Emotion Recognition in Conversations Using Brain and Physiological Signals

Nastaran Saffaryazdi
zsaf419@aucklanduni.ac.nz
Empathic Computing Lab, Auckland Bioengineering Institute, University of Auckland
Auckland, New Zealand

Yenushka Goonesekera
ygoo781@aucklanduni.ac.nz
University of Auckland
Auckland, New Zealand

Nafiseh Saffaryazdi
nsaffaryazdi@gmail.com
Empathic Computing Lab, Auckland Bioengineering Institute, University of Auckland
Auckland, New Zealand

Nebiyou Daniel Hailemariam
nebhailemariam@gmail.com
Addis Ababa University
Addis Ababa, Ethiopia

Ebasa Girma Temesgen
ebagirma8@gmail.com
Addis Ababa University
Addis Ababa, Ethiopia

Suranga Nanayakkara
s.nanayakkara@auckland.ac.nz
Augmented Human Lab, Auckland Bioengineering Institute, University of Auckland
Auckland, New Zealand

Elizabeth Broadbent
e.broadbent@auckland.ac.nz
Department of Psychological Medicine, Faculty of Health Psychology, University of Auckland
Auckland, New Zealand

Mark Billinghurst
mark.billinghurst@auckland.ac.nz
Empathic Computing Lab, Auckland Bioengineering Institute, University of Auckland
Auckland, New Zealand

ABSTRACT

Emotions are complicated psycho-physiological processes that are related to numerous external and internal changes in the body. They play an essential role in human-human interaction and can be important for human-machine interfaces. Automatically recognizing emotions in conversation could be applied in many application domains like health-care, education, social interactions, entertainment, and more. Facial expressions, speech, and body gestures are primary cues that have been widely used for recognizing emotions in conversation. However, these cues can be ineffective as they cannot reveal underlying emotions when people involuntarily or deliberately conceal their emotions. Researchers have shown that analyzing brain activity and physiological signals can lead to more reliable emotion recognition since they generally cannot be controlled. However, these body responses in emotional situations have been rarely explored in interactive tasks like conversations. This paper explores and discusses the performance and challenges of using brain activity and other physiological signals in recognizing emotions in a face-to-face conversation. We present an experimental setup for stimulating spontaneous emotions using a face-to-face conversation and creating a dataset of the brain and physiological activity. We then describe our analysis strategies for recognizing emotions using Electroencephalography (EEG), Photoplethysmography (PPG), and Galvanic Skin Response (GSR) signals in subject-dependent and subject-independent approaches. Finally, we describe new directions for future research in conversational emotion recognition and the limitations and challenges of our approach.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**; **HCI design and evaluation methods**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Human-Computer interaction (HCI), Multimodal Emotion Recognition, Conversational emotion recognition, Electroencephalography (EEG), Galvanic Skin Response (GSR), Photoplethysmography (PPG), Multimodal fusion

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03…$15.00
https://doi.org/10.1145/3490099.3511148

ACM Reference Format:
Nastaran Saffaryazdi, Yenushka Goonesekera, Nafiseh Saffaryazdi, Nebiyou Daniel Hailemariam, Ebasa Girma Temesgen, Suranga Nanayakkara, Elizabeth Broadbent, and Mark Billinghurst. 2022. Emotion Recognition in Conversations Using Brain and Physiological Signals. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3490099.3511148

1 INTRODUCTION

Emotions are multimodal processes that are related to numerous external and internal activities. We are increasingly surrounded by smart devices, computers, and digital agents, and our need for greater interactions with these systems is growing. The role of emotion in human-human and human-computer interaction (HCI) is significant [99]. For example, in remote e-learning [34] or intelligent assistants [47] or humanoid robots [5] it is useful to measure the user's emotions to improve the quality of their interactions. Similarly, in Empathic Computing applications, it is important to measure the emotions of people teleconferencing together and use the result to improve remote communications [59].

Emotional recognition can be essential in intelligent user interfaces where researchers want to design interactive systems that use multimodal sensing to identify and implicitly respond to a user's emotional state. Aranha et al. [3] reviewed software with intelligent user interfaces capable of recognizing emotions in various fields, including health, education, security, and art. According to that review, emotion recognition has often been used for adjusting sounds, user interfaces, graphics, and content based on user emotion. Zepf et al. [98] discuss the importance of emotion-aware systems for cars. Hu et al. [25] introduced an intelligent emotion-aware conversational agent who recognizes emotions through acoustic features of speech. Chin et al. [14] showed the importance of empathy between conversational agents and humans in improving people's aggressive behaviors. Schachner et al. [70] provide a literature review describing research in developing intelligent conversational agents for health-care, and especially chronic diseases.

A large and growing body of literature has focused on conversational emotion recognition using facial expressions [42], audio and text analysis [68]. Most studies in this area are based on collected video and audio data from acted scenarios, TV clips, or monologues. However, recognizing emotion in a real-life conversation is more complex than in acted scenarios [23]. Moreover, behavioral emotion recognition faces challenges like cultural differences, age, and gender variety. Also, behavioral modalities can often be controlled by the user and easily faked. So, we need new approaches for robust methods to recognize emotion in conversation reliably.

Some studies have suggested using unconscious physiological measures for emotion recognition, like Electroencephalography (EEG) [1], Galvanic Skin Response (GSR), heart rate variability (HRV) [84], blood volume pressure (BVP), body temperature (BT), respiration rate (RR), and so on [15]. Many researchers have shown that although EEG and physiological signals are weak, they can be used to reliably recognize underlying emotions [6, 73]. Also, a combination of various modalities can improve the recognition result [27, 100].

Most EEG-based studies used passive tasks to induce emotions, like watching videos or images, listening to music, or recalling memories [1]. Although inducing emotions with these tasks creates natural emotional responses, they are not able to cover a wide range of complex emotional states like emotional responses in a conversation [26, 90]. Daily life interactions include many activities such as talking and body movement, reasoning and decision making, language production and processing, and feeling emotions [60]. These activities activate different parts of the brain and affect different physiological signals [22]. Recognizing emotions in these activities is a challenging task because of the presence of muscle artifacts and a wider range of emotional responses [26]. However, to the best of our knowledge, no research has specifically studied brain activity and physiological signals for recognizing emotion in a conversation.

Our primary goal in this paper is to explore the effectiveness of brain activity and physiological signals for recognizing emotions in a conversation with an emotional subject. In this paper, we created a conversational setting to create spontaneous emotions similar to real-life conversations and collected PPG, EEG and GSR in conversations dataset (*PEGCONV*) [1]. We analyzed the EEG and physiological data and found that these signals could effectively recognize emotions in conversations. We evaluated our work on both discrete and two-dimensional emotion models. To summarize, our research contributions in this work are as follows:

- Creating an effective experimental setting for inducing spontaneous emotions in a conversation and collecting multimodal data.
- Recognizing emotions using EEG and physiological data in conversation.
- Identifying the limitation and challenges of using physiological signals for recognizing emotions in conversations

2 RELATED WORKS

Researchers have modeled emotions in two different ways: a discrete emotion model and a dimensional emotion model. The well-known discrete emotion model was introduced by Ekman [18] and categorized emotions into six basic types; happiness, sadness, surprise, anger, disgust, and fear. The dimensional emotion model [41] considers emotions as a continuous combination of several psychological dimensions like arousal and valence. Valence represents the level of positivity or negativity of emotion, while arousal represents the level of excitement or calmness.

Humans express their emotions using a series of actions, including facial expressions, gestures, head and body movements, speech, and some physiological responses like changes in heart rate, temperature, and brain activity [8, 79]. Affective computing uses image processing, text and speech analysis and, biological signal processing to detect emotion in computer interfaces [58, 62]. Some research has focused on using behavioral modalities like eye-tracking[44], facial expressions [42], gesture [56], text and audio processing [73], while others has focused on physiological emotion recognition [73]. Fusing various modalities in emotion recognition is believed by many researchers to improve the results of emotion recognition and to enable better human-human or human-machine interaction [79]. In the following sections, we review state of the art in behavioral, physiological, and multimodal emotion recognition.

