IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 June 2022, accepted 19 August 2022, date of publication 24 August 2022, date of current version 2 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3201342

RESEARCH ARTICLE

1D Convolutional Autoencoder-Based PPG and GSR Signals for Real-Time Emotion Classification

DONG-HYUN KANG[1], AND DEOK-HWAN KIM[2], (Member, IEEE)
[1]Department of Electrical and Computer Engineering, Inha University, Incheon 22211, South Korea
[2]Department of Electronic Engineering, Inha University, Incheon 22211, South Korea

Corresponding author: Deok-Hwan Kim (deokhwan@inha.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through the Korean Government [Ministry of Science and ICT (MSIT)], XVoice: Multi-Modal Voice Meta Learning, under Grant 2022-0-00641; and in part by Inha University Research Grant.

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 of Inha University under Approval No. 170403-2AR.

ABSTRACT To apply emotion recognition and classification technology to the field of human-robot interaction, it is necessary to implement fast data processing and model weight reduction. This paper proposes a new photoplethysmogram (PPG) and galvanic skin response (GSR) signals-based labeling method using Asian multimodal data, a real-time emotion classification method, a 1d convolutional neural network autoencoder model, and a lightweight model obtained using knowledge distillation. In addition, the model performance was verified using the public DEAP dataset and the Asian multi-modal dataset 'MERTI-Apps'. For emotion classification, bio-signal data were window-sliced in 1-pulse units, and the label was reset to reflect the characteristics of the PPG and GSR signals. Simple data pre-processing, such as the prevention of loss and waveform duplication, was performed without using handcrafted features. The experiment showed that the accuracy of the proposed model using MERTI-Apps was 79.18% and 74.84% in the case of arousal and valence, respectively, for 3-class criteria, and the accuracy of the proposed model using DEAP was 81.33% and 80.25% in the case of arousal and valence, respectively, for 2-class criteria. The accuracy of the lightweight model was 77.87% and 73.49% in the case of arousal and valence, respectively, for 3-class criteria and its calculation time was reduced by more than 80% compared to the proposed 1d convolutional autoencoder model. We also confirmed that the proposed model improved computational time and accuracy compared to previous studies using MERTI-Apps and the lightweight model used in limited hardware environments enabled fast computation and real-time emotion classification.

INDEX TERMS PPG, GSR, 1D convolutional autoencoder, real-time, knowledge distillation.

I. INTRODUCTION

Real-time emotion recognition and classification are essential for efficient human-robot interaction. Human emotions are sensitive to external influences, complex thoughts, and physical characteristics. Therefore, the acquired data require no information loss, a maintainable pre-processing process, and a compact model setting that can quickly produce results. Real-time emotion recognition-based research on human facial expressions and voice signals that appear on the surface has been actively conducted, but facial expressions and voices are human-controlled somatic nervous systems (SNS) and cannot be viewed as accurate inner emotion recognition. Therefore, accurate inner emotions can be recognized using autonomic nervous system (ANS) signals, which cannot be controlled by humans [1], [2]. It is possible to research, develop, and apply human-robot interaction using inner human emotions in various fields. Complex-correlated emotions, which result from internal and external environmental factors and personal characteristics, affect the accuracy of emotion recognition and classification. Therefore, using the fusion technology of bio-signals (ANS), facial expressions,

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

and voice data (SNS), it is possible to build a strong emotion recognition model by improving the internal and external emotion recognition and classification performance [3]. For example, an accurate internal emotion recognition technology can provide individual standards according to the pilot's emotional bio-signals during training for complex airplane piloting [4] and can also be applied in the medical industry for patient rehabilitation [5].

To acquire various bio-signals without loss of information, invasive and non-invasive methods are available. In the invasive type, an accurate signal can be measured by inserting a sensor and needle directly into the subject's body; however, in this case, side effects and safety concerns arise. Therefore, the non-invasive method in which electrodes and sensors are attached to the skin of a subject, is the most commonly used approach [6], [7]. As shown in Figure 1, ANS signals acquired from human skin can be of different types depending on the body part and acquisition method.

