IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 August 2023, accepted 21 September 2023, date of publication 27 September 2023, date of current version 5 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320053

RESEARCH ARTICLE

PhyMER: Physiological Dataset for Multimodal Emotion Recognition With Personality as a Context

SUDARSHAN PANT, HYUNG-JEONG YANG, EUNCHAE LIM, SOO-HYUNG KIM, AND SEOK-BONG YOO

Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Hyung-Jeong Yang (hjyang@jnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] under Grant RS-2023-00219107, in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) through the Artificial Intelligence Convergence Innovation Human Resources Development grant funded by the Korea Government (MSIT) under Grant IITP-2023-RS-2023-00256629, and in part by the Regional Innovation Strategy (RIS) through NRF funded by the Ministry of Education (MOE) under Grant 2021RIS-002.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board, Chonnam National University under Application No. 1040198-210401-HR-045-02.

ABSTRACT Physiological signals are widely used in the recognition of affective status. Recording of such physiological signals involves elicitation of emotions through different stimuli including video-based stimulus. Considering that the same stimulus videos often induce different emotions in different individuals, emotion recognition in such a scenario requires consideration of the individual differences in the consumption of the stimulus content. With this as our goal, we present a Physiological dataset for Multimodal Emotion Recognition (PhyMER) for studying emotion through physiological response with personality as a context. The PhyMER dataset consists of electroencephalogram (EEG), electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature along with the personality traits of 30 participants. We collected the video-based stimulus dataset for emotion elicitation and developed a web-based annotation tool for labeling felt emotions. We compared the stimulus labels and the self-annotation of felt emotions labeled during physiological data recording. Correlation among personalities was analyzed to study the impact of personality on the intensity of emotions in arousal and valence dimensions. Finally, we proposed a baseline model for the classification of emotions using physiological signals. The dataset is publicly available to the academic community for analysis of affective states and the development of emotion recognition models.

INDEX TERMS Physiological signals, EEG, emotion classification, personality traits.

I. INTRODUCTION

Emotions are behavioral phenomena that occur in response to an event or stimulus and are expressed through various behavioral and physiological changes. Emotions have been studied as discrete categories, continuous values in various dimensions such as arousal, valence, and dominance, and in terms of changes in a set of components based on different subjective qualities. Ekman [1] introduced seven basic emotions which are widely accepted as emotions independent of race, culture, or geography [2], [3]. Similarly, Parrot's [4] tree structure of emotions is a popular example of a discrete representation of emotions where emotions are hierarchically organized as primary, secondary, and tertiary emotions. Plutchick [5] organized the discrete emotions in a wheel, known as Plutchick's wheel of emotion, with 8 basic emotions towards the center and fine-grained sub-categories towards the edge of the wheel. On the other hand, the

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

dimensional view of emotions refers to the organization of emotions as continuous values across dimensions such as arousal, valence, and dominance, where the emotional states are interrelated systematically. For instance, Russel's Circumplex Model of Affect [6] includes emotions in the dimensions of arousal and valence. Arousal represents the degree of an individual's excitement while valence indicates the level of pleasant or unpleasant feelings. The component process model [7] is based on the coordinated changes in the individual in terms of components such as appraisal, motivation, physiology, and expression. Several studies such as [8], [9], and [10] demonstrated the multi-componential emotions in different scenarios. Mohammadi and Vuilleumier [10] used a component model to show the role of personality in the recognition of discrete emotions. The theory of constructed emotion [11] states that emotions are invisibly constructed by the brain based on the situation. The body-budgeting regions in the brain predict the experienced world by tweaking the neurons based on past experiences and such predictions are sent to the rest of the body controlling the physiological processes such as heart beats and respiratory rate [12].

Emotions are felt and expressed in a significantly different manner among individuals [13]. Similarly, electroencephalogram (EEG) signals, one of the physiological signals used in emotion recognition, are variable across individuals [14], [15], [16]. Therefore, physiological datasets with precise and fine-grained labels of such emotions in consideration of individual differences are essential for the accurate analysis of human emotions.

One way to take individual differences into account is through personality. However, only a few studies have considered personality during emotion recognition. Emotions generate both behavioral and physiological changes which include variations in facial expression, posture, or alterations in physiological activities such as heartbeat, neural activations, perspiration, and body temperature. Unfortunately, among the existing physiological emotion datasets, only a few consider individual personality as a crucial factor in identifying emotions [17], [18], [19]. Most of the existing datasets either use one of the categorical or dimensional emotions or have coarse annotation for different levels of emotion dimensions. The annotations from both categorical and dimensional perspectives are important for emotion recognition research. Moreover, to study individuals' emotions more precisely, a multimodal dataset labeled with fine-grained emotions is required.

In this paper, we present a Physiological dataset for Multimodal Emotion Recognition (PhyMER) that encompasses a wide range of physiological signals recorded during video viewing as emotional stimulus. Physiological signals are not only obtained from visual or physical evaluation but can also be obtained through off-the-shelf non-invasive consumer-grade devices which offer a convenient and cost-effective means of acquiring physiological data. Although basic emotions are considered universal, the observation and consumption of the stimulus video content may differ among individuals. Therefore, we include personality traits in the dataset to provide individual context for emotion recognition.

The PhyMER dataset consists of physiological signals obtained from 30 Korean participants (15 male and 15 female) using two different wearable devices. To avoid the bias due to the stimulus comprehension the participants with similar age group were selected. The participants were university students aged between 20 and 30 years. The dataset was collected with video-based emotion stimuli, where the participants watched 23 stimulus videos of varying lengths (1 to 3 minutes). The modalities collected include EEG, blood pulse volume (BVP), electrodermal activity (EDA), and skin temperature (TEMP). A custom annotation tool was developed for self-assessment of the felt emotions and for recording the experiment times and emotion annotations to synchronize the signals collected from different devices. To ensure the quality of the dataset, two annotation experiments were conducted. Firstly 28 evaluators labeled the stimulus videos to verify if the stimulus videos collected by the experimenters could induce the expected emotions in the participants. Secondly, the participants of the physiological data collection experiment labeled their felt emotions while they watched the stimulus videos. The physiological signals collected using commercial equipment as well as personality traits are publicly available for academic research. The main contributions proposed by this paper are as follows.

- We present a physiological signal dataset with multiple physiological signals collected from 30 participants. We recorded EEG, EDA, BVP, and temperature information with annotations in both categorical and dimensional views along with the individual personality traits of the participants for the study of emotions in presence of individual personality differences.
- We present the analysis of the emotion elicitation following a video-based stimulus. We conducted experiments to analyze both stimulus and physiological signal annotations using two similar experiments involving participants of similar age groups. The video stimulus data based on Korean movies were collected and evaluated using inter-rater agreement analysis. Moreover, we analyzed the correlation between the felt emotions and personality traits to see how different personality traits affect emotion elicitation.
- We present an emotion recognition framework as a baseline method for the classification of seven basic emotions and the prediction of arousal and valence values. In this case, we performed both subject-dependent and subject-independent experiments for classification and prediction.

The rest of the paper is organized as follows; in section II, we discuss the existing studies on multimodal emotion recognition, emotion recognition datasets, and personality as a context. Section III describes the overall dataset-building process, including criteria for participant selection,

experiment scenario, the stimulus video selection process, and devices used for the data collection. Section IV provides an overview of the dataset and the statistical analysis of the dataset. In section V, we explain the classification and regression experiments for basic emotions and dimensional emotion values respectively. In section VI, we discuss the contribution and potential applications of the dataset, limitations, and future research directions. Finally, we conclude the paper in section VII.

II. RELATED WORK

Several research studies on emotion recognition using biosignal data have been conducted over the past few years. In this section, we have summarized related studies involving physiological datasets and personality as a context.

The publicly available datasets have enabled rapid advancement in emotion recognition research. Over the past few years, several datasets based on a variety of modalities have been published. In this section, we review emotion recognition datasets involving the use of physiological signals, which are relevant to this study.

SJTU emotion EEG dataset (SEED) [20] dataset includes EEG recordings collected from 15 subjects while they watched movie clips for emotion elicitation. It includes 64-channel EEG data annotated for negative (−1), neutral (0), and positive (+1) emotions. MAHNOB-HCI [21] is a multimodal dataset with EEG, Electrocardiogram (ECG), Galvanic Skin Response (GSR), Respiration Amplitude (RA), and Skin Temperature collected from 27 participants stimulated using video clips and images. It includes self-annotated labels for arousal, valence, and dominance dimensions on a 5-point scale, and 20 emotional categories. DEAP dataset [22] includes physiological signals including EEG, ECG, electrooculogram (EOG), RA, GSR, blood volume, skin temperature, and electromyogram (EMG) of 32 participants stimulated by watching music videos. The annotations include 4 categorical emotions (neutral, sad, fear, happy) and 5-dimensional (arousal, valence, liking, dominance, familiarity) emotion labels.

Similarly, DREAMER [23] dataset consists of EEG and ECG signals collected from 23 participants stimulated by 18 stimulus videos. It is labeled for arousal, valence, and dominance dimensions on the 5-point scale. The DECAF database [24] includes multiple modalities including, ECG, EOG, EMG, MEG, and near-infrared (NIR) video. It includes self-reported scores for valence, arousal, and dominance. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis (BP4D+) [25] consists of multiple modalities, including 2D and 3D videos, thermal scans, Respiration, Blood pressure, GSR, and heart rate. It was collected from 140 participants who self-reported felt emotions of 10 discrete emotions elicited using various tasks such as interviews, watching videos, pain induction using ice, and smelly odor. KEmoCon [26] includes audio, video, and physiological signals including EEG, ECG, EDA, BVP, and TEMP during paired debates on a political topic. Emotions were labeled with 20 discrete emotions and arousal and valence dimensions on a 5-point scale.

Different studies have interpreted the context in diverse manners, encompassing aspects such as multimodality, inter-agent relationships within the scene, socio-cultural dynamics, and personality [19], [27]. Emotion expression is different in individuals as it is affected by several factors, including personality [28]. Personality refers to human characteristics which explain or predict individuals' behavior [29]. The relationship of personality with various emotional states has been studied in the past, for example, the use of the personality model with a textual modality for emotion reasoning [30], [31]. Considering the potential variations in psychophysiological changes among individuals, incorporating personality into the analysis of physiological signals for emotion recognition can offer supplementary contextual information. Personality traits have been found to have an impact on perception, causing different reactions to emotional perception [10]. It is essential to examine the connection between personality and emotions to understand how personality traits influence emotional experiences. The widely used Big-5 personality trait model offers a valuable approach for identifying and characterizing human personality, comprising five key qualities: extraversion, neuroticism, conscientiousness, agreeableness, and openness. The commonly used personality assessment method includes Neuroticism, Extraversion, and Openness Five-Factor Inventory (NEO-FFI) [32], the Goldberg Adjectives Scale [33], and Newcastle Personality Assessment (NPA) [34].