2.1 Behavioral Modalities

The research on recognizing emotion using facial expressions, speech recognition, and text analysis has a long history and has achieved considerable accuracy by developing a variety of machine learning strategies [42, 54, 69]. For example, Li et al. [42] reviewed

[1]The PEGCONV dataset will be made available to the research community. https://pegconv.nastaran-saffar.me/

facial expression datasets and investigated the progress and challenges in this area. They showed that although facial expression recognition has achieved more than 95% accuracy, there still needs to be more research to make accurate models for in-the-wild emotion recognition using facial expression because of culture, age, and gender varieties. EmotioNet [20] and Affectnet [51] are two datasets for facial expression in-the-wild which collected data from the internet. Although these datasets improve the recognition strategies, they are built on top of intense facial expressions, which is not appropriate for all real-life situations where people usually show more subtle expressions [97].

A large and growing body of literature on Emotion recognition in social interactions has focused on speech analysis by analyzing audio signals or natural language processing techniques. Recognizing emotions from speech has considerably improved through the use of deep learning strategies[54]. Since the performance of deep learning strategies highly depends on the number of data [76], some research focuses on collecting data for training models. There are many widely used multimodal datasets, such as IEMOCAP [11] which is a multimodal dataset that collected audio and video data from actors in a scripted scenario. In the SEMAINE dataset [49] ordinary people played pre-scripted roles with emotional content. MELD [61] and CMU-MOSEI [96] are large scale datasets which has been widely used in deep learning-based studies [50, 53]. Over 1400 dialogues from the Friends TV series have been collected in the MELD dataset. The CMU-MOSEI dataset is the largest dataset in emotion and sentiment analysis which collected 23453 annotated segments with various topics and speakers. In this dataset, monologue videos have been collected from people who posted their opinions on social multimedia websites.

Although emotion recognition from audio and video has attracted considerable attention, it still faces some limitations. Most of the datasets and research are based on acted scenarios or monologue talks, and the data has been annotated by human annotators. Although annotators are trained, they may not be able to reliably recognize underlying emotions, and their annotations depend on their perspectives [62], and people can control or fake their behaviors.

2.2 Physiological Modalities

Human physiological responses are involuntary reactions that contain salient information about human emotional responses [40]. To overcome behavioral emotion recognition weaknesses, a considerable amount of the current research on emotion recognition pays particular attention to analyzing EEG and physiological signals. Generally, people cannot control physiological responses, so these signals can reveal hidden emotions [29]. For example, Kolodziej et al. [37] showed that GSR can reflect the intensity or the arousal level of the emotional state of humans. Setyohadi et al. [71] recorded physiological signals every second and used this data as the features for classifying positive, neutral, and negative emotional states using a Support Vector Machine (SVM) [86]. Guo et al. [24] used HRV features for recognizing five emotional states. Overall, there has been considerable progress in the literature in using EEG signals for recognizing emotion, as reviewed by Wagh et al. [89].

Although single physiological modalities have been widely used, combining these signals can produce even better results. For example, Yazdani et al. [93] have used EEG, BVP, BT, RR, EOG, Electrooculography (EOG), and Electromyography (EMG) for affect recognition in people watching videos, achieving the highest accuracy of 61.7% and 53.3% in subject-independent approach using EEG signals. In a similar work, Yang et al. [92] used physiological data to recognize the emotions of people playing video games. Shu et al. [73] reviewed physiological emotion recognition in the literature and identified the impact of each emotion on each physiological cue. Chaparro et al. [13] and Matlovic et al. [48] found an improvement in emotion recognition by combining facial expression and EEG emotion recognition.

In most studies that rely on EEG and physiological emotion recognition, emotion has usually been stimulated in a passive task like watching videos or images, listening to music, or recalling memories to minimize body movements and muscle artifacts[1, 8]. For example, Soleymani et al. [75] created the MAHNOB-HCI dataset of EEG signals, eye movements, and respiration rate data while people were watching videos. In a similar study, Zheng et al. [99] created the SEED-IV dataset of EEG signals and eye-tracking data in a video watching task. The DEAP dataset [36] is the most popular dataset in EEG and physiological emotion recognition, which provides facial expression data as well. In this dataset, people watched short music videos, and ground truth labeling was based on a continuous emotion model using self-report questionnaires. However, people's daily activities include body movements that can generate muscle artifacts in physiological data, making it more difficult for emotion recognition. In the next section, we review emotion recognition in more natural settings.

2.3 Physiological Emotion Recognition in Social Interactions

A limited number of researchers have studied physiological responses in non-passive tasks, including body movements. Ringeval et al. [65] introduced the multimodal RECOLA dataset of video, audio (in French), EDA, and ECG signals from 46 participants who collaborated to solve a dyadic task. Boating et al. [7] investigated the quality of dyadic interactions based on physiological signals collected by a smartwatch in real-time. They developed a mobile application that uses smartwatch multimodal data and a machine learning method to recognize emotions in female partners with 73.4% accuracy. Youssef et al. [95] have identified a promising relationship between behavioral and brain activities during a natural conversation using Functional Magnetic Resonance Imaging (fMRI) signals. They found a connection between functional brain areas and speech and face perception.

Although some research has focused on exploring physiological signals in non-passive tasks, to the best of our knowledge, none of them have used EEG and physiological signals to recognize spontaneous emotions in a non-acted face-to-face conversation. Our research is the first step toward creating a validated dataset of emotions from physiological signals in a conversational setting, which could be helpful in creating emotion-aware interactive systems and intelligent user interfaces.

The following section describes the experimental setup we used to induce spontaneous emotions in non-acted conversations and collect EEG and physiological signals. Then we perform a preliminary analysis on the collected data and show the potential of EEG and physiological signals for recognizing the underlying emotions in conversation. Our approach could be used in many applications. For example, in the future, interactive machines like humanoid robots [64], assistant agents [45], and emotion-sensitive car safety systems [28] could be equipped with this system to be emotionally intelligent and respond more effectively when people are engaging in conversations.

3 INDUCING EMOTIONS

There are many different ways of inducing emotions; however, not all of them are equally effective. Siedlecka et al. [74] have classified emotional stimuli into five methods; (1) watching visual stimuli like images and videos, (2) listening to music, (3) recalling personal emotional memories, (4) accomplishing psychological procedures, and (5) imagination of emotional scenes. Quigley et al. [63] have added words, body movements, physiological manipulators like caffeine, and Virtual Reality (VR) tasks to this set. Dyadic interaction tasks are another way for eliciting emotions, which is widely used among psychologists [66]. Siedlecka et al. [74] empirically showed that recalling and imagining memories is one of the most robust ways of inducing emotions.

In order to recognize emotions in conversation, we need robust models trained using human responses in similar situations. These models could be used for recognizing emotions from everyday interactions like remote collaboration or interactions with machines. For this reason, we needed to create a conversational setting where people could feel emotions during their conversations, and we could collect natural responses. In this study, we created a dyadic conversational setting and collected physiological data. An interviewer guided the conversation by imagining, recalling, and expressing personal emotional memories to stimulate emotions. The interviewer was a pre-intern psychologist trained in consulting and therapy conversations and had skills to navigate an emotional conversation with each participant. She was trained in how to connect with people, which helped the participants be more open in expressing themselves and feeling emotions [2].

We targeted five emotions; happiness, sadness, anger, fear, and a neutral state. The interviewer guided the conversation for each target emotion by asking questions about their feelings in emotional situations. Although the details of each conversation and follow-up questions depend on the participant and participants' memories, the main questions were similar for all participants. The following list shows the main questions:

- *Tell me about a time when you felt the most [emotion of interest].*
- *What was going on for you at the time? (Asking for details about the event)*
- *Can you tell me what it was about the situation that made you so emotional?*
- *Tell me about how you felt that experience in your senses (giving them an example of feeling hot when you're angry or starting to shake when you're nervous/anxious)*
- *If you close your eyes now, and picture the event/time/place that made you [emotion of interest], can you tell me what is going on for you (senses, imagination, etc.)*
- *How do you feel reflecting on that moment now?*

4 EXPERIMENTAL SETUP

4.1 Experiment Protocol

We recruited 23 volunteers (13 female and 10 male) aged between 21 and 44 years old ($\mu = 30, \sigma = 6$) from university students and staff. All participants talked about their memories and experiences in different emotional situations with the expert interviewer. The expert interviewer guided the conversation with each participant for about 50 minutes, 10 minutes for each of the five target emotions. The order of emotions was randomly chosen, except for the last one, which was always happiness, to induce a positive mood at the end of the study and ensure that participants were not leaving in a distressed state. There was nobody else in the room except for the participant and the interviewer. In order to maintain a close and comfortable distance, the interviewer sat in front of the participant with a one-meter distance. She asked participants to be open in expressing themselves as much as possible and assured them that nobody would listen to their recorded audio. This research was approved by the University of Auckland Human Participants Ethics Committee (UAHPEC).