- Electroencephalogram (EEG) shows the electrical activity of the brain
- Electrocardiogram (ECG) measures the electrical activity of the heart
- Photoplethysmogram (PPG) is a measure of the reflected light from blood vessels
- Galvanic skin response (GSR) measures sweat output and changes in body temperature.
- Electromyography (EMG) measures electrical signals generated by skeletal muscles

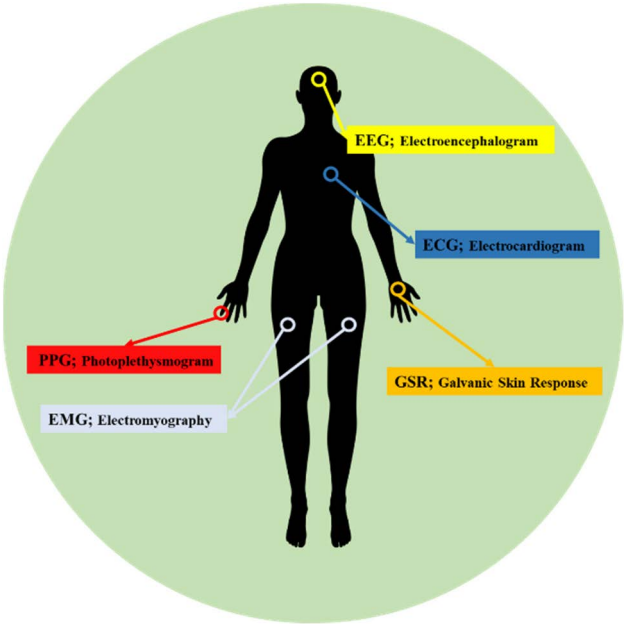

FIGURE 1. Bio-signals of an ANS based on measurement location and type.

As described above, depending on the measurement method for each bio-signal, the user's convenience, commercialization, and specific regularity of the signal appear differently. In addition, the user's usability and efficiency of these methods obtained by the users may significantly vary depending on the number and locations of the electrodes used to acquire various bio-signals. Because the information carried by these signals as well as the methods of utilizing this information are diverse, several trials and studies have been conducted from various perspectives. EEG signals use many electrodes, with up to 64 channels, and contain complex information. Therefore, to extract features for emotional information, Mantini *et al.* used the power spectrum density (PSD) of the EEG signals [8]; moreover, Topic *et al.* extracted the features of sustained emotional information using their topography [9] The ECG signal, which measures the electrical activity of the heart, includes blood pressure and heart rate information. Golgowski *et al.* studied abnormal heart detection using wavelet transformation [10], and Cai *et al.* and Puurtinen *et al.* classified emotions using discrete wavelet transform (DWT) and tabu search (TS) algorithms [11], [12]. The PPG signal is a record of volumetric variations in blood circulation and includes information on changes in blood flow and regular heartbeat based on the luminance reflected from the blood vessels. Rakshit *et al.* extracted heart rate variability (HRV) from PPG signals and classified them based on emotions [13]. Lee *et al.* performed rapid emotional classification using simple pre-processing of short-length PPG signals and a 1D convolutional neural network (CNN) model [14]. The GSR signal includes information on sweat output and body temperature changes according to emotional changes and is closely correlated with the physical domain [15]. Ganapathy *et al.* proposed a CNN model to extract 38 features from the time-frequency domain using Fourier transforms [16], and Susanto *et al.* conducted an emotional classification using a 1D CNN and residual bidirectional GRU [17].

This paper proposes a multimodal emotion classification method that can improve the accuracy and stability of emotion classification [18] and user convenience by using two bio-signals. Furthermore, developed models can be used in real-time and in limited hardware environments. This paper introduces related research and techniques in Section II, and Section III describes a bio-signal-based database. Section IV proposes real-time emotion classification and deep learning model techniques based on PPG and GSR signal pretreatment, considering user convenience. Section V describes the experimental environment and the performance results. Finally, conclusions and future research plans are presented in Sections VI and VII.

II. BACKGROUND

This section describes emotion recognition and classification in SNS and ANS and deep learning in a limited-hardware environment.

A. EMOTION RECOGNITION AND CLASSIFICATION BASED ON THE SNS

The most commonly used data for recognizing and classifying emotions in high-dimensional domains are facial

expressions and voices. Emotion and intention recognition based on facial expressions exhibits high performance and various performed. The face contains information about human conditions such as shape of eyes, facial expressions, changes in skin color, and the forehead. The main steps for facial expression recognition and classification include facial image pre-processing, feature extraction, and feature classification [19]. There are two feature extraction methods: geometric and face-shape feature extraction. The geometric features include facial components such as the mouth, eyes, nose, and chin, whereas the facial shape features consider specific areas, including changes in the face such as wrinkles.

Cootes *et al.* proposed a method for matching statistical appearance models with images [20]. The method utilizes the learned correlation between the errors in model parameters and those in results to construct an efficient iterative matching algorithm that controls changes in the shape and gray-level learned in the training set. Phavish *et al.* explored the classification of human facial expressions through a deep-learning approach using CNN models. The shape and location of the facial composition information were extracted using an edge detection framework and a stochastic classifier was used for facial expression classification [21]. Collecting photo- and video-based facial expression data is expensive, and there are concerns regarding personal information leakage.