AMIGOS [18] dataset uses personality and mood information for emotion recognition of individuals and groups. It was collected from 40 participants while they watched 16 emotional video clips from several movies. It consists of 7 discrete emotions, 5-point annotations of arousal, valence, and dominance, and binary labels for Liking and familiarity. The results showed weak linear correlations between emotions and personality. Personality has been used for behavior analysis in Mission Survival II corpus dataset [35], which consists of video and audio data labeled with task area functional roles and socio-emotional functional roles. The data was collected during meetings of 4 participants. The personality information was obtained using the Ten Item Personality Inventory [36]. The analysis showed the correlation between extraversion and audio features such as pitch and energy, indicating the need for further research in emotion recognition in presence of personality traits. Similarly, ASCERTAIN [17] includes multiple physiological signals such as EEG, ECG, GSR, and facial videos. It was collected from 58 participants while they watched short movie clips. Personality information was recorded through a big-5 marker scale personality questionnaire. The study showed there is a weak correlation between emotions and personality. MEmoR [19] dataset includes personality information of TV characters to reason the emotions. It is a multimodal dataset with video, audio, and text modalities focusing on emotion reasoning based on contextual information. This

TABLE 1. Summary of the existing emotion recognition datasets based on physiological signals.

Dataset	Participants	EEG Channels	Modalities	Stimuli	No. of Stimulus videos	Annotations	Personality
SEED [20]	15	62	EEG	Movie clips	6	Negative, Neutral, Positive	No
DEAP [22]	32	32	EEG, EOG, EMG, GSR, RA, BVP, Temp, Video	Music videos	40	Neutral, Sad, Fear, Happy Arousal, Valence, liking and dominance (1-9)	No
DREAMER [23]	23	14	EEG, ECG	Movie clips	18	Arousal, Valence, Dominance (1-5)	No
KEmoCon [26]	32	2	ECG, EDA, Audio, Video, Skin Temperature, hand motion	Conversation (debate)	-	20 discrete emotions, Arousal Valence (1-5)	Yes
MAHNOB-HCI [21]	27	32	EEG, GSR, RA, Skin Temperature, Video, Audio, Eye-gaze data	Videos + questionnaire	20	Arousal and valence (1-5) 20 emotional categories	No
ASCERTAIN [17]	58	3	EEG, Video, GSR, ECG	Movie clips	36	Arousal, valence, dominance, predictability (1-9)	Yes
DECAF [24]	30	-	MEG, NIR, EMG, EOG, ECG	Movie and music	76	Arousal, Valence, Engagement, Liking, Familiarity	No
AMIGOS [18]	40	14	Video	Videos	20	Pleasant, Unpleasant, calm, Excited Valence, Arousal, Control, Familiarity, like/dislike	Yes
BP4D+ [25]	140	-	Video (2D and 3D), thermal scans, GSR, RESP, BP, HR	Videos, physical stimulus, questionnaire	1	13 discrete emotion classes (relaxed, in pain, happy, disgusted, skeptical, embarrassed, nervous, scared, sad, frustrated, angry, startled, surprised)	No
PhyMER (Proposed)	30	14	EEG, EDA, BVP, Temperature	Movie clips	23	Arousal and Valence (1-9) 7 discrete emotions (Angry, Happy, Sad, Surprise, Neutral, Fear, Disgusting)	Yes

dataset consists of categorical labels (8 primary emotions and 24 fine-grained emotions), labeled by multiple annotators.

While several datasets use physiological signals and personality as a context, the existing datasets either deal with categorical emotions or provide coarse annotations in dimensional space. Therefore, in the present study, we present the physiological dataset with both discrete emotions and 9-point ratings on arousal and valence dimensions for increased precision of measurement unit and consideration of sensitivity towards the changes in physiological signals. A 9-point scale not only measures more precise levels of emotions but also allows for finer distinctions between small changes in physiological signals. Moreover, in this study, we focus on emotion recognition of Koreans as induced by the video content in native language of the participants which ensures the better elicitation of the emotions. We collected physiological data set after evaluation of the stimulus dataset by the evaluators from same culture and age group. The characteristics of the existing databases related to this paper have been summarized in Table 1.

III. DATASET BUILDING

The dataset construction was approved by Chonnam National University Institutional Review Board (IRB); a dataset construction protocol and consent form containing the information on the data collection procedure, the purpose of data collection, and the type of data to be collected was approved by IRB. The participants were briefed on the overall experiment both in verbal and written form before signing the consent documents. The participants provided written consent for the disclosure of the physiological signals as a public dataset. However, the data did not include any Personally Identifiable Information (PII) such as audio-visual information. For statistical purposes, only the age and gender of candidates are published along with the anonymized dataset. The overall experiment was conducted in three steps; selection of the stimulus video by two experimenters, annotation of the stimulus videos by 28 evaluators, and collection of physiological data from 30 participants as discussed in the following sections as shown in Fig 1.

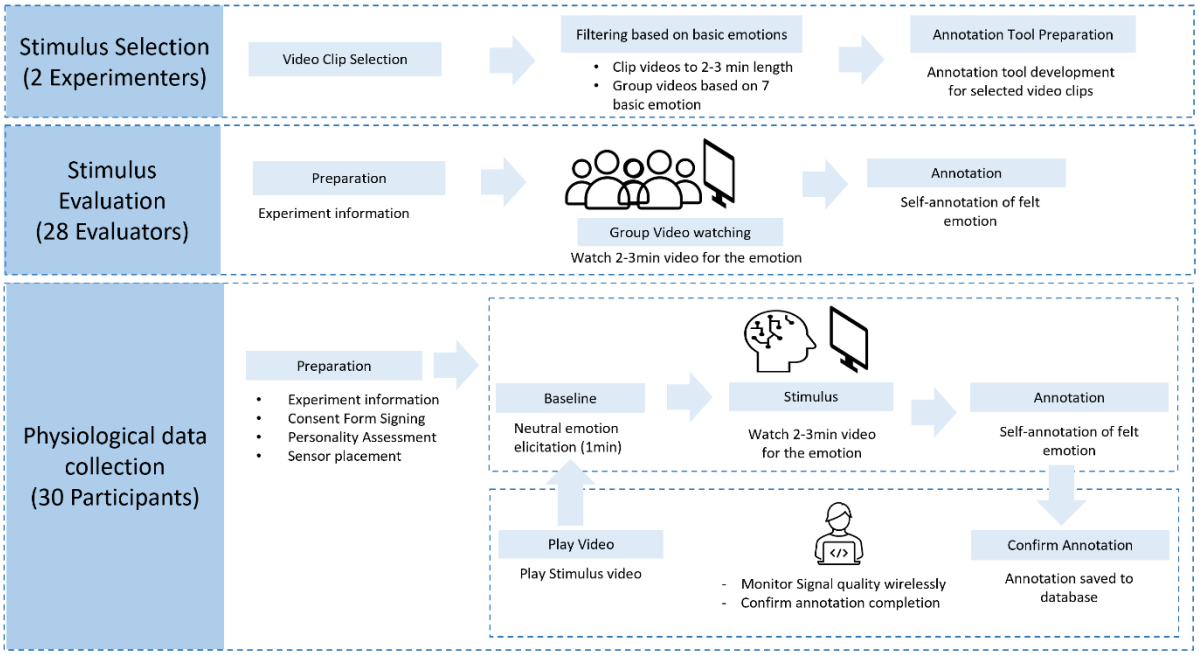

FIGURE 1. Overall data collection process for physiological data collection.

A. STIMULUS DATA

As the reliability of the dataset depends primarily on the elicitation of emotions, the selection of the stimulus is a crucial step in dataset building. In this work, we prepared a stimulus video dataset for emotion stimulation and evaluated it using a multi-rater annotation. We choose movie clips as the emotional stimuli as they are highly effective in evoking emotions [17], [24], [37]. As the experiment was conducted on Korean subjects, we decided to use Korean videos as a stimulus for emotion elicitation to avoid any issues in perceiving the content due to linguistic and cultural differences. Initially, we acquired the video clips from Korean Video Dataset for Emotion Recognition in the Wild (KVDERW) [38] based on 7 basic emotions (Happy, Sad, Angry, Surprise, Fear, Neutral, and Disgust). The KVDERW dataset is designed for emotion recognition using facial expressions in the scene, and the length of the stimulus videos is less than 10 seconds. Despite their brevity, we chose the KVDERW dataset for two main reasons. Firstly, the videos were in the Korean language, which is the native language of our study participants. Secondly, the dataset allowed us to extend the video clips to our desired duration since they were sourced from movies. As the KVDERW dataset was constructed using Korean movie clips, we searched on the web for the extended video clips for those clips and collected 25 video clips based on the availability of clearly distinguishable target segments with a single emotion. To complement the stimulus video set, five additional video clips were obtained from YouTube based on emotion-related tags on the video clips. The emotion-related keywords were prepared based on the basic emotion labels and were searched on YouTube. Five movie clips tagged with such emotion-related keywords were selected based on the observation of two researchers. We trimmed the clips to approximately 2 minutes each to ensure they contained a single emotion. The trimming process was determined by identifying the start and end of the target scene, and initially confirmed to have a single emotion by two experimenters through observation. Consequently, the clip durations varied, ranging from 61 to 122 seconds.

Two researchers selected the initial video sets with 30 clips through visual observation, and the clips were further validated through multi-rater validation. To ensure the quality and appropriateness of the clips, we sought validation from 28 evaluators, comprising 15 males and 13 females, with ages ranging between 20 and 26 years, and a mean age of 23.18 years. To maintain consistency in emotion perception, we specifically recruited evaluators from the same age group (20-30 years) as the participants involved in recording physiological data for the stimulus video evaluation. This approach aimed to ensure that emotions were perceived in a similar manner across both sets of evaluators. Out of the 30 videos, five videos were randomly selected and held out to demonstrate the process and familiarize the evaluators with the annotation task. For the evaluation of the videos, 28 evaluators watched the videos in a group. The stimulus videos were displayed on a 60-inch screen in an auditorium where the evaluators were seated with enough spacing to prevent any interaction among the evaluators. Each of the 30 videos, including 5 test videos, were annotated by evaluators for 7 basic emotions and the arousal and valence dimensions using a 9-point continuous scale. The SAM interface was

used to annotate the dimensional emotions on a scale of 1-9. The evaluators were asked to label 5 test videos at the start of the annotation process to familiarize themselves with the annotation interface. During the annotation of these test videos, the evaluators interacted with the experimenter and asked questions, and the videos were frequently paused. This interaction helped the evaluators understand the annotation process and ensured that the annotation of the remaining 25 videos was smooth and uninterrupted. As our goal is to collect the physiological signals labeled with the emotions felt by the participants, we instructed the evaluators to annotate with the emotions felt by themselves rather than the emotions exhibited by the actors in the clips. Fig. 2 shows the screenshots from the randomly selected clips from each category to illustrate the type of content present in the stimulus videos.

