list of start and stop frame numbers for each scene, labeling them numerically in order of appearance. This list was then refined using harmonization to highlight key camera angles—specifically when the camera focused on the client, the clinician, or both. Next, based on the poor-quality frame criteria, all poor-quality scenes were labeled as zero in the list of start and stop frames. This refined list was then used to remove frames from the analyzed data after facial affect recognition. As a result, excluding poor-quality and side-angle frames led to the removal of 29,909 frames out of 136,096 frames of analyzed data—approximately 21.9% of the original data. Of the 29,909 frames 11,909 constituted frames where no face could be detected, while 18,000 constituted side angles and other poor quality frame criteria. This corresponds to approximately 2.68 hour of footage from 72 videos totaling over 13 hours of information. We do not believe the loss of these frames had significant effects on the final results given the extracted features are averaged for the duration of the video consultation. However we do believe the inclusion of the side shot camera angles specifically would cause significant alterations in the landmark feature signals described in Section 2.4.2, given artificial displacement caused by the extreme change in camera angle when comparing side angle frames to clinician or client frames.

2.3 Image Processing

2.3.1 Facial Affect Recognition

As seen in Figure 2, once the video frames are prepossessed and quality controlled, each quality frame is used for facial affect recognition. Taking advantage of publicly available toolboxes for FAR, we utilized the Python Facial Expression Analysis (Py-Feat) Toolbox [9], developed at Dartmouth College, Hanover, New Hampshire. Py-Feat is an open source toolbox for Facial Expression (FE) analysis containing modules for facial recognition, facial landmark detection, AU detection, and emotion detection. The Py-Feat detector framework chosen for AU extraction was a combination of FaceBoxes [49] for facial detection, Practical Facial Landmark Detector (PFLD) [50] for landmark detection, and an XGBoost classifier trained on Histogram of Oriented Gradients (HOG) [51] for action unit detection, with all other settings left at their default values. Faceboxes were utilized for facial detection due to the algorithm's stated and presented high accuracy. Resmasknet [52] was utilized for FE classification due to its exceptional accuracy. PFLD was utilized for landmark detection as videos utilized for the validation experiment were recorded with web cameras rather than mobile devices. Logistic Regression was used for AU detection as this classifier outputs regression AUs, rather than binary outputs for AU activation.

Py-Feat utilizes 20 AUs to classify 7 categorical emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. These AUs are obtained using a combination of landmark features and HOGs. Landmark features included the coordinates of 68 facial landmarks obtained using a face bounding box from the face detector.

2.3.2 Facial Affect Recognition Segmentation

To alleviate computational overhead, each of the original videos were segmented into 5 minute clips and analyzed for FAR in series. Larger videos were found to require in excess of 50 gigabytes of RAM, and were not practical for analysis on localized machines or cloud machines. Shorter videos, while lowering the computational requirements further, introduced excess additional frames that resulted in erroneous labeling of client/clinician segments, and poor quality data.

2.4 Facial Affect Recognition Feature Extraction

In FAR, the outputted FEs emotion measures and landmark locations are utilized for video features extraction. Features were determined at every 10th frame of video and aggregated to encompass the whole video consultation.

2.4.1 Emotion Features

The outputted frame to frame affect data from the Py-feat pipeline was utilized to determine three sets of feature tables. First, affect frequencies were calculated based on the following simple equation:

$$Affect\_score_i = \frac{Affect\_freq_i}{\sum_{j=1}^{7} Affect\_freq_j} \tag{6}$$

where the $Affect\_freq$ is the frequency count of an affect during the video, and the $Affect\_score$ is the scoring of the specific affect compared to the total frequency count for all seven affects. Second, taking inspiration from [53] and [54] we incorporate emotional dynamics and transitions as features, due to their importance in mental health and well-being [53] and their capacity to predict the nature of interactions and emotional changes during conversation [54] which we believe to be highly important for engagement quantification, especially in medical and therapeutic environments. However, we do not derive the emotion transition matrix using Markov modeling as in [53], [54], we instead calculated based on the empirical data of transitions following equation for frame to frame affect changes:

$$T_{e_i e_j} = \frac{\sum_{t=1}^{N-1} f_t^{e_i} \cdot f_{t+1}^{e_j}}{\sum_{t=1}^{N-1} f_t^{e_i}}, \quad \text{for } e_i, e_j \in E \tag{7}$$

where N is the total number of frames and $T_{e_i e_j}$ represents the probability of transitioning from emotion e_i at time t to emotion e_j at time $t+1$, and $f_t^{e_i}$ is the binary classification of emotion e_i at frame t. This matrix provides key information on the likelihood of one emotion shifting to another for either the clinician or client. The finalized transition matrix over the course of the video is represented as follows:

$$E = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{bmatrix}$$

where the main diagonal values such as e_{11} represents the likelihood of the emotion transitioning to itself, and the off diagonals values at the (i, j) position like e_{ij} represent the likelihood of the given emotion transitioning to any of the another six emotions. In addition, we also incorporate the emotional rate or change dynamics from [53] for each categorical emotion based on the normalized frequency of change from frame to frame calculated using the following equation:

$$D_e = \frac{1}{N-1} \sum_{t=1}^{N-1} |f_{t+1}^e - f_t^e|, \quad \text{for } e \in E \tag{8}$$

where D_e represents the rate of change for emotion e.

2.4.2 Landmark Features

Utilizing the 68 landmarks extracted from the Py-feat framework, we leverage chronic and acute head motion from [34] as an additional stream of information on engagement during conversations. First we utilize landmarks to generate two signals for feature extraction. Landmarks are averaged to a single x-y coordinate which is utilized to interpret the general movement of a participant's head. This information is then utilized to generate the signals of chronic and acute head motion, which are defined by the following equations:

$$\bar{p}_f = \frac{1}{68} \sum_{i=1}^{68} (x_{f,i}, y_{f,i}) \tag{9}$$

$$\bar{p}_{video} = \frac{1}{F \times 68} \sum_{f=1}^{F} \sum_{i=1}^{68} (x_{f,i}, y_{f,i}) \tag{10}$$

$$CH_f = \sqrt{(\bar{x}_f - \bar{x}_{video})^2 + (\bar{y}_f - \bar{y}_{video})^2} \tag{11}$$

$$AH_f = \sqrt{(\bar{x}_{f+1} - \bar{x}_f)^2 + (\bar{y}_{f+1} - \bar{y}_f)^2} \tag{12}$$

where CH_f and AH_f are the chronic and acute head motion signals over the video, respectively. The average landmark coordinate of a frame is $\bar{p}_f$, $\bar{p}_{video}$ is the mean averaged landmark coordinate of the video, $\bar{x}_f$ and $\bar{y}_f$ are the x and y components of $\bar{p}_f$, and $\bar{x}_{video}$ and $\bar{y}_{video}$ are the x and y components of $\bar{p}_{video}$. From these signals we extract common statistical features such as mean, variance, energy, mean-squared, and root-mean-square for every video.