4.2 Data Collection

Since the brain is the source of a lot of activity in our body, monitoring and analyzing its activity can be used to recognize emotions. Electroencephalography (EEG) is one of the brain monitoring methods that measure the electrical activity of cortical neurons in a safe and non-invasive way [17]. We used All-in-One EEG Electrode Cap Starter Kit with a cyton-daisy board from OpenBCI [2] to record brain activities during the experiment with a sampling rate of 125 Hz. It has 16 electrodes located on the scalp based on the 10-20 system of electrode placement [55]. The EEG data was streamed to a PC via Bluetooth using a USB dongle. This EEG cap provides high-quality data while it is affordable, easy to set up, and has open-source software and hardware.

Previous studies have shown a connection between the nervous system and sweat glands on human skin. Changes in the level of sweat secretion because of emotional arousal lead to changes in skin resistance [81]. This variation has been known as the Electrodermal Activity (EDA) or Galvanic Skin Response (GSR). Changes in heart rate is another physiological response to emotions [38]. Measuring Heart Rate Variability (HRV) is a suitable method for analyzing medical and mental states. Photoplethysmography (PPG) is a novel method for measuring Blood Volume Pulse (BVP) using infrared light [19]. It has been shown that HRV can be extracted accurately from PPG data. We also collected GSR and PPG using the Shimmer3 GSR+ module [3]. It is a wearable, reliable, lightweight, high-precision wireless sensor platform for wirelessly transmitting GSR signals from fingers. It can be used for biomedical research applications [10]. An optical pulse sensor attached to a GSR+ module was used to record PPG signals from the index finger. With

[2] https://openbci.com
[3] http://www.shimmersensing.com/

the Shimmer3, we can stream data via Bluetooth to a PC or locally store it. It is also equipped with an accelerometer, gyroscope, magnetometer, and altimeter sensors. Each participant wore a Shimmer3 sensor on their non-dominant wrist (only one participant was left-handed). The sensors were attached to their three middle fingers. The experiments were conducted in a temperature-controlled room to minimize the effect of the environment on the physiological signals.

We also recorded audio and video data during the experiment. One Intel Realsense camera with a frame rate of 30 Hz was used to record facial video, and a Logitech webcam recorded audio. We used this data as metadata to evaluate data labeling, track each session, and further explorations in the future. Figure 1 shows the experiment setup. We used an Asus laptop (TP410U) to run the experiment scenario and record data. We designed a multi-platform and open-source python module to synchronously record data from different devices [4].

Figure 1: The experiment setup

4.3 Scenario

The scenario starts by displaying a timer alongside a randomly selected emotion from the desired emotion list on a screen in front of the expert interviewer. The interviewer then talked to the participant for approximately 10 minutes about the target emotion. The interviewer resets the timer after completing the conversation about a particular emotion and moves on to the next emotion. When the button is pressed, the software sends a marker to all sensors and then resets the timer for the next emotion and repeats the process. Figure 2 shows the scenario.

We had to keep the length of the experiment as short as possible because the EEG headset was not comfortable enough to wear for a long time. Also, after around one hour, the EEG signal quality decreases as the electrode gel dries, and we did not want to distract participants by refilling the gel. We also did not consider any gap between two emotions to decrease the length of the experiment; instead, the interviewer softly changed the topic at the end of each emotion to switch to the next emotion. This is consistent with natural conversations, in which subjects can change quickly.

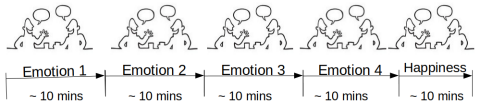

Figure 2: The experiment scenario. Emotions 1 to 4 were selected randomly from sadness, fear, anger and neutral

4.4 Ground Truth Labeling

The ground truth for training and testing the models were based on self-report data where the participants listened to the conversation and labeled their emotion. Since the length of each trial was long and emotional states were not long-lasting, instead of one label, we collected three labels for each conversation. Based on our observation, we roughly divided each trial into 3 equal parts to consider participants' emotions at the beginning (initializing with the topic), in the middle (immersing in the subject), and at the end (topic transition or cooling down). Although more labels would have been better, it was time-consuming for participants to listen to their recorded audio.

After finishing the experiment, we explained this partitioning method to participants. We asked them to listen to some parts of their recorded audio from the beginning, middle, and end of each emotional topic. They did this to remember their emotional state during the conversation and to fill out a self-report questionnaire (Figure 3) for each part. Using these three labels, we also covered the situation where the emotional effect of the previous conversation remained. During each conversation, the emotion may disappear and change to neutral or even other emotions, or it may take time to reach the target emotion. Although we don't know precisely when these transitions happen, using these three labels is more reliable than only one.

5 EMOTION RECOGNITION STRATEGY

Several emotion recognition steps, including preprocessing, feature extraction, feature selection, and classification [8] were applied to predict emotions. First, we removed noise from the data. Since our machine learning methods use trials of the same size, we considered the length of the smallest trial, which was around 1 minute, as the size of each trial. The shortest conversations were around 5 minutes which, after splitting based on the three-part labeling method, the length of each part was less than one and a half minutes. To prepare the data for analysis, we divided each trial into 1-minute segments. Trials' sizes were diverse because participants had different personalities. Some were talkative, while others used short sentences. Using the three-part labeling method, we then labeled each new segment according to its temporal location in the main trial. We ignored the last window if the length was less than one minute and considered the middle part as the most prominent part if the length of the trial was not a multiple of three. For example, if the size of the main trial was 10 minutes and 25 seconds, we ignored the last 25 seconds and tagged 3 minutes with the beginning label,

[4] https://octopus-sensing.nastaran-saffar.me/

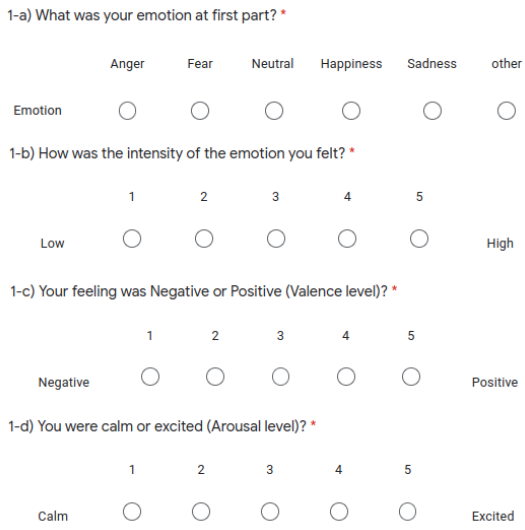

Figure 3: The emotion self-report questionnaire. Participants answered these questions for each emotion topic for each part (beginning, middle and end).

4 minutes with the middle label, and three last minutes with the end label.

In the next step, similar to previous studies [84] we partitioned each new trial into smaller windows and labeled them with the trial's label. We did this to increase the amount of data. We used 5 seconds as the size of each window and used a moving window with 4 minutes overlap with the window data. We chose 5 seconds as the size of each window to ensure that we considered physiological delays in response to external stimuli. Then we extracted some hand-crafted features from each signal in each window. Finally, we used these feature vectors as the input of two different methods for recognizing emotions.

In the first method, we used a Random Forest Classifier (RFC) [9] to classify feature vectors. In the second method, we considered all feature vectors of consequence windows in each trial as the input of a Long Short Term Memory (LSTM) network to extract temporal features in each trial and then used a dense layer to classify the emotional states. Finally, we used a weighted decision level fusion method to fuse various modalities. The following sections describe our strategy in more detail.

5.1 Preprocessing

To clean the EEG data, we applied bandpass filters, and extracted frequencies between 1 to 45 Hz [43]. Then bad channels were removed and interpolated with signals in good locations based on the spherical spline method [57]. In the next step, we referenced channels to a common average reference [43]. We considered the first 5 seconds of each conversation as the baseline data, and baseline normalization similar to [80]. In this method, they calculated the average of each second of the baseline data and made a vector for each channel. The size of each vector equals the baseline length in seconds. Then we subtracted these values from each window of data when the length of the window equaled the baseline size.

We used a bandpass filter in the range of 0.1 to 15 Hz to remove noise from the GSR data[84]. Then we applied a median filter to remove rapid transient artifacts. Similarly, considering the PPG frequency range, which is between 0.5 Hz to 4 Hz [52], the noise was removed by applying bandpass filters with a low-cut frequency of 0.5 and a high-cut frequency of 4 Hz. Finally, we applied baseline normalization to the PPG and GSR signals similar to the EEG signals.