Finally, the labels provided by the evaluators on 25 videos were checked for agreement by calculating the percentage of evaluators who agreed on a single emotion. Based on this agreement, as well as the agreement with the original labels in the KVDERW dataset, 23 out of the 25 clips were selected as emotion stimulus videos. Table 2. shows the average arousal and valence labels for stimulus videos grouped by corresponding basic emotions. The percentage of the agreement was observed to verify how well the emotions were labeled for the expected emotions. For videos with highly contrasting views among the evaluators, the video labels were compared with the original labels in KVDERW dataset, as the selection of the emotions based only on the percentage of the agreement would lead to the inclusion of the videos which are likely to have content with multiple emotions. It is possible that the same video clip can evoke slightly different emotions in different people, due to the influence of their personality traits. As shown in Table 2., the videos marked with an asterisk (*) were not included in the stimulus dataset for emotion elicitation due to a lack of dominant consensus among evaluators. The clips needed to have agreement on non-contrasting emotions. For instance, although the VID06 had 50% agreement for sad, 39.3% of evaluators voted for anger which was the expected emotion. Similarly, VID22 (Introduction) was selected by the experimenters as a stimulus for 'happy' which was labeled as happy by 46.43% of the evaluators, while 53.57% of the evaluators labeled it as neutral. In such a case, the stimulus videos were excluded for having the possibility of eliciting contrasting emotions which are not close in arousal and valence space. Contrary to this, the videos with emotions closely situated in arousal-valence space, for example, anger and fear, were not excluded. For example, VID21 was not excluded despite having a low agreement percentage of 32.1% for anger because the second most highly annotated emotion was fear with 28.6%. Videos such as VID12 and VID23, despite having conflicting emotion labels, were included in the study based on majority voting with agreement of 42.86%.

B. PARTICIPANT SELECTION

Thirty participants between 20 to 30 years of age, (mean 23.56 years; 15 males, 15 females) were recruited two weeks before the beginning of the data acquisition experiments. All the participants were students from Chonnam National University. To determine the number of participants, we calculated the required sample size using GPower [39] for windows for one sample t-test with the smallest effect size (d) of 0.5, an alpha risk of 0.05, and a power of 0.80. A priori power analysis with one tail t-test suggested 28 as the minimum number of samples required, 30 participants were recruited for the experiments through an online advertisement on the university web portal and the participant was provided with remuneration of approximately $15 per hour for their time. To avoid the impact of abnormal emotion elicitation, the participants were inquired about their medical condition to ensure the absence of mental illness in the recent past. Other criteria included the absence of any signs and symptoms of common illness such as a minor headache or common cold. The Beck Depression Inventory (BDI) [40] test was conducted during the recruitment process to confirm that the participants did not have any psychological abnormalities. Due to the pandemic situation, further special precautions were taken such as screening for fever or headache before the study and wearing face masks during the experiments.

C. EXPERIMENT SCENARIO

After signing the consent form, the participants sat in front of a 22-inch screen with a resolution of 1920 × 1080 pixels. Stereo speakers were switched used for audio output, and the volume level was adjusted based on the participant's preference before the experiments. The participants were asked to position themselves comfortably in front of the screen and watch the stimulus video as shown in Fig. 3. An experimenter assisted them to wear an Emotiv Epoc X [41] EEG headset and an Empatica E4 [42] wristband. The EEG headset includes movable electrodes with wet saline terminals. The experimenter applied saline water to the contact points of the headset and placed the electrodes on the scalp. Similarly, the wristband was positioned correctly and connected to the experimenter's smartphone using a Bluetooth connection for recording. The EEG signals were wirelessly recorded at 256Hz using a vendor-provided USB dongle attached to the experimenter's machine. The impedance of the electrodes was maintained by saline-based felts attached to the electrodes, and the signal quality was confirmed on the recording software. EEG quality was verified using vendor-provided recording software which indicated contact quality and EEG signal quality. The participants were asked to avoid any movements to maintain the EEG quality. The experimenter commenced the experiment using a web-based annotation tool that we developed for the experiments. Each participant contributed to about 44 minutes of recording divided into three sessions of about 15 minutes each with at

TABLE 2. Stimulus video clips with corresponding emotions labeled by evaluators.

Expected Emotion	Video ID	Clip Title (Movie)	Duration (sec)	Agreement Percentage	Arousal	Valence
Happy	VID02	Maria (200 Pounds Beauty)	121	78.6%	5.79±2.11	6.96±1.48
	VID10	Little happiness (The Attorney)	121	71.4%	2.86±1.92	6.75±1.24
	VID12	School days (Sunny)	151	42.9%	3.57±1.82	6.71±1.60
	VID14	Cheating on you (Tazza-The Hidden Card)	121	57.1%	4.82±1.85	6.36±1.32
	VID22*	Introduction (Money)	121	53.6%(neutral)	3.64±1.93	6.36±1.04
Sad	VID05	I miss my dad (Miracle in Cell No.7)	181	85.7%	6.39±1.70	2.79±1.54
	VID13	Heartbreaking wounds (200 Pounds Beauty)	122	82.1%	5.00±1.73	3.82±1.23
	VID17	Lonely death (A Taxi Driver)	120	82.1%	6.29±1.36	2.36±1.20
	VID24	Because of my wife (Confidential Assignment)	121	82.1%	5.57±1.55	3.43±1.35
Angry	VID06 *	Hard day (A Taxi Driver)	120	50%(Sad)	5.43±1.78	3.39±1.63
	VID15	Forced labor (The Battleship Island)	121	71.4%	6.61±1.68	2.18±1.28
	VID21	Murderer (Dark Figure of Crime)	121	32.1%	5.86±1.53	3.61±1.35
	VID23	Words that only hurt (200 Pounds Beauty)	121	42.9%	4.57±1.52	3.43±1.32
Surprise	VID08	Furious fight (The Outlaws)	122	50%	7.57±1.15	3.43±1.88
	VID18	Dangerous guest (Alive)	63	53.6%	7.43±1.40	2.82±1.39
	VID25	Fight in space (Space Sweepers)	121	64.3%	6.61±1.57	6.25±1.43
Fear	VID03	Scary principal (Silenced)	121	42.9%	7.75±1.18	1.60±1.11
	VID11	Handle it quietly (A dirty Carnival)	121	50%	6.93±1.85	2.71±1.81
	VID16	Home without parents (Horror Stories)	120	78.6%	8.14±0.95	2.32±1.51
Neutral	VID04	Police academy (Midnight Runners)	63	78.6%	3.18±1.54	5.5±1.12
	VID09	A suspicious plan (Gangnam Blues)	121	78.6%	4.25±2.01	4.21±1.21
	VID19	Stranger things (Money)	121	67.9%	4.79±1.86	5.00±0.71
Disgust	VID01	Protein block (Snowpiercer)	61	78.6%	5.82±1.63	3.11±1.47
	VID07	Terrible bug (Deranged)	121	78.6%	5.71±1.81	3.18±1.67
	VID20	Zombie (Alive)	65	67.9%	6.86±1.51	2.96±1.48

least 10-15 minutes of break to avoid emotional fatigue or possible discomfort due to the prolonged wearing of the EEG headset. Participants wore the EEG headset for a maximum of 15 minutes during each session. However, due to the time needed for remounting the EEG headset after each break, the total time contribution per participant was about 3 hours.

The stimulus videos were not fully randomized to avoid having the same basic emotions in subsequent videos. To determine the order, we shuffled the videos with a constraint that no two consecutive videos contain the same emotion. Two sets of predefined orders were used alternately for the subjects. After each video, a color bars video containing calm music was displayed for 1 minute to neutralize the emotion; a color bars video is considered to have a soothing effect [26], [37]. Three test videos were used to familiarize the participants with the experiment process and annotation tool. During the annotation of these test videos, the participants occasionally adjusted their posture and interacted with the experimenters to ask about the annotation tool. It was observed that the annotators were able to annotate the third test video without movement and without disturbing the EEG signals. Using the test videos before the annotation of the stimulus videos was found to be effective in reducing potential noise in the signals due to lack of familiarity with the annotation tool.

An annotation interface was shown on the video player at the end of each video to report the emotions that they felt while watching the video. The participants were asked to report their emotions immediately; however, no time limit was imposed for labeling the emotions. Self-assessment Manikin (SAM) [43] was used for annotation as shown in Fig. 4. SAM-based annotation interface was displayed only after the stimulus video and recording of physiological signals were stopped. Participants selected a discrete emotion from a list of 7 basic emotions displayed in the first row of the annotation tool. Similarly, for a dimensional view of the

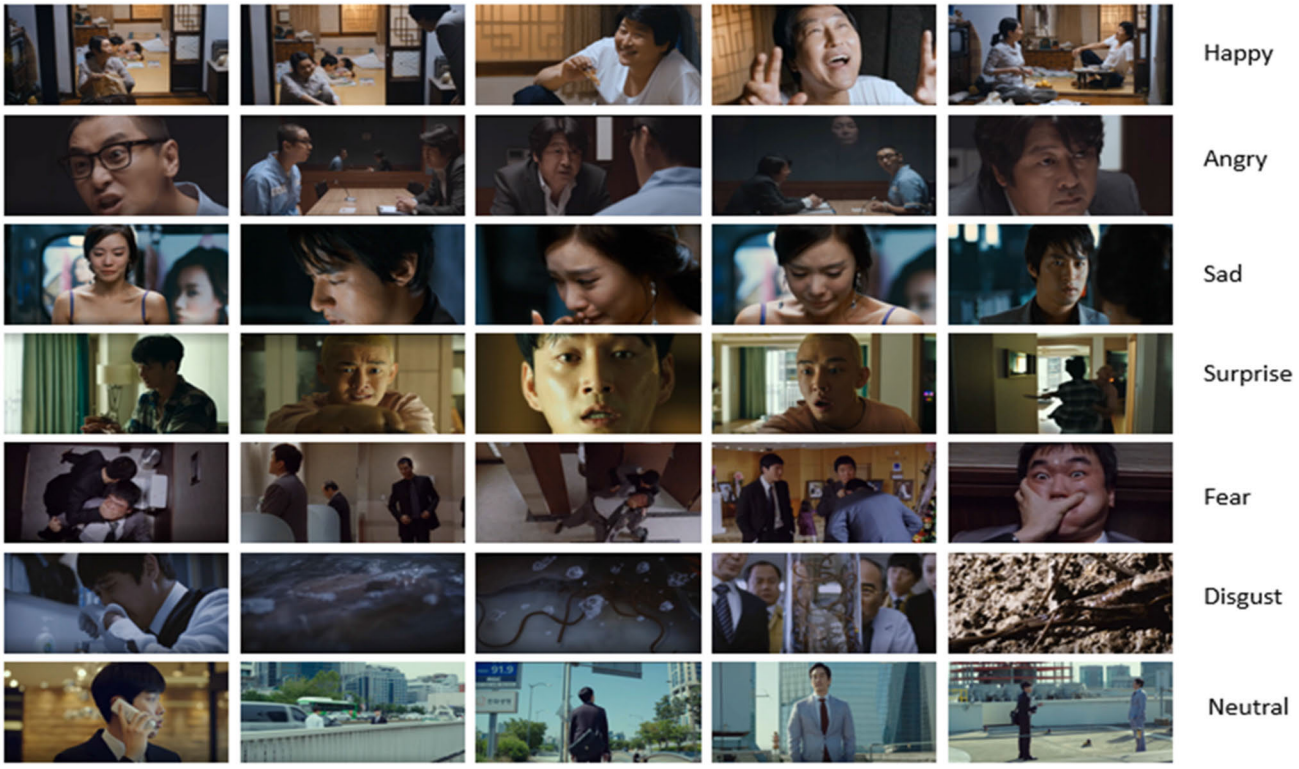

FIGURE 2. Example clips for seven basic emotions.