2.5 Natural Language Processing

2.5.1 Preprocessing

Because the AnnoMI Dataset already included cleaned and extracted text transcriptions from the videos, minimal additional preprocessing was necessary. This lack of significant preprocessing is reflected in the overall pipeline diagram presented in Figure 2. The minimal preprocessing involved three straightforward tasks. Firstly, we cleaned the video titles by removing or replacing characters such as colons (:), quotations ("), and specific phrases such as "NEW VIDEO:" and "Batches." This step was essential to ensure consistency between the titles and the names of videos downloaded from sources such as YouTube and Vimeo. Secondly, we implemented label binarization. This involved transforming 8 columns of label data, which contained various subtypes, into 7 categorical labels representing four primary clinician behaviors (input, question, reflection, and other (any behavior that does not fit the first 3 categories)) and three main client talk types (neutral, change, and sustain). Lastly, we addressed the issue of repeat utterances, which occurred when videos received more than one label for a single utterance from MI experts. We resolved this by applying mean opinion scoring to these repeated utterances.

2.5.2 RoBERTa Fine Tuning

For the purposes of semantic analysis and classification of utterances, we employed the existing Robustly Optimized BERT Pretraining Approach (RoBERTa) [55] model from Huggingface, and fine-tuned its parameters for MI. RoBERTa is a state-of-the-art large language model that improves on the capabilities of BERT [56] through the use of dynamic masking, full sentences without next sentence prediction, large mini-batches, and a larger byte level byte-pair encoding. We accomplish our fine-tuning by using the base RoBERTa model and tuning two separate models, the client model and the clinician model. The separation of these models was required as the combined model struggle to distinguish the overlap in client talk types and clinician behaviours. For example, a client question may be mistaken for a clinician question and vice versa. The primary distinction between the models lies in their NLP features. The clinician model is built around four key therapist behaviors—input, question, reflection, and other—with normalized frequencies computed for each. In contrast, the client model categorizes speech into three types—neutral, change, and sustain. Aside from these differences in NLP feature sets, both models share the same landmark and emotion features. Both models incorporated Low-Rank Adaptation (LoRA) through the parameter-efficient fine tuning (peft) library in python which dramatically decreases the number of parameters required for training from 120+ million to roughly 600,000 parameters. This dramatically decreases training type while producing minimal if any affects on the model accuracy. To train the model we utilized a 60/40 split in the data with 60% of the data being used for fine-tuning, while the rest was used to validate the effectiveness of the model.

2.5.3 Feature Extraction

Feature extraction was conducted following the fine-tuning of RoBERTa into our clinician and client variants. We utilized our fine-tuned semantic analysis RoBERTa models to classify these utterances and utilized their normalized frequency as a feature based on observations in [40], where noticeable distinctions in clinican behaviours were found to be associated with high and low quality MI, with reflection and question being more pronounced in high quality, and input and other being more prominent in low quality. The frequency was based on the following equation:

$$Utterance\_score_i = \frac{Utterance\_freq_i}{\sum_{j=1}^{7} Utterance\_freq_j} \tag{13}$$

Where $Utterance\_score_i$ is simply the normalized frequency of each utterance.

2.6 Data Augmentation - ADASYN Oversampling

Due to an imbalanced dataset with only 23 of the 133 videos labeled low quality MI while the rest were high quality, a means of compensating for the small and imbalanced dataset was required. Since the dataset was already rather limited, downsampling to have an equal number of high and low quality MI interviews was not an option. To compensate for this we chose to create synthetic data based on the existing data using Adaptive Synthetic Sampling (ADASYN) [57]. ADASYN is an improved variant of Synthetic Minority Oversampling Technique (SMOTE) Data Augmentation which adaptively generates synthetic data samples that do not alter the general density distribution of the given minor class' data points. ADASYN generates points by calculating the degree of imbalance between minority and majority classes. It then computes the density distribution for each point in the minority class, and then generates a specified number of synthetic samples based on a position created by the selection of a random number k-nearest minority class neighbours. The outputted data points will provide a balanced representation of the data distribution and will also force ML algorithms to focus on difficult to learn examples. This effectively allows the system to limit bias, while maximizing relevance of the data samples, though at the cost of assuming the limited number of data points in the minority class is a good representation of all low quality MI. Using this approach we increased the sample size of low quality MI data points from 17 to 50 using ADASYN, increasing our sample size from 72 to 105. However, as some videos had little if any data from the FAR pipeline for either the clinician or the client, we effectively ended with two feature tables of size 105×87, and 105×88 for the client and clinician models respectively.

2.7 Feature Selection - t-test

To reduce computational overhead and enhance the robustness of our results, we implemented a t-test statistical analysis. This approach was chosen specifically to identify and select the most statistically significant features, setting a significance level at $P\leq0.05$. By employing this methodology, we were able to significantly streamline our feature set. Originally comprising 88 or 87 features for the clinician and client models respectively, the application of the t-test reduced the number of relevant features to a more manageable range of 10 to 20. This reduction not only simplifies the computational process but also focuses the analysis on the most relevant features, effectively increasing the predictive accuracy and reliability of the models.

2.8 Decision Fusion Classification

Decision fusion is based on an intermediate stage learned fusion, whereby a set of video based features are extracted and combined into a vector with their NLP based features to create one feature row per participant, this is done right before data augmentation and classification. Following feature extraction, the clinician and client models have 88 and 87 features, respectively. As shown in Figure 2 the final decision on low or high is based on classification with tested classifiers including Random Forrest(RF), Support Vector Machines (SVM), Extreme Gradient Boost (XGB), Gradient Boost (GB), Naive Bayes (NB), and Logistic Regression (LR) for both models with each result averaging 200 iterations using a 10-fold CV, and t-test statistical significance as a feature reduction mechanism with a p value of 0.05. Accuracy, sensitivity, specificity, and F1-score were used as performance metrics for both stages of classification based on [58] as well as confidence metrics such as Kappa, AUC, and Matthew Correlation Coefficient (MCC). The MCC is a statistical measure for model evaluation that predict performance as a value between -1 and 1. A value of -1 is interpreted as total disagreement between predictions and observations, a value of 0 presents predictions that are no better than random chance, and a value of 1 indicates perfect prediction agreement. It effectively measures the difference between predicted and actual values. The Kappa coefficient is a measure of inter-rater reliability on a scale of 0 to 1. It effectively compares classification results to values assigned by chance. Values closer to 0 present poor agreement where the values are no better then chance selection, and values closer to 1 are in perfect or near perfect agreement that the prediction is reliable.

3 RESULTS

3.1 Results

3.1.1 Scene Change Detection Validation

TABLE 1: Frame Change Count

Key Point Detector	Scene Change Frames	Useful Frames	Irrelevant Frames	Analysis Time
ORB	166	92	74	8.31 Hours
SIFT	125	74	51	14.43 Hours
N/A (Human)	151	39	112	1 Hour

We tested two key point detectors, Oriented Fast and Rotated BRIEF (ORB) and Scale Invariant Feature Transform (SIFT), in order to determine the optimal choice for key point detection. We utilized 26 of the available 72 videos to test the reliability and analysis time of the scene change detection system. The chosen videos were considered to have the best similarity to remote therapy sessions and the best resolution quality, and were relatively balanced in their MI labels with 12 videos being low quality MI, and 14 videos being high quality MI. Our findings, shown in Table 1, present that while ORB is roughly twice as fast, it generates significantly more useful and irrelevant frames. Our findings underline the fact that SIFT tends to be more accurate at the expense of time, but it also shows that SIFT generates lower computational overhead for the scene harmonization section of the pipeline. It was also noted that ORB tended to miss important scene changes and overemphasized scenes with varying poses and camera zoom levels, issues that SIFT handled with greater robustness.