5.2 Feature Extraction

We extracted some hand-crafted features of the EEG, GSR, and PPG signals as the feature set, which was common in previous studies [8]. We extracted the EEG power bands Delta (1-3 Hz), Theta (4-7 Hz), Alpha (8-11 Hz), Beta (12-30 Hz), and Gamma (31-45 Hz) using Fast Fourier Transforms for each window of data. Then, we calculated the Power Spectral Density (PSD) of each band for all channels and made a feature vector including 80 features (16 channels * 5 power bands) for each trial's windows.

Similar to previous studies, we extracted some statistical features from each window of the GSR signals [84]. We then used the Pearson correlation method [87] to find the highest correlated features and used these as the final feature set (see Table 1).

To extract features from the PPG signal, we used the Heartpy python library [85] to calculate heart rate variability and heart rate features which were statistical features of consequences inter-beats. Also, similar to GSR statistics, we measured some statistics of the PPG signals in the time-domain. Finally, similar to GSR feature extraction, we used the Pearson correlation to find the best subset of features that accurately describes the signal. Table 1 shows the list of the final feature set that we used.

5.3 Classification and Fusion Strategy

We used two different methods to classify emotional states. In the first method, we used a Random Forest Classifier (RFC)[9] to classify feature vectors extracted from each window of the EEG, GSR, and PPG signals. In the second method, we considered feature vectors extracted from consequences windows in each trial as a new time series and fed them to an LSTM network to extract temporal features. The LSTM network has a two-layer stacked LSTM, while the first layer has 80 neurons and the second one has 30 neurons. Finally, a dense layer and Adam optimizer [35] was used to classify trials.

We used a decision-level fusion strategy to improve recognition performance. Fusion strategies have been categorized into two main types, (1) early fusion or feature fusion and (2) late fusion or decision fusion [88]. We used the weighted fusion strategy, which is a decision-level fusion method. We measured the weight for each modality in the training step. Then we used these weights as the weight of each modalities' predictions in the test step. The final predictions are a weighted sum of the various modalities' predictions. Equation 1 shows the fusion strategy. In this equation, p shows the calculated probability using each modality for each class x. The coefficients of each modality, including a, b and c, have been measured based on training data.

Table 1: Extracted features from each window for the EEG, GSR and PPG signals. Since the length of each trial is 60s and we partitioned them to 5 second windows with 4 seconds overlap, we have 56 windows for each trial.

	Features
EEG	PSD(alpha), PSD(beta), PSD(gamma), PSD(delta), PSD(theta) for each channel [1]
GSR	average, median, maximum, minimum, variance, standard deviation, frequency of local extremum, variance of local extremum's amplitude, standard deviation of local extremum's amplitude, skewness of local extremum's amplitude, kurtosis of local extremum's amplitude, max abstract amplitude of local extremum, sum of positive derivatives, sum of negative derivatives [37, 84, 92]
PPG	average, median, maximum, minimum, hrv metrics (ibi, bpm, pnn50, sdnn, rmssd, sd1) [72]

$$p_o^x = a \times p_{EEG}^x + b \times p_{GSR}^x + c \times p_{PPG}^x \quad (1)$$
$$x \in [labels]$$
$$a + b + c = 1$$

6 RESULTS

6.1 Emotion Stimulation Evaluation

We employed two different sources of data to assess the effectiveness of our experimental setup in evoking the intended emotions. The first source was the reported emotions in the self-report questionnaires, and the second source was the content of the conversations. For classifying emotional states based on the arousal-valence model, we classified arousal and valence levels according to low and high levels, as in previous studies. We used a 5-point scale to measure arousal and valence in this study. We used three (the middle score) as the threshold for categorizing arousal and valence levels. Scores less than three were considered low, and scores above three were deemed to be high. We assumed three as high (positive) for valence and low (calm) for arousal to keep the data balanced as much as possible.

The illustrated bar charts in figure 4 and 5 show the number of participants who felt the desired emotions from the two different emotions models based on their self-report. Participants reported three labels for each conversation for the beginning, middle, and end sections. In these figures, we chose the label of the part with the highest intensity to evaluate if we could reach the target emotion or not. As can be seen in 4, almost all participants felt happiness when the target emotion was happiness. For other emotions, most of them felt the target emotion. For the neutral state, although most people felt a neutral emotion, a considerable number of them reported happiness. One reason could be the topic of the neutral conversation, which was about describing their feeling when they prepared breakfast, which was enjoyable for some participants. As described before, in the questionnaire, we included "other emotions" as an option. A limited number of participants felt other emotions mostly instead of neutral emotions. We ignored trials with reported emotion as others in the method evaluations.

According to Figure 5 most of the participants reported high arousal and high valence in the conversation with the happiness subject. Most of the participants were in a positive and relaxed mood during the neutral conversations. In the sad conversations, mostly, the arousal and valence levels were low. Although the fear and anger conversations were less negative than the sadness conversations, the number of participants with negative feelings was still higher than those with positive emotions. Also, in the fear and anger conversations, the level of arousal was higher. According to these figures, on average, we could reach the target emotion at least in some parts of the conversations based on the definition of two emotion models.

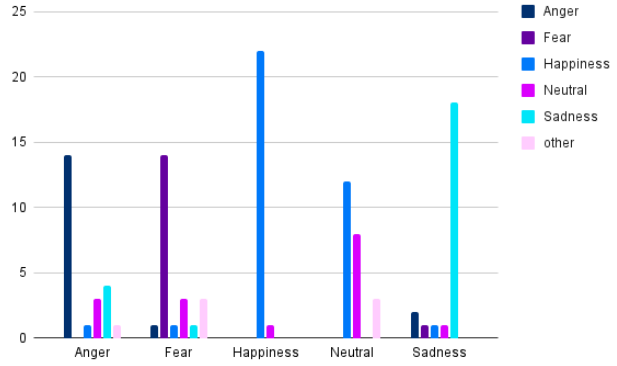

Figure 4: Comparison of self-report emotions and target emotions in the conversations for all participants. Each bar shows the number of participants who reported a special emotion for the conversation with the subject of desired emotion.

We also evaluated the content of conversations. To evaluate conversations with their content, we first converted conversational audio data to text using the Otter online tool [5]. Then we removed the sentences in which the interviewer talked and extracted the participants' sentences. Finally, we used the LIWC2015 [6] software which is a Linguistic Inquiry Word Count (LIWC) tool that counts words in psychologically meaningful categories [82] and has been widely used in psychology [4, 12]. LIWC2015 compares each word against a user-defined dictionary and reports some statistics about the text content [82]. We used the LIWC2015 built-in dictionary to extract some statistics about the emotional content of each conversation. The output of LIWC2015 is a table where each column is the percentage of the presence of words related to different categories in LIWC.

In LIWC2015, there are only positive, negative, fear, anger, sadness, and affect categories related to emotions. After measuring the LIWC statistics, we measured each conversation's emotion based

[5] https://otter.ai/
[6] https://liwc.wpengine.com/

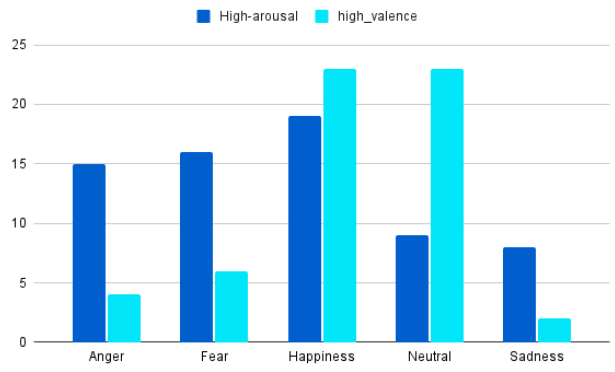

Figure 5: Comparison of the arousal and valence in the self-report questionnaire with target emotions in the conversations for all participants. The bars show the number of participants with high-arousal and high-valence levels for each desired emotion

Table 2: The result of LIWC2015 for the content of each conversation

	Positive>Negative	Anxiety	Anger	Sadness
Anger	8	1	**21**	0
Fear	8	**22**	0	0
Happiness	**23**	5	4	9
Neutral	**22**	2	6	7
Sadness	14	3	0	**20**

on valence, anger, sadness, and anxiety. We rated valence as high when the positive value was more than negative. Also, we rated conversation as sad, when the sad value was more than anger and anxiety. For anger and anxiety, we used the same process. Table 2 compares the measured values based on LIWC2015's output for all participants in different conversations. As can be seen, happiness and neutral conversations contained more positive words than negative words, whereas sad, anger and fear conversations contained more negative terms. In the case of sadness, the number of sad words was considerably higher than other emotions. For anger conversations for most participants, the number of anger words was more than others, while for fear, the words related to anxiety were increased. This result is almost directly related to the desired emotional states. In our experiment, neutral states were more positive than negative, which is the same as the reported emotional state in which some of them felt happiness.