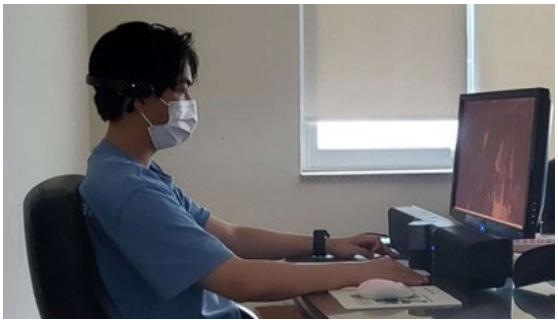

FIGURE 3. A participant watching a stimulus video during the experiment. The participant wore an EEG headset and a wristband for data recording.

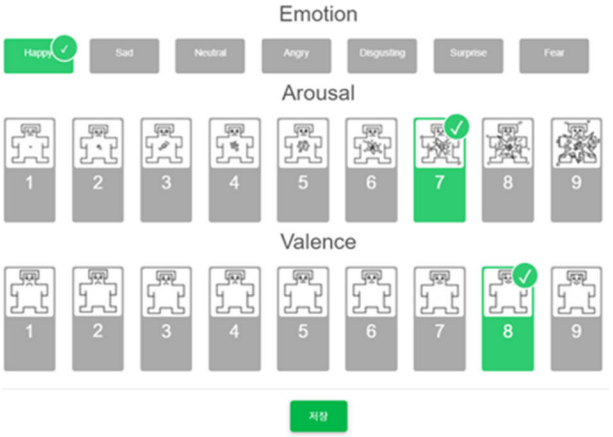

FIGURE 4. The Emotion annotation interface based on Self-Assessment Manikin (SAM) for labeling 7 basic emotions and two emotion dimensions on a 9-point scale.

emotions, the participants rated their felt emotions on a Likert scale of 1-9 represented by SAM icons on the annotation interface.

D. PHYSIOLOGICAL SIGNALS

We recorded 4 types of physiological signals: EEG, EDA, BVP, and TEMP. Electroencephalograms (EEG) capture electrical activity in the brain through electrodes placed on the scalp while EDA, BVP, HR, and TEMP were recorded using a wrist-worn sensor. Although various physiological can be used for emotion recognition, these modalities were selected based on their unobtrusiveness and ease of availability as consumer products. The analysis of emotions using consumer-grade devices leads to the implementation of emotion recognition wider application domain.

In EEG, the electrodes capture signals generated due to the movement of ions during the activation of neurons. Such activations are directly related to cognitive processes and various emotions [44]. EDA, also known as Galvanic Skin Response (GSR), refers to the change in electric potential in the skin in response to perspiration; it measures the effect of neural activities on the permeability of the sweat glands. The

TABLE 3. Wearable devices used in the experiments.

Devices	Collected Data	Sampling Rate
Expatica E4 Wristband	EDA	4
	BVP	64
	Temperature	4
Emotiv Epoc X	EEG	256

EDA signal represents the activity of the synthetic nerve on eccrine sweat glands [45]. It is a non-invasive measurement of skin conductance as it uses a constant supply of low voltage [46]. Due to this non-invasiveness, affordability, and convenient way of acquisition, EDA has been used widely for several applications in affective computing, including smart and intelligent wearable devices.

Blood Volume Pulse (BVP) recorded using photoplethysmography is useful in emotion recognition studies as the change in heart rate affects arousal and valence [47]; heart rate has been found to have a positive correlation with valence [48]. Several features, including heart rate variability (HRV), IBI, and spectral features of the BVP, can be used for emotion recognition.

Peripheral skin temperature fluctuates with the change in emotional states and is used for emotion recognition studies [21], [26]. Notably, the Skin temperature can be recorded continuously in a non-intrusive way using wrist-worn sensors.

E. DEVICES

To record the EEG data, we used Emotiv Epoc X, a wireless EEG headset with 14 electrodes (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) and two reference electrodes located at the left and right mastoid process. Emotiv X includes the saline-based wet electrodes with a 10/20 international system arrangement as shown in Fig. 5. Epoc X allows the recording of EEG signals with a frequency of 128Hz or 256 Hz.

To collect EDA, BVP, and temperature, we used the Empatica E4 wristband, a wrist-worn wearable device with multiple sensors. The E4 device captures various signals including EDA at 4Hz, BVP at 64Hz, and peripheral skin temperature at 4Hz as listed in Table 3. The device is also equipped with an accelerometer and gyroscope sensors. However, movement information is not included in the dataset because the movement of the body parts was prohibited during the experiment to avoid mechanical disturbances during EEG recording. Physiological signals from the E4 device were recorded using a mobile device (Galaxy Z Flip 5G, 256GB, 6.7 inches, 1080 × 2636) through a Bluetooth connection. Fig. 6 illustrates the devices used for data collection.

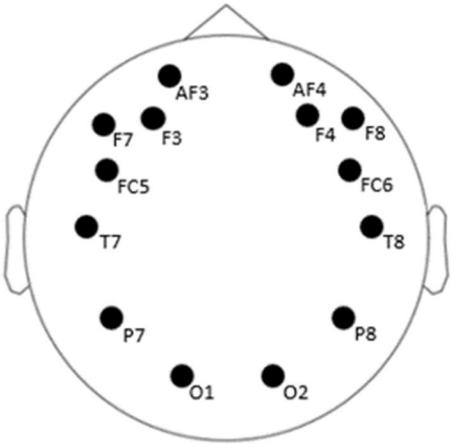

FIGURE 5. Position of EEG channels on the scalp.

FIGURE 6. Devices used for data collection.

IV. DATASET CHARACTERISTICS

The data gathered from multiple sensors in the experiment was associated with UTC timestamps, ensuring accurate time reference. To align the sensor signals with the actual duration of the experiment, the timestamps recorded using the annotation tool were used to clip the signals appropriately. The details of the collected dataset are presented in Table 4, providing an overview of the data acquired. In order to evaluate the stimulus data, a group of 28 individuals, referred to as evaluators, participated in the assessment process. These evaluators were responsible for reviewing and analyzing the stimulus videos.

Additionally, the physiological data and corresponding emotion labels were collected from a separate group of 30 participants, referred to as annotators. These individuals were involved in providing annotations and labeling the emotions exhibited in the stimulus videos. By involving both evaluators and annotators, the study aimed to gather comprehensive insights into the stimulus data and its impact on physiological responses and emotional states. The utilization of multiple participants ensures a diverse range of perspectives and enhances the reliability and validity of the collected data.

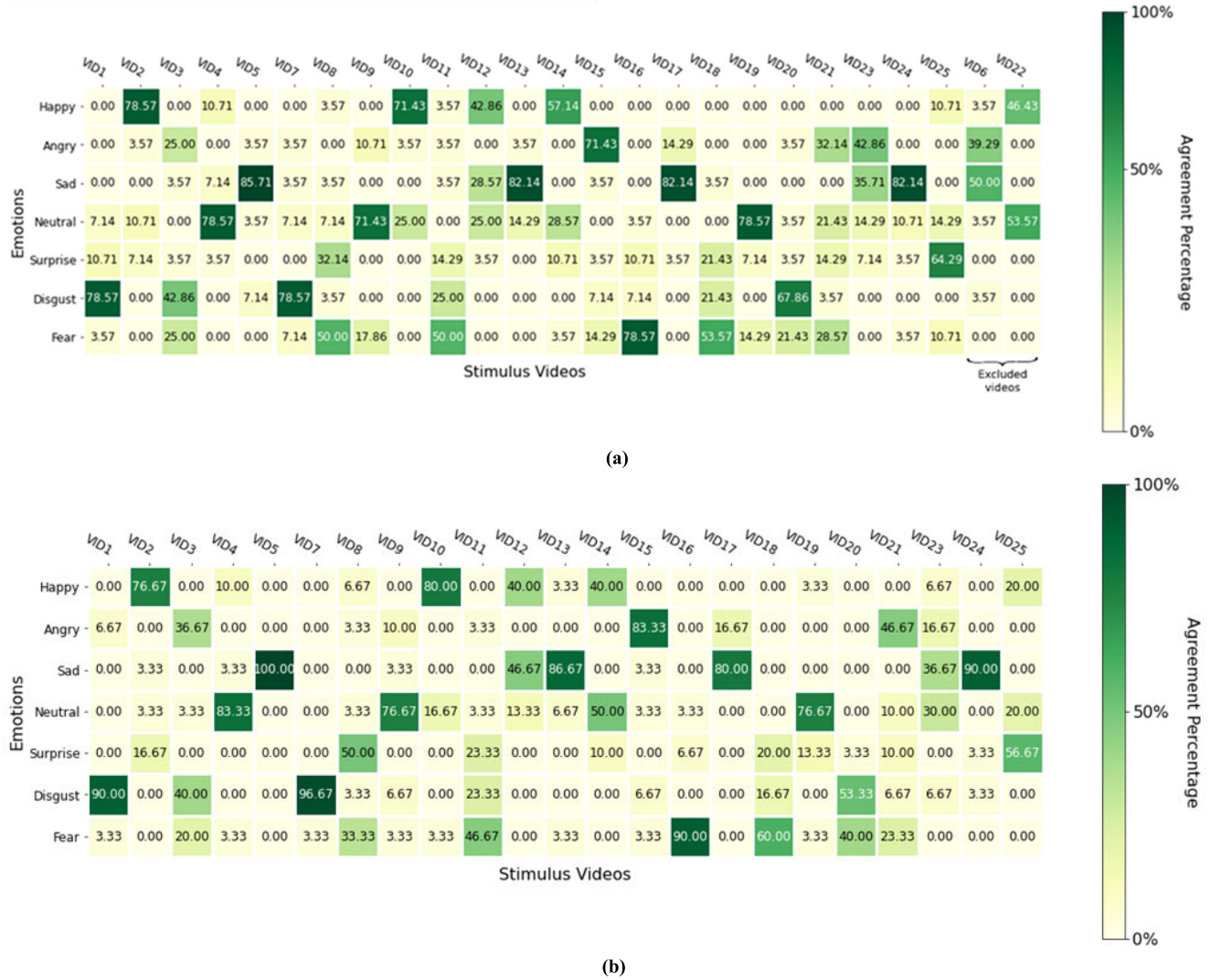

FIGURE 7. Heatmap showing the agreement percentage during (a) stimulus evaluation and (b) annotation during physiological data collection.