3.1.2 Scene Harmonization Validation

For scene harmonization we tested four facial detectors, Dlib's Face Detector, OpenCV's Cascade [59] and YUNET

Decectors, and Retinaface. The facial detectors were accompanied by a number of facial recognizer in the form of LBPH, Dlib's Face Recognizer [60], FaceNET [61], and DeepFace [62]. Our findings, shown in Tables 2 and 3, where the accuracy is based on a comparison of the expected labels for each scene comparison, derived by manual evaluation of the save scene change frames from each detector. The

TABLE 2: Data Harmonization Labeling Results

Face Detector	Face Recognizer	Labeling Accuracy	Total Analysis Time
Dlib Detector	LBPH	62.3%	5.56 Hours
Cascade	LBPH	61.4%	1.32 Minutes
YUNET	LBPH	66.5%	1.43 Minutes
Retinaface	LBPH	87.6%	5.21 Minutes

accuracies in Table 2 are based on the same 26 videos from the scene change detection, and underline the increased performance of the Retinaface neural network architecture compared to more conventional algorithms such as YUNET, Cascade, and the Dlib Face Detector. While it is evident that Retinaface takes longer compared to the low computational systems of YUNET and Cascade, it has a dramatic improvement on performance for labeling frames as useful and irrelevant. This higher accuracy is still maintained even after the incorporation of the total 72 video dataset, dropping only to 75.08% as shown in Table 3. To further validate our choice in

TABLE 3: Facial Recognition Accuracy Results

Face Detector	Face Recognizer	Labeling Accuracy	Total Analysis Time
Retinaface	LBPH	75.1%	23.35 Minutes
Retinaface	Dlib Face Recognizer	70.4%	30.55 Minutes
Retinaface	FaceNET	74.9%	31.67 Minutes
Retinaface	DeepFace	77.2%	54.34 Minutes

systems, we tested LBPH against other commonly utilized open source frameworks, and found that it had comparable performance even to neural networks like FaceNET and DeepFace, however LBPH took far less time to complete the analysis and is far more explainable which is a key point in our framework. We attribute the loss in accuracy chiefly to the lower quality of the other 46 videos, but believe that the labeling accuracy is sufficient for the purposes of the work, and to present the capabilities of the algorithm.

3.1.3 Preprocessing Validation

To determine the effects of the preprocessing step on the final results we tested the classification with each classifier without the application of the preprocessing step. Tables 4 & 5 present the results of this classification.

TABLE 4: Performance Metric for Clinician Model without preprocessing step

Model	Accuracy	Specificity	Sensitivity	F-1 Score
RF	0.637 ± 0.174	0.649 ± 0.186	0.649 ± 0.186	0.599 ± 0.185
XGB	0.783 ± 0.126	0.790 ± 0.133	0.790 ± 0.133	$\mathbf{0.767 \pm 0.137}$
SVM-RBF	0.658 ± 0.144	0.646 ± 0.149	0.646 ± 0.149	0.615 ± 0.158
GB	$\mathbf{0.789 \pm 0.123}$	$\mathbf{0.797 \pm 0.125}$	$\mathbf{0.797 \pm 0.125}$	0.588 ± 0.250
NB	0.712 ± 0.142	0.706 ± 0.152	0.706 ± 0.152	0.684 ± 0.156
LR	0.718 ± 0.138	0.714 ± 0.144	0.714 ± 0.144	0.690 ± 0.151

TABLE 5: Performance Metric for Client Model without preprocessing step

Model	Accuracy	Specificity	Sensitivity	F-1 Score
RF	0.708 ± 0.172	0.725 ± 0.174	0.725 ± 0.174	0.682 ± 0.187
XGB	0.770 ± 0.134	0.774 ± 0.138	0.774 ± 0.138	$\mathbf{0.753 \pm 0.144}$
SVM-RBF	0.760 ± 0.132	0.759 ± 0.139	0.759 ± 0.139	0.734 ± 0.149
GB	$\mathbf{0.797 \pm 0.126}$	$\mathbf{0.802 \pm 0.129}$	$\mathbf{0.802 \pm 0.129}$	0.601 ± 0.252
NB	0.713 ± 0.140	0.713 ± 0.146	0.713 ± 0.146	0.690 ± 0.150
LR	0.707 ± 0.144	0.703 ± 0.154	0.703 ± 0.154	0.681 ± 0.159

When comparing these results to Tables 11 and 12, we observed an average accuracy increase of 6.34% across all six classifiers, including both client and clinician models. Notably, the clinician models experienced an average accuracy boost of 9.57%, while the client models saw a 3.12% improvement, suggesting that the preprocessing step may have a greater impact on the clinician model. The most pronounced effect was seen in the RF model, which exhibited a 23% increase in the clinician model—indicating that irrelevant frames had previously introduced significant noise into its results. Interestingly, although the SVM model also encountered additional noise from irrelevant frames, this led to a slight improvement; however, in general, the presence of these "noisy" frames resulted in decreased accuracy overall. These findings show the importance of curating the training data, especially for models sensitive to noisy inputs. The larger performance boost in clinician models indicates that the preprocessing step is crucial for achieving more accurate predictions. Although some classifiers can manage moderate noise, the evidence clearly shows that removing irrelevant frames leads to consistently better accuracy, specificity, sensitivity, and F1 scores across most of the evaluated algorithms, and thus justifying our employment of the preprocessing algorithm to the data.

3.1.4 Data Augmentation Validation

To determine the effects of data augmentation on the final results we tested the classification with each classifier without ADASYN oversampling. Tables 6 & 7 present the results of this classification.

TABLE 6: Performance Metric for Clinician Model without Data Augmentation

Model	Accuracy	Specificity	Sensitivity	F-1 Score
RF	0.753±0.126	0.506±0.124	0.506±0.124	**0.461±0.253**
XGB	**0.787±0.099**	0.500±0.0	0.500±0.0	0.438±0.034
SVM-RBF	0.546±0.187	0.420±0.171	0.420±0.171	0.389±0.139
GB	0.730±0.162	**0.734±0.169**	**0.734±0.169**	0.469±0.338
NB	0.620±0.249	0.466±0.139	0.466±0.139	0.386±0.129
LR	0.779±0.106	0.499±0.053	0.499±0.053	0.441±0.059

TABLE 7: Performance Metric for Client Model without Data Augmentation

Model	Accuracy	Specificity	Sensitivity	F-1 Score
RF	0.730±0.132	0.501±0.135	0.501±0.135	**0.454±0.131**
XGB	**0.771±0.101**	**0.498±0.010**	**0.498±0.010**	0.433±0.034
SVM-RBF	0.546±0.243	0.420±0.221	0.420±0.221	0.388±0.182
GB	0.746±0.155	0.762±0.165	0.762±0.165	0.528±0.315
NB	0.742±0.134	0.516±0.137	0.516±0.137	0.469±0.140
LR	0.769±0.107	0.498±0.026	0.498±0.026	0.434±0.043

When comparing these results to Tables 11 & 12 the

accuracy results are raised by an average of approximately 5.42% with some of the most extreme jumps occurring in the GB and NB classifiers which have accuracy jumps of 20%, pointing to the potential that these classifiers tend to emphasize the internal biases of the minority class. The least affected classifier by the data augmentation is XGB, which only has minor improvements to accuracy pointing to a far more robust classification framework.