Overall, our two evaluation results show that we could mostly achieve the target emotions for most participants. This indicates that in this experimental setup, we were able to induce emotions in a non-acted conversation.

6.2 Methodology evaluation

We used two evaluation methods, including subject-dependent and subject-independent, to evaluate each modality in recognizing emotions. In subject-dependent evaluation, we used the leave-some-trial-out approach for each participant separately. We used 3-fold cross-validation over each participant's one-minute trials and measured the f-score of the classification for discrete emotions, arousal, and valence levels. We split the data into training and test sets before windowing. We used the leave-one-subject-out cross-validation technique to break data into training and test sets in the subject-independent approach. In each iteration, we considered one participant in the test set and the remaining in training set in this method. The reported result is the average of the calculated F-score over all iterations. We used f-score to consider imbalanced data [94].

The four main metrics in evaluating models are accuracy, precision, recall, and F-Score or F1 [46]. All of the reported results in this paper are based on the F-Score because this shows the model's performance more reliably when we have imbalanced data. Equations 2 show how to calculate the evaluation metrics. In these equations, TP is True Positive which means correctly predicted positive, TN is True Negative or correctly predicted negative class, FP is False Positive or incorrectly predicted positive, and FN is True Negative or incorrectly predicted negative. We chose the F-Score for evaluating our methods which are appropriate for imbalanced data. [77].

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \quad (2)$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 = 2 \times (Recall \times Precision)/(Recall + Precision)$$

We used grid-search cross-validation parameter tuning [16] to measure the hyperparameters for RFC and LSTM . Hyper-parameters were tuned using 5-fold cross-validation when we split data based on the cross-subject technique, which scrambled all participants' trials and then randomly split in training and test datasets. We got the best result when we considered 500 estimators for RFC. For LSTM, we considered 32 as the batch size and used a reduced learning rate in the range of 0.001 to 0.0001, which drops down at the rate of 0.5 if the validation loss is not changing.

Figures 6, 7 and 8 show the result of the RFC classifier for 23 participants based on arousal (low and high), valence (low and high), and discrete emotions including anger, fear, happiness, neutral and sadness. As can be seen, recognizing emotions for some participants was more difficult than others. These figures show that human responses can be completely different and unpredictable in different situations. In some of them, EEG signals showed a different pattern in various emotional states, while for others, physiological signals were more effective in recognizing emotions. The importance of multimodal emotion recognition arises from this diversity when various modalities will overcome a single mortality's weaknesses. Based on these figures, the F-Score varied from 62% to 92% for arousal, 60% to 98% for valence and less than random to 97% for emotion when fusing all modalities. The possible reasons for this wide range could be the variety in personalities in response to stimuli and the variety in physiological responses in different people. Also, in conversations, some people move their bodies, head, and hands more than others. These movements directly affect the quality of the signal and contaminate emotional signals with signals related to movements, which may lead to a decrease in the F-Score level. Furthermore, each person's physical features affect the quality of

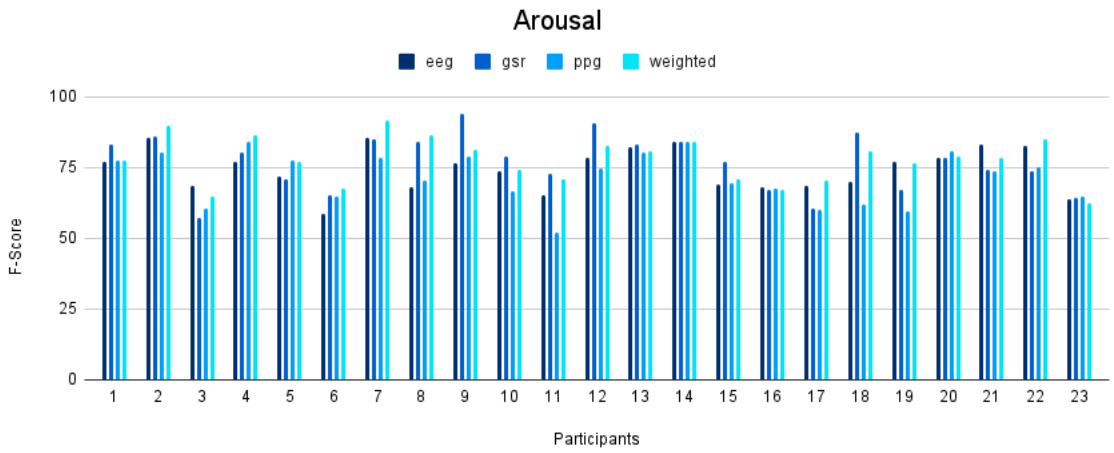

Figure 6: F-Score of the RFC method in classifying arousal levels for 23 participants

the signal. For example, the quality of the EEG signals for people who had long and thick hair was worse than others.

According to these figures, fusing various modalities always does not create the best result. Still, as shown in table 3, on average the F-Score of fusion strategy is higher than single modalities when we are using the RFC classifier. Also, recognizing emotional states based on the 2-dimensional emotion model is more accurate than recognizing emotions based on the discrete emotion model. One reason could be that all discrete emotions are related to each other. Based on the Circumplex Model [67] it is not correct to categorize them in discrete emotions because the human emotional state always is a mixture of several emotions. So, when people report fear as their emotion, their emotion may be a mixture of excitement, joy, and scared or a combination of negative feelings and fear. So, in positive and negative scary situations, the pattern of the brain and physiological signals are not the same, and categorizing them in a single class leads to incorrect recognition.

Table 3 shows the F-Score average for subject-dependent and subject-independent evaluations. As can be seen, the evaluation result of RFC in subject-dependent evaluation is considerably higher than subject-independent evaluations. In the subject-independent approach, defining a general model which can identify unseen participants' emotions is challenging because of differences in physiological responses in different individuals. We need more robust models to extract more deep features to improve these low F-Score.

For this reason, we used an LSTM model, which could extract more deep features. The improvement in LSTM F-Score compared to the RFC method is significant, at around 20% improvement for arousal, valence, and emotion levels in the subject-independent approach. This shows that although simple machine learning models

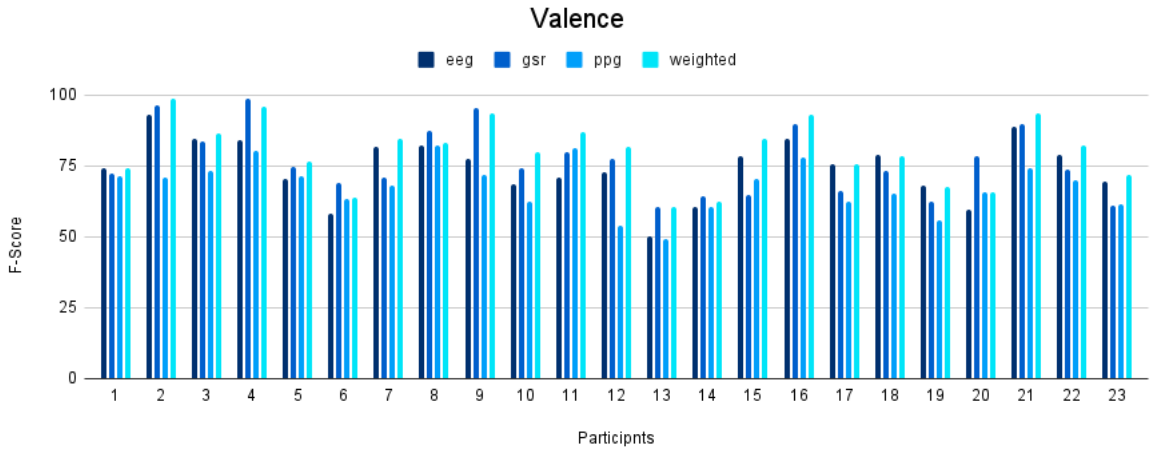

Figure 7: F-Score of the RFC method in classifying valence levels for 23 participants