TABLE 4. Dataset characteristics.

Attribute	Description
Number of stimulus videos	23
Duration	61s - 181s
Number of the participants	30 (15 males, 15 females)
Age of the participants	20-30 (mean: 23.56, std:2.77)
Types of emotions	Happy, Sad, Angry, Surprise, Fear, Neutral, Disgust; Emotion Dimensions (Arousal and Valence)

A. STIMULUS DATASET

We constructed a stimulus dataset using video clips from various sources and further evaluated it through multiple annotations. A group of 28 evaluators labeled the videos with 7 basic emotions and dimensional emotions. To assess the relative accuracy of the annotations, we calculated the agreement among the evaluators. The selection of stimulus videos was based on the agreement among the evaluators and two videos with low agreement were removed from the stimulus set. We adopted Cronbach's Alpha [49] statistic as a measure of agreement among the evaluators. Cronbach's alpha is a commonly used interobserver agreement measure for continuous labels. For stimulus labeling the Cronbach's alpha of 0.97 and 0.96 were observed for valence and arousal respectively. Similarly, for the categorical annotations we used Fleiss Kapa [50] as a metric for inter-rater agreement and a moderate inter-rater agreement of 0.40 was observed.

B. INTER-OBSERVER AGREEMENT IN THE DATASET

As the same stimulus videos were shown to all the participants, the annotations were validated through inter-annotator agreement analysis. Cronbach's alpha of 0.84 for

arousal and 0.89 for valence was observed, indicating strong inter-annotator reliability among 30 participants for arousal and valence.

For categorical annotations of seven emotions, a moderate inter-rater agreement was observed with a Fleiss Kappa value of 0.50. We also analyzed the agreement of the categorical annotations based on the number of annotators agreeing on one of the seven basic emotions. As shown in Fig. 7, we computed percentage scores for each stimulus video. The results showed a high agreement among the annotators for most of the videos; 19 videos had an agreement of over 50%. The VID05 video had the highest agreement; a 100% agreement among the annotators. Videos VID23 and VID03 had the lowest agreement, with an agreement of 36.67 and 40 percent, respectively. Such low agreement might be due to conflicting emotions resulting from participants' personality differences; the same stimulus video may impart slightly different emotions to different individuals. For example, in the case of the stimulus VID23, the video contains a conversation between two characters which caused 36.67% of the annotators to feel sad which was the expected emotion, while 30 percent felt neutral and 16.67% felt angry as seen in Fig. 7 (b). Similarly, VID03(scary principal), where a school principal shows abusive behavior towards a female student, may induce fear if the subjects perceive the video from the student's perspective, while based on the general observation of his abusing behavior towards a female student, the subject may get angry. These differences in viewers' perspectives and the manner in which video content is consumed by the individuals suggest that individual differences among the participants may have influenced their focus on different characters in the scene, indicating the need for personality-based profiling in video-based stimuli for eliciting emotions.

C. DISTRIBUTION OF EMOTIONS IN AROUSAL-VALENCE SPACE

Stimulus videos were evaluated by 28 evaluators. Both the evaluators and the participants of the data collection experiments labeled the videos in both categorical and dimensional views. Based on the inter-rater agreement of annotations by stimulus evaluators, we selected 23 video clips for the experiments for data collection experiment.

As shown in Fig. 8 (a) and (b), both stimulus video evaluators and physiological data annotators were labeled similarly for arousal and valence. Fig. 8 (a) shows the distribution of the average values of annotations in arousal and valence space as annotated by 28 evaluators for the stimulus videos. The majority of the videos were labeled as high arousal and low valence. This is due to the selection of stimuli based on the categorical labels. Fig. 8(b) shows the average values of the annotations during the physiological data acquisition experiment while watching the stimulus videos. The videos for happy and surprise can be seen on the positive valence quadrants while videos with surprise were found to be labeled both positive and slightly negative. It can

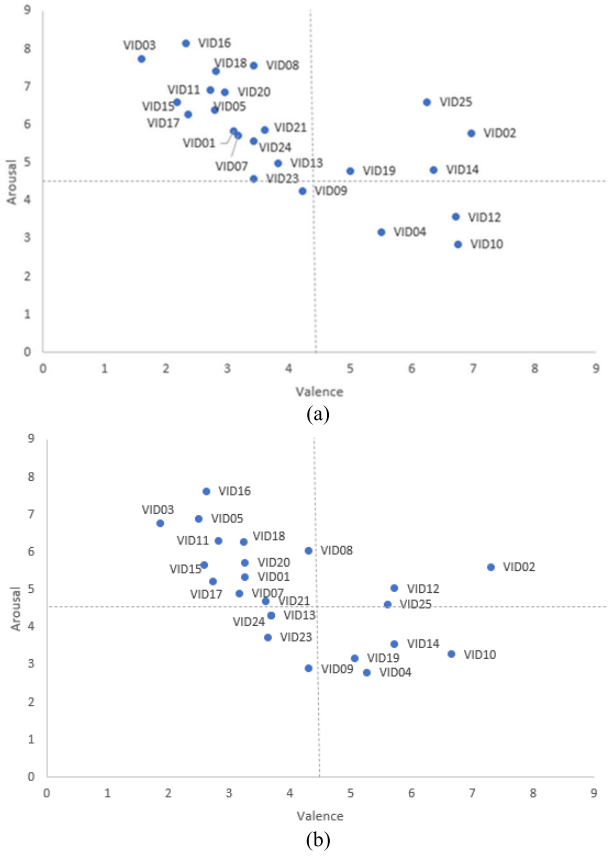

FIGURE 8. Distribution of the mean arousal and valence in stimulus dataset (a) and those labeled by annotators during the physiological data collection experiment (b).

be concluded that the videos were labeled following a similar trend and the participants of the data acquisition experiment expectedly annotated the felt emotions.

D. COMPARISON OF LABELS WITH THE STIMULUS VIDEO ANNOTATIONS

To evaluate the annotations, we calculated Spearman's correlation coefficient between the mean of the participants' annotations from the physiological data collection experiment and the annotations made by the evaluators for each video. As the videos were labeled by a different set of annotators with different numbers, we calculated the arousal and valence for each video. We observed a high correlation of 0.9708 for valence and 0.8702 for arousal. A strong linear correlation was observed between the annotations and the stimulus labels. This shows that the participants annotated the videos in the same way as the stimulus video dataset. The strong linear correlation between the 9-point labels of arousal and valence indicates that the participants in the data acquisition experiment voted in the same way as the participants of the stimulus labeling experiment. The participants were provided with proper information on the annotation process and rating scales, and the selected stimulus video set is appropriate for emotion stimulation.

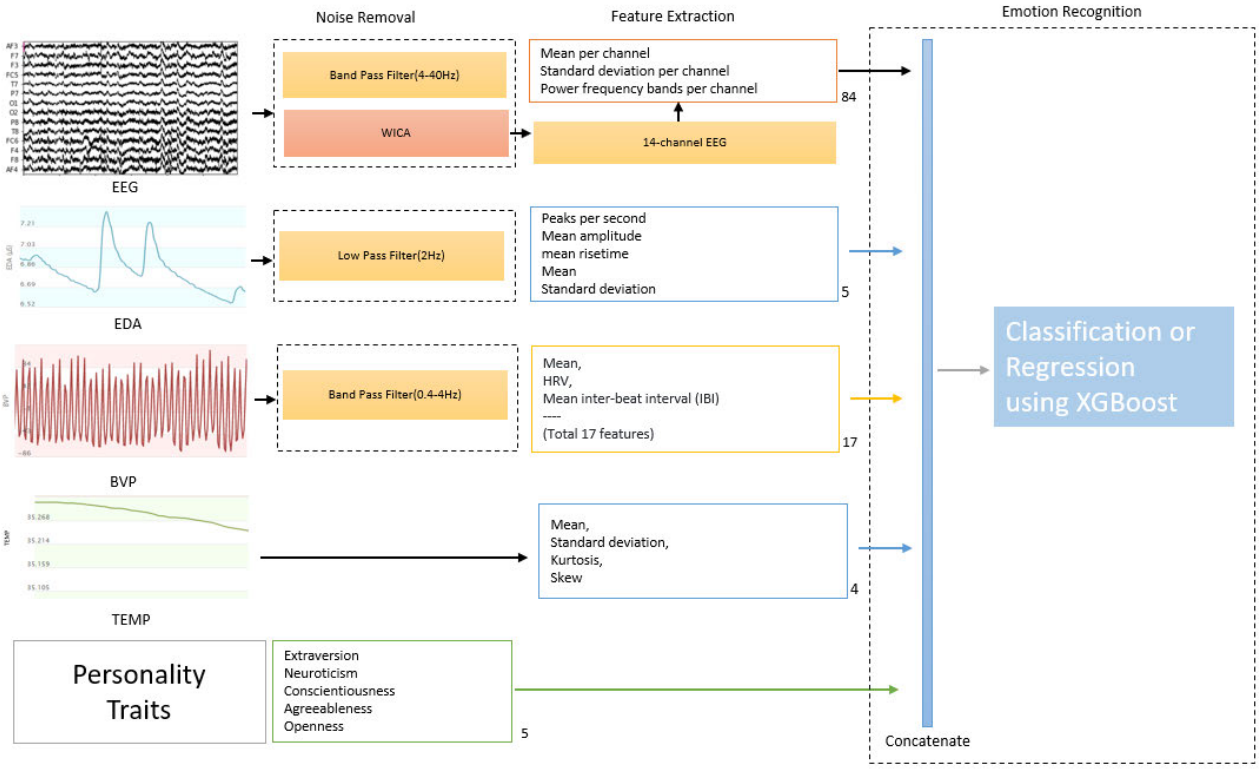

FIGURE 9. The overall framework for emotion recognition using physiological signal. The feature extraction part includes extraction of the features from each channel of EEG and other modalities. The emotion recognition module includes classification of seven basic emotions and prediction of 9-point labels for arousal and valence.