3.1.5 RoBERTa Fine-Tuning Results

We employed 15 training epochs for the RoBERTa Models and achieved the results presented in Figure 3 and Figure 4. We derive the performance and result confidence metrics

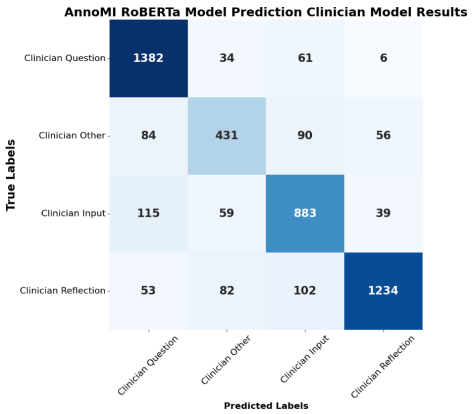

Fig. 3: Clinician Model Confusion Matrix Results after 15 epochs of training on RoBERTa.

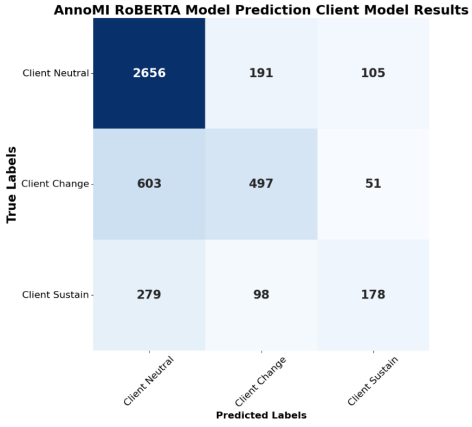

Fig. 4: Client Model Confusion Matrix Results after 15 epochs of training on RoBERTa.

of the two models and present them in Tables 8 and 9, respectively.

TABLE 8: Performance Metric for Fine-tuned RoBERTa Models using AnnoMI Dataset

Model	Accuracy	Specificity	Sensitivity	F-1 Score
Clinician	0.834	0.944	0.807	0.809
Client	0.715	0.793	0.551	0.577

TABLE 9: Result Confidence Metrics for Fine-tuned RoBERTa Models using AnnoMI Dataset

Model	MCC	Kappa	AUC(ovo)	AUC(ovr)
Clinician	0.773	0.772	0.947	0.953
Client	0.406	0.392	0.785	0.801

It is evident that the RoBERTa model struggled with interpreting the Client Talk Types, which we chiefly attribute to the labels being more vague as any response not directly sustaining or changing the targeted behaviour was labeled as neutral. This created a large imbalance in the existing dataset, and made interpretation of the finite differences between change, neutral, and sustain talk types more difficult to interpolate. However, despite this lower accuracy and confidence we believe both the Clinician and Client models in their current state are sufficient for the needs of the AES framework.

3.1.6 Modality Contribution

To better understand how each data stream contributes to overall model performance, we conducted an ablation study. In this analysis, we systematically removed feature groups associated with specific data streams—namely, Landmarks, Emotions, and NLP—and observed how these changes affected classification performance. Table 10 summarizes the average impact of each feature group on the four key performance metrics by comparing results with and without the respective features. These figures represent average percentage changes calculated across all six classifiers and both the Clinician and Client models, with values presented as decimals.

TABLE 10: Modality Ablation Study Results

Modality	Model	Accuracy	Specificity	Sensitivity	F-1 Score
Emotion	Clinician	0.1124	0.1168	0.1168	0.1480
	Client	0.0394	0.0434	0.0434	0.0504
Landmarks	Clinician	0.0168	0.0111	0.0111	0.0197
	Client	0.0970	0.0938	0.0938	0.1157
NLP	Clinician	0.0541	0.0913	0.0913	0.1252
	Client	0.0415	0.0765	0.0765	0.1051

Table 10 reveals that, on average, emotion-based features—particularly those capturing emotional transitions, dynamics, and frequencies during therapy sessions—have the most significant impact on performance. These features are closely followed by NLP features, while landmark features appear to contribute less overall. However, these averages do not tell the whole story; the effects vary significantly across different classifiers. For example, when emotional features were removed, the Clinician model experienced notable declines. Across all six classifiers, accuracy dropped by an average of 11.24%, while both specificity and sensitivity fell by 11.68%, and the F1-score decreased by 14.8%. More strikingly, in three of the classifiers (LR, NB, and XGB), the reductions exceeded 15%—surpassing the standard deviation of the results—which emphasizes the substantial influence of emotional features on model performance. In contrast, landmark features had a more pronounced effect on the Client model. Removing these features resulted in an average accuracy decline of 9.7%, with specificity and sensitivity each decreasing by 9.38%, and the F1-score falling by 11.15%. The RF classifier, in particular, showed an 18% drop in all metrics for the Client model—approximately 7% more than the typical deviation—and similar trends were observed in the Clinician model. Although classifiers such as NB and LR experienced declines of 11% in all metrics, these were closer to their error margins. Interestingly, the SVM classifier demonstrated a significant performance improvement in the Clinician model, with all

four metrics increasing by roughly 19%, even though it still underperformed in the Client model. This indicates that while landmark features are relevant, their overall impact is less pronounced than that of emotional features, and that its removal provides improvements for the SVM classifier that underline negative impacts on performance for these features. The removal of NLP features had a smaller effect on overall accuracy—decreasing it by 5.41% for the Clinician model and 4.15% for the Client model—but resulted in dramatic declines in other metrics. In particular, the GB classifier was severely affected: in the Clinician model, accuracy fell by 15%, specificity and sensitivity dropped by 26%, and the F1-score plunged by 55.4%. For the Client model, the GB classifier saw a 12% decrease in accuracy, a 32.4% reduction in both specificity and sensitivity, and a 41.7% decline in the F1-score. Although NLP features generally had a lower impact on the overall results, their removal significantly compromised the performance of the top classifier (GB), highlighting their key role in achieving accurate classification.

3.1.7 Decision Fusion Results

To determine the quality of MI as either low or high, we evaluated two models—the Clinician Model and the Client Model—using six distinct classifiers: RF, XGB, SVM, GB, NB, and LR. Of these, the GB classifier consistently outperformed the others. The performance metrics, averaged over 200 iterations of 10-fold cross-validation with all six classifiers, are detailed in Tables 11 & 12. Our results present high performance metrics, showing our Decision Fusion approach is highly effective for both models.