Because voice signals are expressed in various forms based on age, region, and language, the acquired data becomes complicated. Therefore, it is necessary to proceed with feature extraction through the standardization and pre-processing of individual characteristics. Albornoz *et al.* conducted speech recognition and classification using a two-stage hierarchical classifier and studied the prosodic and spectral characteristics of speech signals and performed speech-based emotion classification using acoustic characteristics [22]. Eyben *et al.* developed an LSTM-RNN model that did not require emotion recognition or separate window slicing; it was based on real-time voice signals. In our previous work, we combined acoustic and linguistic features through feature extraction and fusion [23]. Voice-based emotion recognition and classification has lower risk of personal information leakage compared to its image-based counterpart. However, it cannot be applied to environmental factors such as libraries, construction sites, and underwater environments, where human behavior is restricted, and to those persons with disabilities who have difficulty in vocalization.

B. EMOTION RECOGNITION AND CLASSIFICATION BASED ON THE ANS

Because the bio-signals acquired from the ANS consist of electrical signals transmitted through muscles and bones, external environmental influences are minimized. Therefore, it is possible to avoid the problems observed in the case of facial expression and voice signals, and technology fused with additional bio-signals exhibits strong emotion recognition and classification performance and performs various functions [24], [25]. Unlike the user-controllable SNS, the ANS can accurately recognize and classify internal emotions using uncontrollable signals. However, using a single bio-signal for inner emotion recognition limits the accuracy and size/quality of learning data. Therefore, a multimodal emotion recognition and classification method based on multiple bio-signals has been proposed for improved performance.

Hui *et al.* proposed a decision recognition fusion model to combine the continuity of dimensional emotion recognition. In [26], first, pre-processing and feature extraction of non-invasive EEG and ECG signals were performed; subsequently, emotion recognition was performed based on the probabilistic neural network (PNN) and the experimental results showed that the results achieved using multiple signals outperformed those obtained using a single signal.

Yang *et al.* extracted EEG and PPG signals into 11 statistical time zones and five frequency zones and performed emotion classification using a 1D CNN. Furthermore, in our previous study, we calculated the pulse transit time (PTT) between the ECG and PPG signals to improve the final classification accuracy [27]. However, it required a minimum signal latency of 10 s and a maximum of 60 s for feature extraction and recognition of emotions. A signal latency of more than 10 s is unsuitable for real-time emotion recognition systems and it limits the usefulness of studies that fuse different signals (images, voices, etc.).

Ayata *et al.* used random forest, k-nearest neighbors (KNN), support-vector machines, and ensemble learning methods used in classification and regression analysis to evaluate PPG and GSR signals. They extracted features from GSR and PPG signals measured for 3 s and 8 s, respectively, resulting in an accuracy of 72.06% based on arousal and 71.05% based on valence [28]. In addition, the correlation between the PPG and GSR signals in the arousal and valence regions was confirmed. Lee *et al.* used the characteristics of the normal to normal (NN) time domain for heart rate variation in PPG signals and the frequency domain characteristics through normalization. They achieved an accuracy of 82.1% based on feature-based arousal and 80.9% based on valence extracted from a PPG signal of 10-s duration [29].

In these studies, although the minimum time required for emotion recognition has been shortened compared to that in previous studies, the researchers agree that additional feature extraction is required for data processing. In a previous study, various feature extraction methods improved bio-signal-based emotion recognition and classification performance, but handcrafted tasks such as time-frequency domain feature extraction still require additional computation time and have the problem of limited usability. A fast data processing method and a method for maintaining signal characteristics are required to solve the existing problems in real-time emotion recognition.

C. USE OF DEEP LEARNING MODELS IN LIMITED-HARDWARE ENVIRONMENTS

Advances in computer hardware and software have significantly influenced the development of artificial intelligence

research. The multilayer perceptron [30], developed in the 1980s, overcomes the limitations of existing linear classifiers by adding hidden layers to the existing perceptron structure. However, accessing it was difficult for various industries and products because of the limitations in terms of hardware performance at the time, learning time. Recently, because high-performance GPUs capable of parallel computation and with multiple cores have been developed, research on artificial intelligence models using rapid learning and more complex algorithms has progressed. However, a model designed to perform operations on a vast amount of data is inappropriate for application to the resources of various embedded devices for human-robot interaction [31]. Various studies have been conducted to utilize the methods that help reduce the number of model parameters, for example, deep learning model compression, knowledge distillation, and pruning, for deployment and use in embedded systems with limited hardware resources, as shown in Figure 2. Han *et al.* pruned layer units with low values based on a specific threshold, and then conducted repeated re-learning [32]. The overall number of parameters was reduced from 61M to 6.7M, which was reduced to approximately 90%, and the accuracy error from the original model was not higher than 0.1%. Han *et al.* pruned layer units with low values based on a specific threshold, and then conducted repeated re-learning [32]. The overall parameter was reduced from 61M to 6.7M, which was a reduction of approximately 10%, and the accuracy error from the existing model was not significantly different, at 0.1%. Knowledge distillation techniques can transfer knowledge from a large model, the teacher, learned through ensemble techniques, to build students, a small model that exhibits the same performance. Ho *et al.* extracted the knowledge of complex teacher models into light student models, after which, they self-trained these student models to perform multilabel lung disease classification [33]. The aforementioned model lightning technology can be utilized in limited-hardware environments using biological signals, embedded boards, and modules for biological signal acquisition. In addition, it could be applied to parallel systems that used video and voice signals together because, compared to existing models, it could minimize accuracy loss and reduce the required memory and computation volume.