The results show that valence is more consistent across participants than arousal.

E. PERSONALITY CORRELATION

For personality information, we use Big5 personality traits, namely extraversion, neuroticism, conscientiousness, agreeableness, and openness using the Newcastle Personality Assessor (NPA) questionnaire [34]. NPA is a short questionnaire with 12 questions to be answered on a Likert scale of 1-5, representing very uncharacteristic to very characteristic. We obtained the personality scores on a 4-point scale representing low, low-medium, medium-high, and high based on the NPA questionnaire interpretation [34]. We calculated the average Spearman's correlation coefficient of the personality scores and the annotated values in valence and Arousal annotated for emotions felt while watching 23 videos. Table 5 shows the correlation coefficients between the personality traits and levels of emotional dimensions. Valence was found to have weak negative significant ($p < 0.05$) correlation with openness while the correlation of valence with other personality traits was not significant. This implies that the personality traits do not necessarily affect the level of valence. However, arousal had a significant negative correlation with neuroticism and agreeableness, while Conscientiousness had a positive correlation with arousal. Among the personality traits, a negative significant correlation was found between neuroticism and conscientiousness indicating and inverse relationship between the two personalities. The correlations between the personality traits and emotions in arousal and valence were not significant for all personality traits, suggesting that personality does not necessarily play a significant role in the intensity of felt emotions.

TABLE 5. Correlation among personality traits and emotion dimensions annotated for 23 videos.

Personality	Valence	Arousal	Ex	Ne	Co	Ag
Ex	0.016	-0.001				
Ne	0.020	**-0.176***	**0.103***			
Co	-0.040	**0.108***	**-0.111***	**-0.385***		
Ag	0.065	**-0.098***	0.036	**0.167***	**0.226***	
Op	**-0.115***	-0.004	**-0.231***	**0.272***	0.010	**-0.317***

Ex, Ne, Co. Ag, and Op stand for extraversion, neuroticism, conscientiousness, agreeableness, openness to experience, significant correlations ($p<0.05$) are in bold.

In general, the analysis of all stimulus videos did not show a significant correlation, however, there was an anomaly

in the annotation as seen in Fig. 7. in the interrelated emotions such as sadness and anger in a conversational context. As seen in VID23, the conversation between two characters may have two different aspects imparting distinct emotions to the viewer. To investigate the role of personality in such a scenario we calculated Spearman's correlation of the personality traits and the annotations on VID23 and VID03 by all the participants. For VID23 and VID03, which potentially could induce low valence, a negative significant correlation of -0.447 between valence and openness was found. Such an inverse relationship between openness and valence suggests that the participants with low openness experienced higher valence. Moreover, in case of VID23 and VID03, which were labeled as anger and fear by most participants, a negative correlation was found between arousal and personality traits, was seen while for the overall dataset, neuroticism showed a negative correlation with arousal. In addition, both openness showed a negative correlation with both arousal and valence, suggesting a lower level of emotion elicitation for the subjects having high scores in openness. These observations were found to agree with the existing studies [12].

F. DATA AVAILABILITY

The dataset is publicly available for academic research at https://sites.google.com/view/phymer-dataset. The data presented in this study were also available as a part of the Third Korean Emotion Recognition Challenge (KERC) 2021. The preprocessed data, divided into training, validation, and the test set was open to the participants of the competition on Kaggle during the com-petition duration (from Aug 30 to Oct 31, 2021). (https://www.kaggle.com/c/kerc2021). The baseline model for the competition was an LSTM-based classification model, achieving an f1-score of 0.55.

V. BASELINE EXPERIMENTS

In this section, we present the implementation of the baseline method for multimodal emotion recognition. We also evaluate the performance of the proposed baseline model in a subject-independent and subject-dependent manner. As the range of physiological data such as heart rate, blood pulse volume, or electrodermal activity may differ among individuals, within-subject normalization was performed for subject-independent emotion analysis. We normalized the extracted features using Robust Scalar [59] which scales the data based on quantile range and is suitable for small data sizes.

A. PREPROCESSING

EEG data is highly susceptible to noise due to its sensitivity towards minor physiological and physical activities such as blinking of eyes, heartbeats, and muscular movements [53]. We applied a band-pass filter of 4Hz-40Hz to include only the frequencies in the range of 4 bands (Theta, Alpha, Beta, and Gamma) of EEG signals as the Delta band is not relevant to this study as it is observed only during sleep. Although the band-pass filter of this range eliminates certain noise components, it is impossible to get rid of the noise using frequency alone as some artifacts may lie within the same frequency range as the EEG signals. We applied AWICA [54] to eliminate such noise.

AWICA is a threshold-based automatic artifact removal technique involving wavelet analysis and Independent Component Analysis (ICA). In this method, each channel of the EEG signal is partitioned into 4 bands of EEG through Discrete Wavelet Transform (DWT). The artifactual components in the WT components are selected quantitatively based on thresholds of Kurtosis and Reny's Entropy [55]. Then ICA is performed on the wavelet components for automatic rejection of the artefactual components. ICA is a commonly used blind source separation technique used for isolating the source signals from the recorded signals. Finally, artifact-free EEG channels are reconstructed through inverse ICA followed and subsequent inverse DWT operation. Noise removal in EDA signals is done using a low pass filter with a 2Hz cutoff frequency [56].

Blood Pulse Volume (BVP) signals recorded using photoplethysmography are susceptible to motion artifacts (MA) [57]. However, in our experiments, we minimized the wrist motion with the sensor to avoid motion artifacts. As the dataset did not include motion information, the preprocessing of BVP signals mainly involved removing out-of-band noise through the fourth-order Butterworth band-pass filter (BPF) between 0.4Hz and 4.0 Hz. For temperature (TEMP), only normalization was performed as the experiment process did not involve activities that could produce significant

TABLE 6. Features extracted from physiological signals.

Modalities	Features
EEG (84 features)	mean per channel (14 features), standard deviation per channel (14 features), band power spectrum (56 features)
EDA (5 features)	peaks per second, mean amplitude, mean rise time, mean, standard deviation
BVP (17 features)	mean, heart rate variability, mean inter-beat interval, multi-scale entropy at 5 levels (5 features), spectral power for 4 bands (4 features), spectral power ratio, tachogram low-frequency spectral content, tachogram medium frequency spectral content, tachogram high-frequency spectral content, tachogram energy ratio
TEMP (4)	mean, standard deviation, kurtosis, skew

TABLE 7. Classification results of seven emotions.

Modalities	Subject Dependent		Subject Independent	
	F1-score	MCC	F1-Score	MCC
EEG	0.7727/ 0.7755	0.7329/0.7359	0.1854/0.1827	0.0432/0.0387
EEG, EDA	0.7778/ 0.7909	0.7387/0.7386	0.1828/0.1901	0.0359/0.0445
EEG, EDA, BVP	0.7641/ 0.7759	0.7225/0.7358	0.1812/0.1872	0.0399/0.0451
EEG, EDA, BVP, TEMP	0.7659/0.7673	0.7251/0.7255	0.1877/0.1809	0.0454/0.0416

The '/' separates the result for modalities without personality and when combined with personality features. The f1-score for the random guess for 7-class classification for subject-dependent and subject-independent was 0.1407 and 0.1474 respectively. MCC for subject-dependent and subject-independent were -0.0021 and -0.0008 respectively

noise in temperature data. Eight samples for Subject SUB10 (SUB10VID09-SUB10VID16) were excluded from the experiments due to the device malfunctioning during the data collection experiment.

B. EMOTION RECOGNITION

To evaluate the dataset, we conducted subject-dependent and subject-independent experiments and compared the performance using different combinations of the features. For subject dependent method we performed a 5-fold cross-validation on the whole dataset split into training and validation sets at the ratio of 80 and 20 percent respectively and compared it with the subject-independent method where the leave-one-subject-out cross-validation, where one subject's data was used for validation.

For the baseline experiments, we extracted various handcrafted features for different modalities. Inspired by the high performance of features based on Toolbox for Emotional feature extraction from Physiological signals (TEAP) [58] features, we followed [26] for feature extraction. For each of the 14 channels of EEG samples, we calculated the power spectrum for theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz), and gamma (30-40Hz) bands, and obtained 56 features (14 channels $\times$ 4 bands). In addition to the band power, we included two statistical features (mean and standard deviation) for each channel. Similarly, for EDA and BVP, and skin temperature (TEMP) we extracted various features as shown in Table 6.

We performed experiments involving the classification of the seven basic emotions and the prediction of arousal and valence values labeled by the participants. Following feature extraction, we employed the Extreme Gradient Boosting (XGBoost) [56] based classifier and regressor models, chosen for their demonstrated effectiveness in various classification and regression tasks [26]. Different combinations of EDA and BVP features (Table 6) extracted in the preprocessing step were concatenated with the personality feature to perform the classification and prediction in the presence and the absence of personality features. The proposed emotion recognition framework is illustrated in Fig. 9. The baseline system consists of handcrafted feature extraction and emotion recognition modules. The feature extraction module involves the preprocessing of different physiological signals where statistical, and signal features are extracted. In the classification module, the extracted features are fused and used for the recognition of both categorical and dimensional emotion labels.

We evaluated the classification results using F1-score and Mathews Correlation Coefficient (MCC) to evaluate the classification task considering the imbalance in the dataset. F1-score is a commonly used evaluation metric for classification tasks representing a harmonic mean of precision and recall. As F1-score does not consider the true negatives, MCC can better represent the classification accuracy [57]. The best MCC score was obtained with the combination of EEG and EDA, where there was no significant improvement while using the personality.

To assess the influence of multimodality on emotion recognition, we conducted a comparison of classifications using different modalities. The results, presented in Table 7, indicate that the incorporation of BVP and TEMP does not necessarily lead to performance improvement. Furthermore, when combining personality features with physiological signals, we observed only a slight increase in performance, suggesting a weak or low impact of personality on emotion labeling. As the use of multiple modalities resulted in an increase in performance, it appears that further research on the effective fusion of the modalities needs to be performed. As the EEG signals are highly subjective there was a huge difference in subject-dependent and subject-independent classification results. Arousal and Valence prediction was evaluated using mean absolute error (MAE) and concordance correlation coefficient (CCC) as shown in Table 8, which shows the inclusion of personality features improves the prediction of both arousal and valence.

Table 9 presents the F1-scores for each categorical emotion comparing the performance when a different combination of modalities is used in the presence and absence of personality features. We can see from Table 9 and Table 10 that the use of all modalities improved the performance for some emotions in some cases. Disgust, happiness, and anger showed higher performance when personality information was used, while

TABLE 8. Subject-dependent and subject-independent prediction of arousal and valence.