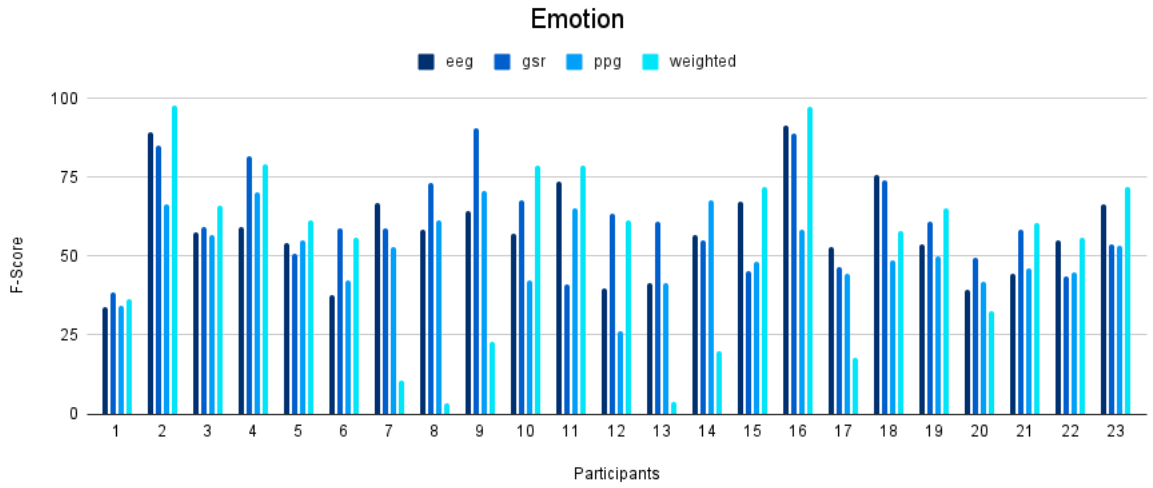

Figure 8: F-Score of the RFC method in classifying 5 emotions for 23 participants

Table 3: The average F-Score of the two strategies using the three evaluation approaches

	EEG		GSR		PPG		Fusion	
	RFC	LSTM	RFC	LSTM	RFC	LSTM	RFC	LSTM
subject-dependent								
Arousal	74.5	78.1	76.7	**79.7**	71.5	78.6	77.6	77.6
Valence	74.4	78.3	76.7	**81.1**	68	72.5	80	80.5
Emotion	58.1	56.2	61.1	51.3	51.7	44.9	52.5	**63.1**
subject-independent								
arousal	45	67.6	48.9	67.6	48.3	68	45.9	**68.6**
valence	52.4	**66**	50.4	57.5	48	60.1	51.2	64.4
emotion	22	**36.2**	19.3	26.2	19.5	25	19.5	34.8

based on hand-crafted features are effective in subject-dependent evaluation, they are not robust enough to create a general and independent model for recognizing emotions in a complex task like a conversation. However, the LSTM model could improve the recognition performance using a single modality.

In the subject-dependent approach, recognizing emotion using GSR signals was more accurate than EEG and PPG signals, and fusing all modalities improved the recognition in arousal and valence levels with 1% and 3.4% improvements to reach the average of 77.6% and 80% using RFC. We could improve the performance of all modalities using the LSTM method and achieved the highest average F-score in the GSR, which was 79.7% for arousal and 81.1% for valence. We achieved an average F-Score of 63.1% for emotion using LSTM in the fusion strategy in recognizing emotion levels.

In the subject-independent approach, EEG emotion recognition classified valence levels of unseen participants' trials more accurately than other modalities. Although the fusion strategy improved the F-score around 1% for arousal, the performance of the fusion method for valence and emotion levels was less than using only the EEG signals. To improve this in the future, we should use a more complex fusion strategy or enhance the performance of the PPG and GSR methods.

7 DISCUSSION

In this study, we conducted an experiment to induce emotions in non-spontaneous conversations. We collected brain activity and a variety of physiological signals from subjects in the experiment. We showed that although physiological signals from the emotional responses in the brain and body were contaminated with other activities like muscle movement and word production, they could still be used as effective cues in recognizing emotions. We could see a trivial improvement after fusing these signals in some cases, but it was not significant. We will explore more complex fusion strategies to overcome single modalities' weaknesses in the future.

This study is one step towards equipping intelligent human-computer interfaces with emotional intelligence. The intelligent systems that interact with humans based on the brain and physiological signals are more reliable than behavioral systems since they cannot be controlled. The designed model based on collected data in this study could be used in various activities such as creating intelligent assistant agents for helping with daily life tasks like driving, education, health and entertainment.

Although we still have a long way to go toward designing a general model for recognizing emotions, creating datasets is crucial for identifying the variety in emotional response and finding the most

reliable modalities to recognize emotion. Despite the possibility of collecting some elements of this study remotely, such as audio and video, the study's use of OpenBCI and Shimmer sesnors would make it challenging to collect data remotely or through crowdsourcing. However, the reduction in the cost of EEG hardware and the use of smartwatches or smartphones to measure physiological signals will make this possible in the future.

In this study, we faced several limitations and challenges which should be considered in future works. The following section describes these in more detail.

7.1 Limitations and Challenges

This research was subject to several limitations and challenges which could be overcome in future work. The major challenges in recognizing emotions using EEG and physiological signals while people are talking are as follows:

7.1.1 Sensor limitations. Since we use a gel-based soft EEG headset, increasing the experiment time too much will decrease the quality of signals due to the electrode gels drying out. This directly affects the quality of signals and, respectively, the recognition results. Also, we had to conduct the experiment in one session because many participants were not happy to come for data collection more than once since the electrode gel makes their hair messy. We decreased the effect of this problem by designing a short experiment scenario. We reduced the length of the experiment as much as possible by reducing the emotion set and using a continuous conversation for various emotions in one session. This issue will be addressed in the future by using newer dry EEG headsets that are more comfortable while having high-quality signals.

7.1.2 Effect of Personality. People have various personalities. Based on the "Big Five Inventory model" of personality [31], the agreeableness level is one of the factors in personality variety. People with a high level of agreeableness tend to be friendly, compassionate, and cooperative. Also, extroversion is another factor, where extroverted people are more talkative and tend to seek the company of others [32]. These varieties in personalities affect the length of the conversation. So we had various lengths for different emotions and different participants, which may lead to unbalanced data. In the future, we will attempt to control this by having participants take personality tests and use the results of these tests to filter out the collected data.

7.1.3 Effect of Mood. Human interactions and emotional experiences are directly influenced by participants' moods [21, 30, 33]. In addition, this affects how we perceive others' emotions [83]. We found that the mood of the participants directly affected their desire to talk and share their emotional state. For example, participants in a negative mood were harder to encourage to collaborate or recall happy memories and express them. Since mood lasts longer than emotion and can be influenced by previous events, we were unable to ensure the same mood for all participants. We can decrease this effect in the future by assessing participants' moods before experiments and taking their moods into account during the data analysis or filtering out the collected data.

7.1.4 Inducing spontaneous emotions. Inducing emotions in a conversation is a challenging task that can be affected by many factors like environmental context, personality, mood, and argumentation logic [62]. In this study, we tried to minimize these factors by using the same topic and interviewer for all of the participants' conversations. Also, the interviewer was a psychologist who had skills in making people relax and encouraging them to be open in feeling and expressing emotions as much as possible. Although the psychologist helped create a relaxed environment for feeling emotions, some people were still uncomfortable talking with a stranger about their feelings. This could be improved by some preparation before the study in the future. For example, a collaborative and fun task like a game between the interviewer and participant before the experiment could create more empathy between the participant and interviewer.

7.1.5 Ground truth labeling. Annotating conversational data with underlying emotions is a significant challenge. Only participants know about their underlying emotions, so labeling with other annotators may give completely different labels. We assured participants that no one would listen to their conversations to induce natural emotions. Although people may sometimes be bad at recognizing their own emotions, self-report questionnaires are more reliable in revealing hidden emotions. Still, filling out a questionnaire cannot be done in the middle of the experiment for a small chunk of data because it will distract participants and affect the conversation flow. So we used a post-experiment questionnaire to collect ground truth labels. In fact, we used a self-annotation strategy by asking participants to listen to parts of their audio and label how they were feeling at the time. This was time-consuming and sometimes not pleasant for participants to listen to all of their conversations. Due to this, we collect only three labels for each conversation to minimize the time.