TABLE 11: Performance Metric for Clinician Model using 10-fold Cross-Validation and ADASYN based Data Augmentation

Model	Accuracy	Specificity	Sensitivity	F-1 Score
RF	0.867±0.113	0.872±0.111	0.872±0.111	**0.855±0.123**
XGB	0.825±0.125	0.831±0.127	0.831±0.127	0.812±0.135
SVM-RBF	0.630±0.156	0.628±0.153	0.628±0.1649	0.590±0.153
GB	**0.895±0.094**	**0.899±0.095**	**0.899±0.095**	0.796±0.185
NB	0.829±0.131	0.830±0.137	0.830±0.137	0.811±0.147
LR	0.825±0.132	0.826±0.137	0.826±0.137	0.808±0.147

TABLE 12: Performance Metric for Client Model using 10-fold Cross-Validation and ADASYN based Data Augmentation

Model	Accuracy	Specificity	Sensitivity	F-1 Score
RF	0.811±0.124	0.816±0.129	0.816±0.129	**0.795±0.136**
XGB	0.801±0.126	0.809±0.128	0.809±0.128	0.785±0.135
SVM-RBF	0.744±0.135	0.746±0.144	0.746±0.144	0.724±0.145
GB	**0.811±0.123**	**0.817±0.128**	**0.817±0.128**	0.629±0.149
NB	0.736±0.142	0.737±0.151	0.737±0.151	0.715±0.153
LR	0.739±0.141	0.739±0.152	0.739±0.152	0.717±0.153

Furthermore the confidence of our results, detailed through the confidence metrics in Tables 13 and 14, suggest a high likelihood of prediction accuracy. The MCC often approaches values closer to 1, indicating strong predictive performance, as values closer to 1 are interpreted as being closer to a perfect prediction. The Kappa coefficient also achieves high values, ranging from 0.41 to 0.9 depending on classifier. Given most classifiers fall in the 0.61 to 0.8 range, this presents a moderate to substantial agreement in results between raters presenting a low likelihood that results were achieved by chance. Additionally, the AUC values which reflect a classifier's ability to distinguish between two or more classes was generally above the 0.8 threshold; this presents robust model interpretation. Notable exceptions to these trends include the SVM classifier for both models and both LR and NB for the client model.

TABLE 13: Confidence Metrics for Clinician Model using 10-fold Cross-Validation and ADASYN based Data Augmentation

Model	MCC	Kappa	AUC
RF	0.742±0.219	0.720±0.232	0.872±0.111
XGB	0.658±0.249	0.634±0.255	0.831±0.127
SVM-RBF	0.270±0.318	0.241±0.290	0.628±0.153
GB	**0.778±0.198**	**0.899±0.094**	**0.886±0.104**
NB	0.661±0.268	0.638±0.273	0.830±0.137
LR	0.654±0.270	0.630±0.275	0.82±0.137

TABLE 14: Confidence Metrics for Client Model using 10-fold Cross-Validation and ADASYN based Data Augmentation

Model	MCC	Kappa	
RF	0.629±0.252	0.604±0.257	**0.816±0.129**
XGB	0.612±0.250	0.585±0.254	0.809±0.128
SVM-RBF	0.489±0.284	0.465±0.277	0.746±0.144
GB	**0.604±0.253**	**0.816±0.128**	0.795±0.133
NB	0.472±0.297	0.449±0.290	0.737±0.151
LR	0.476±0.299	0.452±0.292	0.739±0.152

3.1.8 Feature Selection Visualization

To enhance system explainability and clarify the feature selection process, we visualized feature selections across 200 iterations using our best-performing classifier, GB, as presented in Figures 5 and 6. These visualizations, derived from our t-test feature selection mechanism (which yielded the most effective results), emphasize the features most pertinent to the therapeutic setting. One figure represents the clinician model, while the other depicts the client model, offering clear insights into feature importance and overall model behavior. Although similar visualizations were produced for all six classifiers, we present only the GB results for brevity; however, key observations across all visualizations will be discussed.

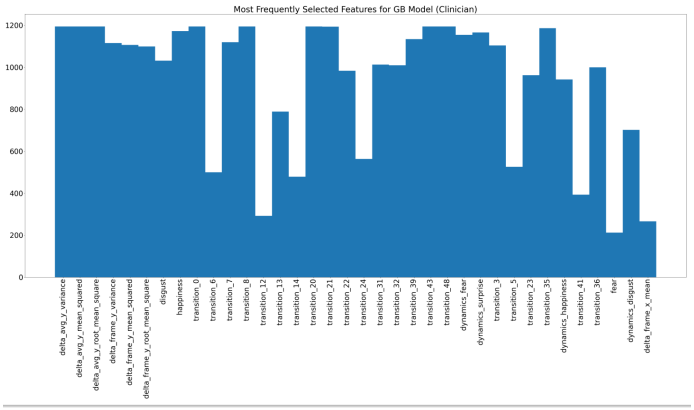

Fig. 5: GB Clinician model feature spread visualization, with only features with more then 200 appearances over 200×10-fold Cross-Validation iterations and hyperparamater tuning

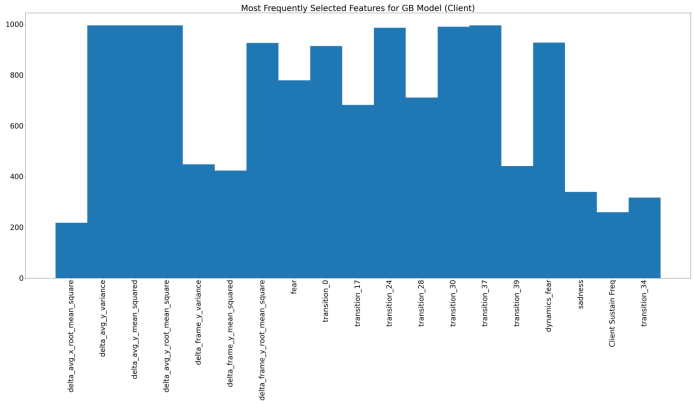

Fig. 6: GB Client model feature spread visualization, with only features with more then 200 appearances over 200×10-fold Cross-Validation iterations and hyperparamater tuning

For clarity, we present only the features that appeared more than 200 times during training and testing. From this information a number of inferences can be made. First, the clinician model consistently selects more features than the client model, indicating that a broader range of attributes is relevant for clinicians, who play a central role in providing information, treatment, and driving the conversation. Second, vertical head motion features appear in nearly every model iteration. This emphasizes the significance of non-verbal gestures such as nodding or reflective poses in the MI interaction. Third, emotion transitions and dynamics strongly influence model decisions. Notably, the persistent anger-to-anger transition (transition 0) is pivotal across nearly all cross-validation splits. Frequently occurring transitions include:

1) **Transition 48:** Neutral to neutral
2) **Transition 43:** Neutral to disgust
3) **Transition 39:** Surprise to sadness
4) **Transition 37:** Surprise to fear
5) **Transition 30:** Sadness to fear
6) **Transition 28:** Sadness to anger
7) **Transition 24:** Happiness to happiness
8) **Transition 17:** Fear to happiness

These indicate that continued neutrality (transition 48) may reflect attentiveness, while negative-to-negative transitions (30 and 28) might distinguish adversarial interactions in low-quality MI from more collaborative ones in high-quality MI. Conversely, transitions from negative to positive (transition 17) could signal high-quality, constructive exchanges. Fourth, the frequencies and dynamics of emotions—particularly fear, disgust, happiness, and surprise, play significant roles in model decisions. Finally, Client talk types and clinician behaviors were not consistently selected by the models, suggesting that the distinction between low- and high-quality MI may not rely heavily on these conversational attributes. This observation warrants further investigation and may prompt the exploration of more discriminative NLP features.