		Subject-dependent		Subject-independent	
Modalities	Metric	Arousal	Valence	Arousal	Valence
EEG	MAE	1.7898/1.7555	1.5261/1.5219	1.9348/1.9250	1.5908/1.5893
	CCC	0.1737/0.1775	0.0972/0.1082	0.0233/0.0163	0.0183/0.0200
EEG, EDA	MAE	1.7854/1.7496	1.5155/1.5132	1.9436/1.9175	1.5915/1.5880
	CCC	0.1846/0.1780	0.1105/0.1171	0.0200/0.0265	0.0167/0.1938
EEG, EDA, BVP	MAE	1.7723/1.7534	1.5137/1.5125	1.9413/1.9148	1.5877/1.5951
	CCC	0.2040/0.2097	0.1006/0.1130	0.0217/0.0247	0.0266/0.1598
EEG, EDA, BVP, TEMP	MAE	1.7786/1.7561	1.5154/1.5134	1.9397/1.9136	1.5960/1.6005
	CCC	0.2035/0.1983	0.1107/0.1175	0.0219/0.0253	0.0160/0.0127

The '/' separates the results for modalities without personality and when combined with personality features. MAE: mean absolute error, CCC: concordance correlation coefficient.

TABLE 9. Subject-dependent classification performance for each class in the presence of personality features.

		disgust	happy	angry	neutral	sad	surprise	fear
EEG	F1-score	0.7307/**0.7465**	**0.7785**/0.7752	0.7718/**0.7765**	**0.7861**/0.7836	0.8187/**0.8206**	0.7115/**0.7208**	**0.7499**/0.7488
	MCC	0.6959/**0.7155**	**0.7500**/0.7473	0.7509/**0.7563**	**0.7412**/0.7375	0.7623/**0.7648**	0.6945/**0.7029**	**0.7123**/0.7112
EEG, EDA	F1-score	**0.7430**/0.7413	**0.7896**/0.7812	0.7464/**0.7784**	0.7957/**0.8040**	**0.8193**/0.8093	**0.7322**/0.7227	**0.7577**/0.7568
	MCC	**0.7123**/0.7110	**0.7606**/0.7522	0.7215/**0.7555**	0.7527/**0.7639**	**0.7623**/0.7497	**0.7161**/0.7061	**0.7214**/0.7205
EEG, EDA, BVP	F1-score	0.7222/**0.7288**	0.7593/**0.7685**	0.7421/**0.7824**	0.7710/**0.7964**	**0.8125**/0.8119	0.7282/**0.7300**	0.7542/**0.7625**
	MCC	0.6873/**0.6938**	0.7286/**0.7390**	0.7170/**0.7624**	0.7236/**0.7534**	**0.7531**/0.7525	0.7107/**0.7115**	0.7163/**0.7268**
EEG, EDA, BVP, TEMP	F1-score	0.7225/**0.7345**	0.7748/**0.7824**	0.7420/**0.7731**	**0.7847**/0.7766	**0.8110**/0.7998	0.7058/**0.7286**	**0.7530**/0.7353
	MCC	0.6875/**0.7027**	0.7492/**0.7535**	0.7156/**0.7516**	**0.7395**/0.7295	**0.7518**/0.7359	0.6873/**0.7113**	**0.7156**/0.6956

The '/' separates the f1-scores for modalities without personality and when combined with personality features.

the performance was not improved in the case of neutral, sad, fear, and surprise. All the experiments were on Nvidia GeForce RTX 3080Ti GPU with 12 GB of memory. The experiment codes were implemented in Python, and to ensure the reproducibility of the experiment results, a fixed random seed of 42 was set for all the experiments.

We believe that performance improvement can be achieved through different feature engineering and multimodal feature fusion techniques. Moreover, this dataset can provide value to multimodal emotion recognition research through further analysis of personality as a context. The dataset contains annotations for both categorical and continuous affects, enabling its utilization in the development of multitask models. The diverse range of annotations within the dataset provides an opportunity to enhance the complexity of models and further improve the performance of multitask models in future studies.

VI. DISCUSSION

This paper primarily focuses on dataset acquisition and annotation protocol. We generated a stimulus dataset, designed a data collection experiment, and developed a web-based data annotation tool for physiological data generation. Moreover, a baseline method for emotion classification was developed to evaluate the dataset.

The experiment protocol was meticulously crafted to simulate the real-world scenario to the greatest extent possible. Although the ideal data acquisition procedure required signal recording while the participants were consciously unaware of the collection process, it is practically impossible for EEG-based datasets as certain restrictions are required to avoid signal noise. Despite having been conducted in laboratory settings, comfortable posture for watching stimulus videos, use of consumer-grade commercial equipment, and personalized audio levels make the data collection natural and close

TABLE 10. Subject-independent classification performance for each class in presence of personality features.

		disgust	happy	angry	neutral	sad	surprise	fear
EEG	F1-score	0.1572/**0.1796**	0.1577/**0.1764**	0.0808/**0.1461**	0.1533/**0.1772**	**0.3231**/0.2392	0.0302/**0.0304**	**0.1757**/0.1504
	MCC	0.0674/**0.0930**	0.0718/**0.0918**	0.0192/**0.0477**	**-0.0033**/-0.0131	**0.0552**/0.0162	**-0.0146**/-0.0170	**0.0814**/0.0631
EEG, EDA	F1-score	0.1653/**0.1698**	0.1506/**0.1816**	0.0897/**0.1445**	0.1644/**0.1914**	**0.3046**/0.2505	0.0253/**0.0305**	**0.1710**/0.1689
	MCC	**0.0780**/0.0766	0.0658/**0.0977**	0.0248/**0.0483**	**0.0111**/0.0064	**0.0241**/0.0195	-0.0208/**-0.0164**	0.0718/**0.0882**
EEG, EDA, BVP	F1-score	0.1718/**0.1827**	0.1294/**0.1635**	0.0611/**0.1246**	0.1621/**0.1792**	**0.3093**/0.2802	**0.0498**/0.0381	**0.1819**/0.1572
	MCC	0.0854/**0.0928**	0.0436/**0.0812**	-0.0031/**0.0228**	**0.0130**/-0.0029	0.0327/**0.0618**	**0.0135**/-0.0007	**0.0848**/0.0755
EEG, EDA, BVP, TEMP	F1-score	0.1634/**0.1754**	0.1403/**0.1432**	0.0776/**0.1319**	0.1549/**0.1685**	**0.3155**/0.2634	**0.0621**/0.0518	**0.1778**/0.1668
	MCC	0.0779/**0.0871**	0.0600/**0.0663**	0.0133/**0.0332**	**0.0107**/-0.0069	**0.0419**/0.0320	**0.0185**/0.0074	0.0825/**0.0865**

The '/' separates the f1-scores for modalities without personality and when combined with personality features.

to a real-world situation. A key feature that sets our dataset apart from existing ones is the inclusion of fine-grained emotion annotation, which utilizes a 9-point scale to provide a dimensional perspective of emotions. This enhances the richness and granularity of emotional information captured in the dataset, offering a more comprehensive view of emotional states compared to other available datasets.

Context plays an important role in emotional representation. The context in emotion recognition can be represented by environmental and socio-cultural factors. As this dataset focuses on the bio-signal data collected in a uniform environment, only the personality traits have been included in the dataset to provide the individual nature of the participants as a context for emotion recognition. The individual differences and their impact on emotion recognition can be studied in the context of various other factors such as age, gender, and ethnicity. However, this dataset focuses on adults from homogenous populations.

With the increasing use of consumer-grade wearable sensors, the collection and analysis of physiological data has become more convenient and non-obtrusive. Several accessories such as wristbands and headsets have been widely accepted as they can provide personalized information. However, behavioral differences among individuals also need to be considered during the analysis of different types of emotions. Individuals with reserved or introverted personalities may not respond to the same stimuli as actively as extroverts. In other words, the degree of emotional expression may not be the same for individuals with different personalities, therefore analysis of fine-grained continuous values in different dimensions is essential.

Therefore, the same stimulus videos often induce different emotions in different individuals. Such differences may arise when the stimulus contains interrelated emotions and participants watch the videos with a focus on a different character. For example, a video with a character showing anger at another character may induce either anger or sadness based on which character the participant focuses on. Therefore, the role of personality in distinguishing closely related emotions requires further exploration, although there is no significant correlation between personality traits and felt emotions in general.

While the dataset's primary intended use is for studying emotions during video content consumption, it holds potential for broader applications concerning the analysis of physiological signal variations among individuals in different emotional scenarios. The dataset, collected using consumer-grade portable headsets in response to video-based stimuli, offers multichannel EEG data that could contribute to advancements in brain-computer interfaces (BCI) by exploring temporal and spatial information. This physiological signal-based emotion recognition method also has potential for application in the analysis of emotions in immersive environments such as virtual reality (VR) based games. We also conducted experiments with VR devices with 360-degree videos as emotion stimuli to study the application of physiological signals in an immersive environment. However, our preliminary experiments showed that both mechanical and electronic noise affect the data collection process due to the simultaneous use of VR and EEG devices. Thus, the application domain of physiological datasets is limited to situations with minimal movement.

The dataset compilation has certain limitations, primarily related to noise interference in EEG sensors. Unavoidable sources of noise, such as blinking of eyes and subtle muscular movements during data collection, may have affected the dataset. To ensure effective utilization of the datasets, appropriate noise removal techniques are crucial. Another limitation is the potential bias arising from the possible elicitation of multiple emotions. Efforts were made to mitigate this bias through careful stimulus video selection and labeling before the experiments. The chosen stimulus video clips were intended to depict a single emotion, and 28 evaluators of the same age group verified this.

However, the personalities of these evaluators might differ from the 30 annotators participating in the physiological data collection experiments. Despite the thorough selection process, one video (VID21) was found to induce both anger and fear emotions, both being negative valence emotions. Considering individual differences, this video was included in the dataset. Consequently, the inclusion of samples recorded through stimulation using such videos is a limitation that may hinder the distinction between emotions like anger and fear. Additionally, the stimulus data collection process involved partial use of an existing emotion recognition dataset. Although the reconstructed stimulus dataset was evaluated by 28 evaluators, the final selection was made by the experimenters based on evaluator agreement. Videos with different emotion labels from the original dataset were excluded, which might conflict with the purpose of the evaluation experiment. However, as the videos in the source dataset were labeled by multiple annotators, it is expected that any decision bias would be minimal.

Similarly, the familiarity with stimulus videos, human limitations in identifying own emotions, and homogenous demography are some of the limitations to be considered during the use of this dataset. The order of the stimulus videos was predefined by conditionally randomizing the order, where the order was changed if the consecutive videos had the same emotion labels. Although the order was made different for two consecutive subjects, the order was the same for alternate subjects. Such partial randomization may have induced order bias in the dataset. Another limitation of the present study is the use of the NPA questionnaire for personality data collection. A short questionnaire might have limited the reliability to some extent in the case of a small sample size. Although multiple modalities were considered in this study, we explicitly excluded the video modalities which are common in multimodal emotion recognition.