We also used LIWC for automatic labeling for each minute of data. However, the performance of LIWC decreases by reducing the size of data, and it is not reliable for input with less than 50 words [82]. So one-minute conversations are not long enough to get correct results from LIWC. In the future, with the advances of deep-learning methods in Natural Language Processing (NLP) [39], we can use pre-trained robust models for labeling smaller chunks of data.

7.1.6 Multimodal data mismatching. Although we collected video and audio data in this study, we did not use them for emotion recognition since behavioral responses are sometimes different from the underlying emotions. For example, people may feel sad while their face shows neutral or even smiling expressions in non-happy situations. Also, they may not express their emotions with words precisely related to their emotions. This needs further analysis in the future to discover these mismatches and the relation of behavioral responses, brain and physiological activities. Extracting facial micro-expressions instead of regular facial expressions is another possibility for improving these mismatches in the future. Facial micro-expressions refer to quick and involuntary changes in the facial expressions, such as raising the inner eyebrows or the wrinkled nose that occur spontaneously in response to external stimuli, typically within a timeframe of 65ms and 500ms [91]. Facial

micro-expressions can be used to detect genuine emotions and are difficult to fake [78].

8 CONCLUSION AND FUTURE WORKS

There are significant applications for conversational emotion detection in intelligent user interfaces, such as interacting with robots or virtual humans or creating the most effective video-conferencing systems in health-care, education, business, and everyday life. Creating datasets in different conditions is crucial in training robust models that can recognize emotions in a conversational setting. This paper introduced an experimental setup and method for conducting a conversation to stimulate emotions similar to those in real-world conversations, and created a multimodal dataset for emotion recognition. We collected EEG and physiological signals and explained our strategy for using these modalities and fusing them for recognizing emotions. We evaluated the conversations' content and showed that we successfully stimulated emotions from a non-acted conversation. Even though there were muscle artifacts in a conversation, we could achieve an average F-Score of 77.6% and 80% respectively for arousal and valence levels in a subject-dependent approach in recognizing emotions. We reached 68% and 64% F-Score for arousal and valence levels in a subject-independent approach. Although this result for a general model is considerable, the recognition performance could be improved in the future by using more complex deep-learning models. We also discussed the limitations and challenges of creating these datasets and recognizing emotions in conversations using EEG and physiological signals and how these could be overcome in the future.

Our result shows that EEG and physiological signals can be used effectively for recognizing emotions in human-human or human-machine conversations. In our research, we are particularly interested in two areas. First, developing intelligent user interfaces for video conferencing where emotions could be recognized and shared during a video call. For example, a therapist could better understand when a client becomes emotional in a remote therapy session. Second, developing systems that include conversational digital humans who can recognize and respond to emotion in human/agent conversations. In the future, we will use our dataset (*PEGCONV*) directly in training these intelligent systems to provide emotion recognition abilities. For example, it could be used in humanoid robots or virtual assistants to understand human behavior more accurately and respond appropriately. We plan to demonstrate this with a digital human in future work. We also plan to use video and audio data combined with EEG and physiological signals and exploit behavioral and physiological modalities to recognize emotion in conversations. We anticipate that the data collection and processing methods that we demonstrated, and the dataset we have, will be helpful for a wide range of intelligent user interface applications.

REFERENCES

[1] Soraia M Alarcao and Manuel J Fonseca. 2017. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing* 10, 3 (2017), 374–393.

[2] Morgan Anvari, Clara E Hill, and Dennis M Kivlighan. 2020. Therapist skills associated with client emotional expression in psychodynamic psychotherapy. *Psychotherapy Research* 30, 7 (2020), 900–911.

[3] Renan Vinicius Aranha, Cléber Gimenez Corrêa, and Fátima LS Nunes. 2019. Adapting software with affective computing: a systematic review. *IEEE Transactions on Affective Computing* 12, 4 (2019), 883–899.

[4] Alexandra Balahur, Saif Mohammad, Veronique Hoste, and Roman Klinger. 2018. Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

[5] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. 2003. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.. In *2003 Conference on computer vision and pattern recognition workshop*, Vol. 5. IEEE, 53–53.

[6] Hayfa Blaiech, Mohamed Neji, Ali Wali, and Adel M Alimi. 2013. Emotion recognition by analysis of EEG signals. In *13th International Conference on Hybrid Intelligent Systems (HIS 2013)*. IEEE, 312–318.

[7] George Boateng. 2020. Towards Real-Time Multimodal Emotion Recognition among Couples. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 748–753.

[8] Patricia J Bota, Chen Wang, Ana LN Fred, and Hugo Plácido Da Silva. 2019. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access* 7 (2019), 140990–141020.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[10] Adrian Burns, Emer P Doheny, Barry R Greene, Timothy Foran, Daniel Leahy, Karol O'Donovan, and Michael J McGrath. 2010. SHIMMER™: an extensible platform for physiological signal capture. In *2010 annual international conference of the IEEE engineering in medicine and biology*. IEEE, 3759–3762.

[11] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[12] Flavio Carvalho, Gabriel Santos, and Gustavo Paiva Guedes. 2018. AffectPT-br: an Affective Lexicon based on LIWC 2015. In *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*. IEEE, 1–5.

[13] Valentina Chaparro, Alejandro Gomez, Alejandro Salgado, O Lucia Quintero, Natalia Lopez, and Luisa F Villa. 2018. Emotion Recognition from EEG and Facial Expressions: a Multimodal Approach. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 530–533.

[14] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[15] Abhishek Chunawale, Dr Bedekar, et al. 2020. Human Emotion Recognition using Physiological Signals: A Survey. *Available at SSRN 3645402* (2020).

[16] Marc Claesen and Bart De Moor. 2015. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127* (2015).

[17] Didar Dadebayev, Goh Wei Wei, and Tan Ee Xion. 2021. EEG-based Emotion Recognition: Review of Commercial EEG Devices and Machine Learning Techniques. *Journal of King Saud University-Computer and Information Sciences* (2021).

[18] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.

[19] Mohamed Elgendi. 2012. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews* 8, 1 (2012), 14–25.

[20] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5562–5570.

[21] Joseph P Forgas, Gordon H Bower, and Susan E Krantz. 1984. The influence of mood on perceptions of social interactions. *Journal of Experimental Social Psychology* 20, 6 (1984), 497–513.

[22] Judith E Glaser and Ross Tartell. 2014. Conversational Intelligence at work. *OD Practitioner* 46, 3 (2014), 62–67.

[23] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.

[24] Han-Wen Guo, Yu-Shun Huang, Chien-Hung Lin, Jen-Chien Chien, Koichi Haraikawa, and Jiann-Shing Shieh. 2016. Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 274–277.

[25] Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2021. Enhancing the Perceived Emotional Intelligence of Conversational Agents through Acoustic Cues. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[26] Xin Hu, Jingjing Chen, Fei Wang, and Dan Zhang. 2019. Ten challenges for EEG-based affective computing. *Brain Science Advances* 5, 1 (2019), 1–20.

[27] Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahui Pan. 2017. Fusion of facial expressions and EEG for multimodal emotion recognition. *Computational intelligence and neuroscience* 2017 (2017).

[28] Maryam Imani and Gholam Ali Montazer. 2019. A survey of emotion recognition methods with emphasis on E-Learning environments. *Journal of Network and Computer Applications* 147 (2019), 102423.

[29] S Jerritta, M Murugappan, R Nagarajan, and Khairunizam Wan. 2011. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. IEEE, 410–415.

[30] Rajiv Jhangiani, Hammond Tarry, and Charles Stangor. 2014. Principles of social psychology-1st international edition. (2014).

[31] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of Personality and Social Psychology* (1991).

[32] Oliver P John, Sanjay Srivastava, et al. 1999. *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*. Vol. 2. University of California Berkeley.

[33] Avery Keith. 2019. Influence of Mood on Language Use in Dyadic Social Interaction. (2019).

[34] Jihen Khalfallah and Jaleleddine Ben Hadj Slama. 2015. Facial expression recognition for intelligent tutoring systems in remote laboratories platform. *Procedia Computer Science* 73 (2015), 274–281.

[35] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[36] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[37] M KOŁODZIEJ, P Tarnowski, A Majkowski, and RJ Rak. 2019. Electrodermal activity measurements for detection of emotional arousal. *Bull. Pol. Ac.: Tech* 67 (2019), 4.

[38] Sylvia D Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological psychology* 84, 3 (2010), 394–421.