3.1.9 Model Comparison

To further prove the validity of our work we compared the AES framework to a handful of other multi-modal models from both the empathy and engagement domains, more specifically the educational and digital subdomains of each. We chose these models as they are closest in relation to our framework, with the most similar model being MET which focuses on therapeutic engagement quantification. Table 15 presents this comparison Our results indicate that AES performs comparably to both empathy-based and engagement-based models. Although its accuracy and F1 score are not the highest, it's important to note that these scores are averaged over 200 iterations, highlighting the robustness of our model. In contrast, the comparison models—which are neural network architectures—have not all reported on the robustness of their results within their respective works. It is unclear whether their reported metrics represent optimal or average performances, and due to the lack of publicly available code, reproducing their results is challenging. The model most similar to ours, MET, relies on non-explainable neural network architectures and reports performance metrics limited to an individual loss function metric. These factors collectively provide confidence that, despite lower performance, our work is more robust and representative. The only other model that reports on robustness is the system described in [38]. However, this is based on an ablation study that removes modalities to test the performance of various parts of the model and reports no variance in the results. Nevertheless, our model performs comparably to the results of this work.

TABLE 15: Comparison of AES Against Related Multimodal Frameworks for Engagement and Empathy Prediction

Model Framework	Best Classifier	Performance Metrics
AES (Ours)	**Gradient Boost**	**Accuracy: 89.5% (T)[a], 81.1% (P)[b]** **F1: 79.6% (T), 62.9% (P)**
MET [27]	Semi-Supervised GAN	RMSE: 0.10
Teacher-Student DNNs [63]	Decision Tree	Accuracy: 90.9% F1: 90.1%
LSTM + BERT [38]	SWFAN	Accuracy: 77.6% (ER)[c], 86.4% (EE)[d], 78.3% (CR)[e] F1: 77.7% (ER), 78.5% (EE), 78.5% (CR)
Student MES [35]	CNN	RMSE: 0.106 Accuracy: 93.6% F1: 92.4%

[a] T: Clinician model
[b] P: Client model
[c] ER: Emotional Reaction
[d] EE: Expression of Experience
[e] CR: Cognitive Reaction

4 DISCUSSION AND CONCLUSION

In this study, we introduced AES, an engagement quantification system designed for therapeutic environments,

and demonstrated its application through a multimodal framework pipeline. This pipeline integrates results from the Py-Feat toolbox for FAR and RoBERTa for semantic analysis. We developed preprocessing techniques to clean and segregate video data of clients and clinicians using algorithms for scene change detection and scene harmonization. Features for engagement classification were defined and evaluated using the AnnoMI Dataset. Our multimodal pipeline performed well, achieving high accuracy in filtering out irrelevant data and in classifying low and high-quality MI, particularly with the GB classifier. The GB classifier reported an accuracy of 89.5% for the Clinician model and 80.1% for the Client model, indicating robust results for a first-of-its-kind work. These outcomes are notably favorable given the small and imbalanced nature of the dataset. Importantly, the persistence of certain features as top contributors in post processing model interpretation suggests that these features, which are rooted in clinical theory, social norms, and established engagement research, are likely useful in distinguishing between high and low motivational interviewing. The fact that features such as emotion transition rates, vertical head movement (e.g., nodding), and utterance categories (e.g., reflective statements or open questions) emerged consistently as influential across classifiers supports the rationale behind their inclusion. These features were selected based on the assumption that they reflect observable behaviors associated with engagement: for instance, dynamic affect may signal emotional involvement, nodding may indicate attentiveness, and certain therapist speech patterns may align with techniques known to foster client participation. The model's reliance on these features in achieving strong classification performance lends further support to their relevance for therapeutic engagement quantification. While the directionality of individual features' contributions cannot be conclusively determined, their prominence reinforces the value of integrating theoretically and clinically motivated behaviors into computational models. In this way, AES not only performs well but does so by leveraging a feature set that clinicians and researchers are likely to find meaningful and interpretable in context. Nevertheless, there are discrepancies between expected and actual results, which we attribute primarily to the dataset being labeled for MI rather than engagement. This misalignment might have influenced the effectiveness of the features in accurately reflecting MI quality. For instance, due to the lack of ground truth labels for intermediate steps and the dataset's design prioritizing natural language processing over video analysis, evaluation is limited to the final classification of each utterance as reported in Sections 3.1.7 and 3.1.5. Consequently, precise error propagation analysis within the system remains challenging, particularly for facial affect recognition where independent frame-by-frame analysis prevents tracking errors propagation due to the fact that the error in the analysis of one frame does not affect the analysis of previous or following frames. Additionally, we have identified several limitations in the system that warrant further investigation. First, our dataset is relatively small at approximately 72 data samples, and this data is heavily imbalanced. While these issues can be addressed to some degree with data augmentation, this approach strengthens pre-existing biases in the data. This means the current system may struggle with in wild data or when shifting to its target domain of mental health consultations. Second, we have not assessed how well AES performs across varied cultural backgrounds and therapeutic approaches, such as cognitive-behavioral therapy versus psychodynamic therapy. Because our current dataset is drawn exclusively from American institutions with limited ethnic diversity, there is a risk that cultural uniformity could undermine accuracy when the system encounters different norms of behavior, gesture, or speech patterns. AES has also not been tested in multiple therapy modalities to determine whether certain behavioral markers require recalibration or weighting adjustments. However, despite these challenges, we remain optimistic that many of our chosen features, such as, nodding to signal agreement, emotional transitions, and dynamics will continue to work as they are grounded in universal psychological principles and have been validated in cross-cultural research. We therefore anticipate that, with thoughtful adaptation and additional training data, AES's core feature set will generalize effectively to a broad spectrum of consultation formats. Third, our approach is currently highly computationally intensive, primarily in the pre-processing phase, due to the scene change detection and harmonization algorithm and the FAR analysis. With both of these systems requiring 3+ days of computational time to analyze 11+ hours of video data on a AMD Ryzen 5 5600X CPU with 64 gigabytes of RAM. This effectively results in an approximate ratio of 12 seconds to analyze every 1 second of video. This ratio would likely increase further if we were required to extract the audio and audio transcript of each video for feature extraction; both of which are intentions of the finalized AES pipeline. Fourth, our decision fusion has a bias towards FAR given 85 of the extracted feature are derived from the FAR analysis while only 3-4 are extracted from the NLP analysis. FAR may not always be possible, even under the expected circumstances of virtual therapy sessions. FAR is also considered a non-specific feature in therapeutic settings as its validity in these environments is still being tested. Currently only NLP is considered a specific feature for therapeutic environments, due to its proven validity in predicting therapeutic outcomes. Thus greater focus on this data stream is recommended for any system trying to measure engagement. Despite these limitations, we believe this is the first step to objectively define a robust engagement quantification system designed as a support tool for clinicians in a therapeutic environment. Our future work intends to overcome the major limitations of our current model starting with our relatively small dataset. We are working to create and secure access to a larger more varied dataset containing recordings of clinical sessions labeled with validated inventories as well as variable therapy types. This dataset will be more balanced, variable, with larger sample sizes to allow for further testing of our developed prototype and its continued improvement. In addition to this we intend to incorporate additional features for engagement in the NLP domain. We believe this will prove integral in creating a more robust and versatile model capable of being deployed in a multitude of therapeutic or consultation environments. In addition to these steps we are also looking into lower computational alternatives for the FAR pipeline, primarily through the employment of the

Graphical Neural Networks to landmark based FAR which is less computationally intensive than AU based FAR.