VII. CONCLUSION

In this paper, we presented a multimodal bio-signal dataset for emotion recognition with both categorical and dimensional perspectives and analyzed the importance of personality as a context. The datasets include various physiological signals obtained from 30 participants, where the emotions were induced through stimulus videos. We observed strong inter-observer agreement among the annotators. To evaluate the dataset quality, we performed baseline experiments using a multimodal classification model which achieved an F1-score of up to 0.73 with multiple physiological modalities.

We plan to continue our research work in multimodal emotion recognition using physiological signals. We will improve the baseline classification and prediction models through improved feature extraction using deep learning techniques. As a further study, the stimulus dataset will be investigated with multiple emotion labels.

The database is publicly available for academic research in the field of emotion recognition with the hope that this dataset would be beneficial in the development of new emotion recognition methods and algorithms.

REFERENCES

[1] P. Ekman, "Are there basic emotions?" *Psychol. Rev.*, vol. 99, no. 3, pp. 505–553, 1992.
[2] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer, M. Tomita, and A. Tzavaras, "Universals and cultural differences in the judgements of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.
[3] P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russel's mistaken critique," *Psychol. Bull.*, vol. 115, no. 2, pp. 268–287, 1994.
[4] W. G. Parrot, *Emotions in Social Psychology: Essential Readings*. Hove, U.K.: Psychology Press, 2001, pp. 30–40.
[5] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Emotion: Theory, Research, and Experience: Theories of Emotion*, vol. 1, R. Plutchik and H. Kellerman, Eds. New York, NY, USA: Academic, 1980, pp. 3–33.
[6] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.
[7] K. R. Scherer, "The component process model: Architecture for a comprehensive computational model of emergent emotion," in *Blueprint for Affective Computing: A Sourcebook*. New York, NY, USA: Oxford Univ. Press, 2010, pp. 47–70.
[8] B. Meuleman and D. Rudrauf, "Induction and profiling of strong multi-componential emotions in virtual reality," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 189–202, Jan. 2021.
[9] M. Q. Menétrey, G. Mohammadi, J. Leitão, and P. Vuilleumier, "Emotion recognition in a multi-componential framework: The role of physiology," *Frontiers Comput. Sci.*, vol. 4, Jan. 2022, Art. no. 773256, doi: 10.3389/fcomp.2022.773256.
[10] G. Mohammadi and P. Vuilleumier, "A multi-componential approach to emotion recognition and the effect of personality," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1127–1139, Jul. 2022, doi: 10.1109/TAFFC.2020.3028109.
[11] L. F. Barrett, "The theory of constructed emotion: An active inference account of interoception and categorization," *Social Cogn. Affect. Neurosci.*, vol. 12, no. 1, pp. 1–23, 2017.
[12] L. F. Barrett, *How Emotions Are Made: The Secret Life of the Brain*. New York, NY, USA: Houghton Mifflin Harcourt, 2017.
[13] J. J. Gross and O. P. John, "Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior," *J. Personality Social Psychol.*, vol. 72, no. 2, pp. 435–448, 1997.
[14] J. Berkhout and D. O. Walter, "Temporal stability and individual differences in the human EEG: An analysis of variance of spectral values," *IEEE Trans. Bio-Med. Eng.*, vol. BME-15, no. 3, pp. 165–168, Jul. 1968.
[15] A. Keil, M. Stolarova, S. Heim, T. Gruber, and M. M. Müller, "Temporal stability of high-frequency brain oscillations in the human EEG," *Brain Topogr.*, vol. 16, no. 2, pp. 101–110, 2003.
[16] R. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles, "The electroencephalogram as a biometric," in *Proc. Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2001, pp. 1363–1366, doi: 10.1109/CCECE.2001.933649.
[17] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using 2. Commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr. 2018.
[18] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr. 2021, doi: 10.1109/TAFFC.2018.2884461.
[19] G. Shen, X. Wang, X. Duan, H. Li, and W. Zhu, "MEmoR: A dataset for multimodal emotion reasoning in videos," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, Oct. 2020, pp. 493–502, doi: 10.1145/3394171.3413909.
[20] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015, doi: 10.1109/TAMD.2015.2431497.

[21] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012, doi: 10.1109/T-AFFC.2011.25.

[22] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis ;Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[23] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[24] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul. 2015, doi: 10.1109/TAFFC.2015.2392932.

[25] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3438–3446, doi: 10.1109/CVPR.2016.374.

[26] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Sci. Data*, vol. 7, p. 293, Sep. 2020. [Online]. Available: https://www.nature.com/articles/s41597-020-00630-y

[27] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-aware multimodal emotion recognition using Frege's principle," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14222–14231, doi: 10.1109/CVPR42600.2020.01424.

[28] P. Chevalier, J. Martin, B. Isableu, and A. Tapus, "Impact of personality on the recognition of emotion ex-pressed via human, virtual, and robotic embodiments," in *Proc. IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2015, pp. 229–234, doi: 10.1109/ROMAN.2015.7333686.

[29] G. Matthews, I. Deary, and M. Whiteman, *Personality Traits*. Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 392–429.

[30] H. Li, N. Pang, S. Guo, and H. Wang, "Research on textual emotion recognition incorporating personality factor," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2007, pp. 2222–2227.

[31] N. Omhenı, A. Kalboussı, O Mazhoud, and A. H Kacem, "Annotation-based learner's personalitymodeling in distance learning context," *Turkish Online J. Distance Educ.*, vol. 17, no. 4, pp. 46–62, 2016.

[32] P. T. Costa and R. R. McCreae, *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Odessa, FL, USA: Psychological Assessment Resources, 1992.

[33] L. R. Goldberg, "The development of markers for the big-five factor structure," *Psychol. Assessment*, vol. 4, no. 1, pp. 26–42, Mar. 1992.

[34] D. Nettle, *Personality: What Makes You the Way You Are*. New York, NY, USA: Oxford Univ. Press, 2007, pp. 249–253.

[35] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of consensus decision making meetings," *Lang. Resour. Eval.*, vol. 41, no. 3, pp. 409–429, Dec. 2007.

[36] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr., "A very brief measure of the Big-Five personality domains," *J. Res. Pers.*, vol. 37, no. 6, pp. 504–528, Dec. 2003.

[37] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cogn. Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[38] T. L. B. Khanh, S.-H. Kim, G. Lee, H.-J. Yang, and E.-T Baek, "Korean video dataset for emotion recognition in the wild," *Multimedia Tools Appl.*, vol. 80, no. 1, pp. 9479–9492, 2021.

[39] F. Faul, E. Erdfelder, A. G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav. Res. methods*, vol. 39, no. 2, pp. 91–175, 2007.

[40] A. T Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression," *Arch. Gen. Psychiatry*, vol. 4, no. 6, pp. 561–571, 1961.

[41] *Emotiv Epoc X*. Accessed: Mar. 2022. [Online]. Available: https://www.emotiv.com/epoc-x/

[42] *Empatica E4 Wristband*. Accessed: Mar. 2022. [Online]. Available: https://www.empatica.com/en-int/research/e4/

[43] J. Morris, "Observations: SAM: The self-assessment manikin; an efficient cross-cultural measurement of emotional response," *J. Advertising Res.*, vol. 35, no. 8, pp. 38–63, 1995.

[44] J. A. Coan, and J. J. B. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biol. Psychol.*, vol. 67, no. 1, pp. 7–50, Oct. 2004.

[45] H. Critchley and Y. Nagai, "Electrodermal activity (EDA)," in *Encyclopedia of Behaviorial Medicine*, M. D Gellman and J. R. Turner, Eds. New York, NY, USA: Springer, 2013, pp. 666–669.

[46] W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson, and D. L Filion, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, Aug. 2012.

[47] M. Swangnetr and D. B. Kaber, "Emotional state classification in patient–robot interaction using wavelet analysis and statistics-based feature selection," *IEEE Trans. Human-Mach. Syst.*, vol. 43, no. 1, pp. 63–75, Jan. 2013, doi: 10.1109/TSMCA.2012.2210408.

[48] P. J. Lang, "The emotion probe: Studies of motivation and attention," *Amer. Psychol.*, vol. 50, no. 5, pp. 372–385, May 1995.

[49] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951.

[50] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[51] X. Jiang, G. B. Bian, and Z. Tian, "Removal of artifacts from EEG signals: A review," *Sensors*, vol. 19, no. 5, p. 987, Feb. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/5/987

[52] N. Mammone, F. La Foresta, and F. C. Morabito, "Automatic artifact rejection from multichannel scalp EEG by wavelet ICA," *IEEE Sensors J.*, vol. 12, no. 3, pp. 533–542, Mar. 2012.

[53] O. Sporns, G. Tononi, and G. M. Edelman, "Connectivity and complexity: the relationship between neuroanatomy and brain dynamics," *Neural Netw.*, vol. 13, no. 8, pp. 909–922, Oct. 2000.

[54] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 73–78.

[55] A. K. Ahmadi, P. Moradi, M. Malihi, S. Karimi, and M. B. Shamsollahi, "Heart rate monitoring during physical exercise using wrist-type photoplethysmographic (PPG) signals," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 6166–6169.

[56] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA, 2016, pp. 785–794.

[57] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[58] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (TEAP)," *Frontiers ICT*, vol. 4, p. 1, Feb. 2017, doi: 10.3389/fict.2017.00001.

[59] *RobustScaler*. Accessed: Sep. 2022. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

SUDARSHAN PANT received the B.S., M.S., and Ph.D. degrees from Mokpo National University, South Korea. He is currently a Post-Doctoral Researcher with Chonnam National University. His research interests include machine learning, affective computing, healthcare AI, multimodal deep learning, AI ethics, and trustworthy AI.

HYUNG-JEONG YANG received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

EUNCHAE LIM is currently pursuing the Ph.D. degree in artificial intelligence convergence with Chonnam National University, South Korea. She was a Researcher with the Department of Artificial Bio-Robot, Osong Medical Innovation Foundation, from 2019 to 2020. Her research interests are exploratory data analysis, emotion recognition, natural language processing, and language model of the dialogue systems.

SOO-HYUNG KIM received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and ubiquitous computing.

SEOK-BONG YOO received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, Republic of Korea, in 2009, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea. He was a Senior Engineer with Samsung Electronics, Suwon, Republic of Korea, from 2015 to 2017, and a Senior Researcher with the Electronics and Telecommunications Research Institute, Daejeon, from 2017 to 2020. He is currently an Assistant Professor with Chonnam National University, Gwangju, Republic of Korea. His research interests include visual intelligence, image processing, computer vision, and immersive media synthesis.