[39] Akshay Kulkarni and Adarsha Shivananda. 2021. Deep learning for NLP. In *Natural language processing recipes*. Springer, 213–262.

[40] Min Seop Lee, Yun Kyu Lee, Dong Sung Pae, Myo Taeg Lim, Dong Won Kim, and Tae Koo Kang. 2019. Fast Emotion Recognition Based on Single Pulse PPG Signal with Convolutional Neural Network. *Applied Sciences* 9, 16 (2019), 3355.

[41] Heather C Lench, Sarah A Flores, and Shane W Bench. 2011. Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitations. *Psychological bulletin* 137, 5 (2011), 834.

[42] Shan Li and Weihong Deng. 2018. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348* (2018).

[43] Zhen Liang, Shigeyuki Oba, and Shin Ishii. 2019. An unsupervised EEG decoding system for human emotion recognition. *Neural Networks* 116 (2019), 257–268.

[44] Jia Zheng Lim, James Mountstephens, and Jason Teo. 2020. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors* 20, 8 (2020), 2384.

[45] Kate Loveys, Gregory Fricchione, Kavitha Kolappa, Mark Sagar, and Elizabeth Broadbent. 2019. Reducing patient loneliness with artificial agents: design insights from evolutionary neuropsychiatry. *Journal of medical Internet research* 21, 7 (2019), e13664.

[46] John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. 1999. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*. Herndon, VA, 249–252.

[47] Samuel Marcos-Pablos, Emilio González-Pablos, Carlos Martín-Lorenzo, Luis A Flores, Jaime Gómez-García-Bermejo, and Eduardo Zalama. 2016. Corrigendum: Virtual Avatar for Emotion Recognition in Patients with Schizophrenia: A Pilot Study. *Frontiers in human neuroscience* 10 (2016), 554.

[48] Tomas Matlovic, Peter Gaspar, Robert Moro, Jakub Simko, and Maria Bielikova. 2016. Emotions detection using facial expressions recognition and EEG. In *2016 11th international workshop on semantic and social media adaptation and personalization (SMAP)*. IEEE, 18–23.

[49] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing* 3, 1 (2011), 5–17.

[50] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1359–1367.

[51] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[52] Jermana L Moraes, Matheus X Rocha, Glauber G Vasconcelos, José E Vasconcelos Filho, Victor Hugo C De Albuquerque, and Auzuir R Alexandria. 2018. Advances in photopletysmography signal analysis for biomedical applications. *Sensors* 18, 6 (2018), 1894.

[53] Ashritha R Murthy and KM Anil Kumar. 2021. A Review of Different Approaches for Detecting Emotion from Text. In *IOP Conference Series: Materials Science and Engineering*, Vol. 1110. IOP Publishing, 012009.

[54] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access* 7 (2019), 19143–19165.

[55] Standard Electrode Position Nomenclature. 1991. American electroencephalographic society guidelines for. *Journal of clinical Neurophysiology* 8, 2 (1991), 200–2.

[56] Fatemeh Noroozi, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *IEEE transactions on affective computing* (2018).

[57] François Perrin, J Pernier, O Bertrand, and JF Echallier. 1989. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology* 72, 2 (1989), 184–187.

[58] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[59] Thammathip Piumsomboon, Youngho Lee, Gun A Lee, Arindam Dey, and Mark Billinghurst. 2017. Empathic mixed reality: Sharing what you feel and interacting with what you see. In *2017 International Symposium on Ubiquitous Virtual Reality (ISUVR)*. IEEE, 38–41.

[60] Stephen W Porges. 2009. The polyvagal theory: new insights into adaptive reactions of the autonomic nervous system. *Cleveland Clinic journal of medicine* 76, Suppl 2 (2009), S86.

[61] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

[62] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* 7 (2019), 100943–100953.

[63] Karen S Quigley, Kristen A Lindquist, and Lisa Feldman Barrett. 2014. Inducing and measuring emotion and affect: Tips, tricks, and secrets. (2014).

[64] Fuji Ren and Yanwei Bao. 2020. A review on human-computer interaction and intelligent robots. *International Journal of Information Technology & Decision Making* 19, 01 (2020), 5–47.

[65] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 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)*. IEEE, 1–8.

[66] Nicole A Roberts, Jeanne L Tsai, and James A Coan. 2007. Emotion elicitation using dyadic interaction tasks. *Handbook of emotion elicitation and assessment* (2007), 106–123.

[67] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[68] Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining* 8, 1 (2018), 28.

[69] Anvita Saxena, Ashish Khanna, and Deepak Gupta. 2020. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems* 2, 1 (2020), 53–79.

[70] Theresa Schachner, Roman Keller, and Florian Von Wangenheim. 2020. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. *Journal of medical Internet research* 22, 9 (2020), e20701.

[71] Djoko Budiyanto Setyohadi, Sri Kusrohmaniah, Sebastian Bagya Gunawan, Pranowo Pranowo, and Anton Satria Prabuwono. 2018. Galvanic skin response data classification for emotion detection. *International Journal of Electrical and Computer Engineering* 8, 5 (2018), 4004.

[72] Fred Shaffer and Jay P Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* 5 (2017), 258.

[73] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.

[74] Ewa Siedlecka and Thomas F Denson. 2019. Experimental methods for inducing basic emotions: A Qualitative review. *Emotion Review* 11, 1 (2019), 87–97.

[75] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2011), 42–55.

[76] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 843–852.

[77] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. 2009. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence* 23, 04 (2009), 687–719.

[78] Madhumita Takalkar, Min Xu, Qiang Wu, and Zenon Chaczko. 2018. A survey: facial micro-expression recognition. *Multimedia Tools and Applications* 77, 15 (2018), 19301–19325.

[79] Jianhua Tao and Tieniu Tan. 2005. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*. Springer, 981–995.

[80] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. 2020. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing* (2020).

[81] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz Jan Rak. 2018. Combined analysis of GSR and EEG signals for emotion recognition. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*. IEEE, 137–141.

[82] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[83] Irene Trilla, Anne Weigand, and Isabel Dziobek. 2021. Affective states influence emotion perception: evidence for emotional egocentricity. *Psychological research* 85, 3 (2021), 1005–1015.

[84] Goran Udovičić, Jurica Đerek, Mladen Russo, and Marjan Sikora. 2017. Wearable emotion recognition system based on GSR and PPG signals. In *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*. 53–59.

[85] Paul van Gent, Haneen Farah, N Nes, and Bart van Arem. 2018. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. In *Proceedings of the 6th HUMANIST Conference*. 173–178.

[86] Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* 10, 5 (1999), 988–999.

[87] B Venkatesh and J Anuradha. 2019. A review of feature selection and its methods. *Cybernetics and Information Technologies* 19, 1 (2019), 3–26.

[88] Gyanendra K Verma and Uma Shanker Tiwary. 2014. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage* 102 (2014), 162–172.

[89] Kalyani P Wagh and K Vasanth. 2019. Electroencephalograph (EEG) based emotion recognition system: A review. *Innovations in Electronics and Communication Engineering* (2019), 37–59.

[90] Haolin Wei, David S Monaghan, Noel E O'Connor, and Patricia Scanlon. 2014. A New Multi-modal Dataset for Human Affect Analysis. In *International Workshop on Human Behavior Understanding*. Springer, 42–51.

[91] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior* 37, 4 (2013), 217–230.

[92] Wenlu Yang, Maria Rifqi, Christophe Marsala, and Andrea Pinna. 2018. Physiological-based emotion detection and recognition in a video game context. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[93] Ashkan Yazdani, Jong-Seok Lee, Jean-Marc Vesin, and Touradj Ebrahimi. 2012. Affect recognition based on physiological changes during the watching of music videos. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 1 (2012), 7.

[94] Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. 2017. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine* 140 (2017), 93–110.

[95] Hmamouche Youssef, Laurent Prevot, Ochs Magalie, and Chaminade Thierry. 2020. Identifying Causal Relationships Between Behavior and Local Brain Activity During Natural Conversation. In *Interspeech 2020*. ISCA, 101–105.

[96] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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.

[97] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2008. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2008), 39–58.

[98] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W Picard. 2020. Driver emotion recognition for intelligent vehicles: a survey. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–30.

[99] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics* 49, 3 (2018), 1110–1122.

[100] Qingyang Zhu, Guanming Lu, and Jingjie Yan. 2020. Valence-Arousal Model based Emotion Recognition using EEG, peripheral physiological signals and Facial Expression. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*. 81–85.