REFERENCES

[1] J. Wosik *et al.*, "Telehealth transformation: COVID-19 and the rise of virtual care," *Journal of the American Medical Informatics Association*, vol. 27, no. 6, pp. 957–962, 2020.

[2] L. Cadel *et al.*, "A scoping review of patient engagement activities during COVID-19: More consultation, less partnership," *PLoS One*, vol. 16, no. 9, p. e0257880, Sep. 2021.

[3] D. Duong, "Five lessons from a year of virtual patient partnerships," *CMAJ*, vol. 193, no. 29, pp. E1145–E1146, Jul. 2021.

[4] M. J. Lambert and D. E. Barley, "Research summary on the therapeutic relationship and psychotherapy outcome." *Psychotherapy: Theory, Research, Practice, Training*, vol. 38, no. 4, pp. 357–361, 2001.

[5] A. Tetley, M. Jinks, N. Huband, and K. Howells, "A systematic review of measures of therapeutic engagement in psychosocial and psychological treatment," *Journal of Clinical Psychology*, vol. 67, no. 9, pp. 927–941, 2011.

[6] M. Hall, Alan Meaden, Jo Smith, Chris, "Brief report: The development and psychometric properties of an observer-rated measure of engagement with mental health services," *Journal of Mental Health*, vol. 10, no. 4, pp. 457–465, 2001.

[7] M. Srinivasan, A. J. Phadke, D. Zulman, S. Thadaney, I. Nelligan, M. Artandi, and C. Sharp, "Enhancing patient engagement during virtual care: A conceptual model and rapid implementation at an academic medical center," *NEJM Catalyst*, no. 09, pp. 1–10, 2020.

[8] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311–324, 2007.

[9] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang, "Py-feat: Python facial expression analysis toolbox," 2021. [Online]. Available: https://arxiv.org/abs/2104.03509

[10] S. Liu, S. Liu, Z. Liu, X. Peng, and Z. Yang, "Automated detection of emotional and cognitive engagement in mooc discussions to predict learning achievement," *Computers and education*, vol. 181, p. 104461, 2022.

[11] E. E. Felix, N. Hernandez, and I. Rebaï, "Exploring the limits of lexicon-based natural language processing techniques for measuring engagement and predicting mooc's certification," in *CSEDU 2022: 14th International Conference on Computer Supported Education*, vol. 2, 2022, pp. 95–104.

[12] X. Solé-Beteta, J. Navarro, B. Gajšek, A. Guadagni, and A. Zaballos, "A data-driven approach to quantify and measure students' engagement in synchronous virtual learning environments," *Sensors*, vol. 22, no. 9, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/9/3294

[13] X. Tang, Y. Gong, Y. Xiao, J. Xiong, and L. Bao, "Facial expression recognition for probing students' emotional engagement in science learning," *Journal of Science Education and Technology*, vol. 34, no. 1, pp. 13–30, 2025.

[14] F. A. Bachtiar, A. M. Mahesa, and E. W. Cooper, "Engagement level detection using facial extraction and multi-stacked convolutional neural network in e-learning settings," in *2024 12th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*. IEEE, 2024, pp. 208–213.

[15] N. Prameela, M. S. Das, and R. D. Borreo, "Multi-head network based students behaviour prediction with feedback generation for enhancing classroom engagement and teaching effectiveness," *International Journal of Information Technology and Computer Science*, vol. 16, no. 5, pp. 81–100, 2024.

[16] G. Morio and C. D. Manning, "An nlp benchmark dataset for assessing corporate climate policy engagement," *Advances in Neural Information Processing Systems*, vol. 36, pp. 39 678–39 702, 2023.

[17] A. Singh, S. Gupta, H. Satyawali, V. Sharma, S. Awasthi, and S. Vats, "Moodsync: Personalized video recommendation based on user face emotion," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*. IEEE, 2024, pp. 975–980.

[18] S. P. Pattar, E. Coronado, L. R. Ardila, and G. Venture, "Intention and engagement recognition for personalized human-robot interaction, an integrated and deep learning approach." IEEE, 2019, pp. 93–98.

[19] D. Mazzei, F. Chiarello, and G. Fantoni, "Analyzing social robotics research with natural language processing techniques," *Cognitive computation*, vol. 13, no. 2, pp. 308–321, 2021.

[20] M. Ninaus, S. Greipl, K. Kiili, A. Lindstedt, S. Huber, E. Klein, H.-O. Karnath, and K. Moeller, "Increased emotional engagement in game-based learning – a machine learning approach on facial emotion detection data," *Computers and education*, vol. 142, p. 103641, 2019.

[21] R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, "Development and application of emotion recognition technology—a systematic literature review," *BMC psychology*, vol. 12, no. 1, p. 95, 2024.

[22] N. Alhakbani, "Facial emotion recognition-based engagement detection in autism spectrum disorder." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 3, 2024.

[23] M. H. Siddiqi, I. Ahmad, Y. Alhwaiti, and F. Khan, "Facial expression recognition for healthcare monitoring systems using neural random forest," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[24] S. Malins, G. Figueredo, T. Jilani, Y. Long, J. Andrews, M. Rawsthorne, C. Manolescu, J. Clos, F. Higton, D. Waldram, D. Hunt, E. Perez Vallejos, and N. Moghaddam, "Developing an automated assessment of in-session patient activation for psychological therapy: Codevelopment approach," *JMIR medical informatics*, vol. 10, no. 11, pp. e38 168–e38 168, 2022.

[25] L. Zhang, O. Arandjelovic, S. Dewar, A. Astell, G. Doherty, and M. Ellis, "Quantification of advanced dementia patients' engagement in therapeutic sessions: An automatic video based approach using computer vision and machine learning," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, vol. 2020. IEEE, 2020, pp. 5785–5788.

[26] K. Vijay, R. Raghakeerthana, S. Thusheel *et al.*, "Ai-powered mental health assessment using emotion detection for real-time analysis," in *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*. IEEE, 2025, pp. 530–535.

[27] P. Guhan, N. Awasthi, , K. McDonald, K. Bussell, D. Manocha, G. Reeves, and A. Bera, "Developing an effective and automated patient engagement estimator for telehealth: A machine learning approach," 2023. [Online]. Available: https://arxiv.org/abs/2011.08690

[28] A. C. Heusser, D. J. DeLoss, E. Cañadas, and T. Alailima, "Leveraging machine learning to examine engagement with a digital therapeutic," *Frontiers in Digital Health*, vol. 5, p. 1063165, 2023.

[29] M. R. Hasan, M. Z. Hossain, S. Ghosh, A. Krishna, and T. Gedeon, "Empathy detection from text, audiovisual, audio or physiological signals: Task formulations and machine learning methods," 2024. [Online]. Available: https://arxiv.org/abs/2311.00721

[30] R. Islam and S. W. Bae, "Facepsy: An open-source affective mobile sensing system-analyzing facial behavior and head gesture for depression detection in naturalistic settings," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. MHCI, pp. 1–32, 2024.

[31] Y. Liu, K. Wang, L. Wei, J. Chen, Y. Zhan, D. Tao, and Z. Chen, "Affective computing for healthcare: Recent trends, applications, challenges, and beyond," 2024. [Online]. Available: https://arxiv.org/abs/2402.13589

[32] G. Pei, H. Li, Y. Lu, Y. Wang, S. Hua, and T. Li, "Affective computing: Recent advances, challenges, and future trends," *Intelligent Computing*, vol. 3, p. 0076, 2024. [Online]. Available: https://spj.science.org/doi/abs/10.34133/icomputing.0076

[33] N. A. Aziz, A. Manzoor, M. D. Mazhar Qureshi, M. A. Qureshi, and W. Rashwan, "Unveiling explainable ai in healthcare: Current trends, challenges, and future directions," *medRxiv*, pp. 2024–08, 2024.

[34] A. Behera, P. Matthew, A. Keidel, P. Vangorp, H. Fang, and S. Canning, "Associating facial expressions and upper-body gestures with learning tasks for enhancing intelligent tutoring systems," *International Journal of Artificial Intelligence in Education*, vol. 30, pp. 236–270, 6 2020.

[35] P. Bhardwaj, P. K. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of deep learning on student engagement in e-learning environments," *Computers & electrical engineering*, vol. 93, p. 107277, 2021.

[36] R. Islam and S. W. Bae, "Facepsy: An open-source affective mobile sensing system - analyzing facial behavior and head gesture for depression detection in naturalistic settings," *Proceedings of the ACM on human-computer interaction*, vol. 8, no. MHCI, pp. 1–32, 2024.

[37] D. V. Rodriguez, J. Chen, R. V. Viswanadham, K. Lawrence, and D. Mann, "Leveraging machine learning to develop digital engagement phenotypes of users in a digital diabetes prevention program: Evaluation study," *JMIR AI*, vol. 3, p. e47122, 2024.
[38] Z. Zhu, X. Li, J. Pan, Y. Xiao, Y. Chang, F. Zheng, and S. Wang, "Medic: A multimodal empathy dataset in counseling," 2023. [Online]. Available: https://arxiv.org/abs/2305.02842
[39] Y. Pan, J. Wu, R. Ju, Z. Zhou, J. Gu, S. Zeng, L. Yuan, and M. Li, "A multimodal framework for automated teaching quality assessment of one-to-many online instruction videos," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1777–1783.
[40] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. Reforgiato Recupero, and D. Riboni, "Anno-mi: A dataset of expert-annotated counselling dialogues." IEEE, 2022, pp. 6177–6181.
[41] S. Pizer, R. Johnston, J. Ericksen, B. Yankaskas, and K. Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," in *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*, 1990, pp. 337–345.
[42] Y. Luo, X. Wang, Y. Liao, Q. Fu, C. Shu, Y. Wu, and Y. He, "A review of homography estimation: Advances and challenges," *Electronics*, vol. 12, no. 24, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/24/4977
[43] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.
[44] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:174065
[45] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," 2019.
[46] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Computer Vision - ECCV 2004*, T. Pajdla and J. Matas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 469–481.
[47] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," vol. 3021, 05 2004, pp. 469–481.
[48] W. Wu, H. Peng, and S. Yu, "Yunet: A tiny millisecond-level face detector," *Machine Intelligence Research*, vol. 20, 04 2023.
[49] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Faceboxes: A cpu real-time face detector with high accuracy," 2018.
[50] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "Pfld: A practical facial landmark detector," 2019.
[51] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893 vol. 1.
[52] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4513–4519.
[53] "New mental health diseases and conditions data have been reported by investigators at duke university (the temporal dynamics of spontaneous emotional brain states and their implications for mental health)," pp. 395–, 2022.
[54] M. A. Thornton and D. I. Tamir, "Mental models accurately predict emotion transitions," *Proceedings of the National Academy of Sciences - PNAS*, vol. 114, no. 23, pp. 5982–5987, 2017.
[55] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv.org*, 2019.
[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.
[57] H. He, Y. Bai, E. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, vol. 10. IEEE, 2008, pp. 1322–1328.
[58] S. Krishnan, "Biomedical signal analysis for connected healthcare," *Elsevier*, 2021.
[59] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–I.
[60] D. Zhang, J. Li, and Z. Shan, "Implementation of dlib deep learning face recognition technology," in *2020 International Conference on Robots Intelligent System (ICRIS)*, 2020, pp. 88–91.
[61] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2015, p. 815–823. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298682
[62] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.
[63] Y. Pan, J. Wu, R. Ju, Z. Zhou, J. Gu, S. Zeng, L. Yuan, and M. Li, "A multimodal framework for automated teaching quality assessment of one-to-many online instruction videos," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 1777–1783.

Martin Ivanov Martin Ivanov is a MASc student enrolled in Electrical & Computer Engineering at Toronto Metropolitan University (TMU), Toronto, Ontario, Canada and is expecting to graduate in 2024. He is specializing in Biosignal Analysis with a focus on affective computing. Currently, he is working as a research assistant at the Signal Analysis Research group at TMU, Toronto, Ontario, Canada.

Alice Rueda Alice Rueda received the B.Sc. degree in electrical engineering and the M.Sc. degree in electrical and computer engineering from the University of Manitoba, Winnipeg, MB, Canada, in 1995 and 1999, respectively. Alice received her (hc) LL.D. degree from Brock University in 2020 and Ph.D. degree in Electrical and Computer Engineering from TMU in 2021. She is currently a post-doctoral fellow at St. Michael's Hospital, Toronto, ON, Canada. Her research interest are inin signal processing and analysis for Parkinson's disease and neuroimaging analysis for major depressive disorder.

Venkat Bhat Dr. Bhat is a staff psychiatrist at St. Michael's Hospital and at the University Health Network. He is an Associate Professor within the Department of Psychiatry at University of Toronto. He envisioned and developed the Interventional Psychiatry Program at St. Michael's Hospital, an interdisciplinary program which offers emerging interventions for neuropsychiatric disorders. He co-developed the multi-institutional Digital Interventions Intelligence Group (DiiG), leads the Health Care AI/Analytics pillar within the Institute for Biomedical Science, Engineering Technology, is a member within Faculties of Medicine Biomedical Engineering, and within the Schwartz Reisman Institute for Science and Technology at the University of Toronto.

Sridhar Krishnan Dr. Krishnan joined Toronto Metropolitan University, Toronto, ON, Canada, in 1999. He is now a Professor of Electrical, Computer, and Biomedical Engineering. From 2007 to 2017, he was a Canada Research Chair in Biomedical Signal Analysis. His research interests are in biomedical signal analysis, audio signal analysis, and explainable machine learning,Prof. Krishnan is a Fellow of the Canadian Academy of Engineering. He is a recipient of the Outstanding Canadian Biomedical Engineer Award, Achievement in Innovation Award from Innovate Calgary, Sarwan Sahota Distinguished Scholar Award, Young Engineer Achievement Award from Engineers Canada, New Pioneers Award in Science and Technology, and Exemplary Service Award from the IEEE Toronto Section.