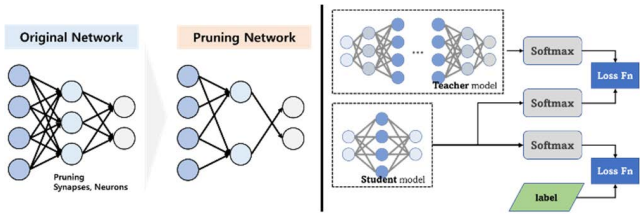

FIGURE 2. Pruning techniques (left) and knowledge distillation (right).

III. BIO-SIGNAL (ANS)-BASED DATA

Psychologists study emotions from two perspectives: general and complex [34], [35], [36]. General emotions are divided into happiness, anger, sadness, and fear, and complex emotions have a multidimensional structure and are constantly changing. As shown in Figure 3, the structure widely used in the study of bio-signal-based emotion recognition and classification consists of the arousal (physical) and valence (emotional) bases of the two-dimensional structure proposed by Russell and James [37]. By expressing emotions in a two-dimensional structure, complex and general emotions can be subdivided into arousal and valence scales. As shown in Table 1, representative bio-signal public data for emotion analysis were labeled with the same two-dimensional domain-based emotion as the DEAP dataset [38] and MAHNOB-HCI database [39]. To create accurate training data for the emotions of the subject, human emotion labels are important [40]. In addition, according to Western standards, two public data (DEAP, MAHNOB-HCI) were mainly used for emotion recognition and classification based on bio-signals.

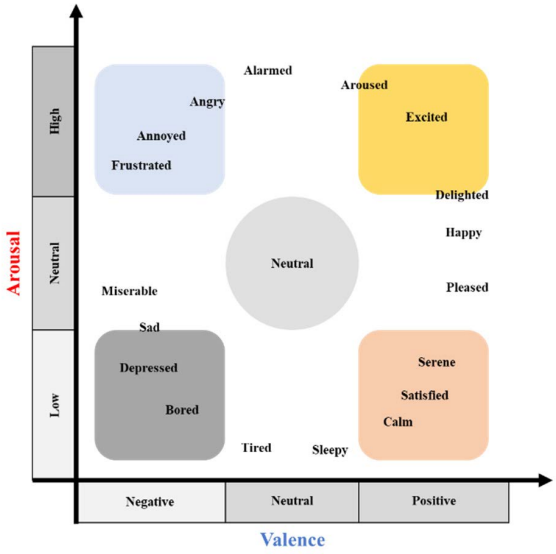

FIGURE 3. Arousal and valence domains in the 2D structure.

TABLE 1. Public data (DEAP, MAHNOB-HCI).

DEAP DATA SET	
Participants	32 (16 male, 16 female)
Video	40 videos (60 second)
Bio-signals	EEG, EOG, EMG, GSR, PPG
Labeling	Self-Assessment Labeling (Arousal, Valence)
MAHNOB-HCI DATABASE	
Participants	27 (11 male, 16 female)
Video	20 videos (34 to 114 second)
Bio-signals	EEG, EOG, GSR
Labeling	Annotation Labeling (Valence)

A. COMPARISON OF LABELING TECHNIQUES FOR A QUANTITATIVE APPRAISAL

To evaluate emotions in multidimensional domains, DEAP uses self-assessment labeling and MAHNOB-HCI uses observer annotation labeling.

For self-assessment labeling, the arousal scale ranged from boredom to excitement and the valence scale ranged from 1 to 9, ranging from sadness to joy. The subjects watched a 60-s-long emotion-evoking video and marked self-assessment labeling [41]. Therefore, it was difficult expressing changes in emotion while watching, and there was a limit to subdividing the emotion standard of the entire video.

Observer annotation labeling showed that fine-grained labeling could be used for short-term bio-signal data, as shown in Figure 4, because trained observers measured emotional changes in real time while viewing the subject's face-facing videos. Consequently, observer annotation labeling can note the change in emotions until the experimenter elicits an emotion. Therefore, observer annotation labeling, in which the learning data labels show variations in emotions, was appropriate for real-time emotion recognition and subdivided emotion classification.

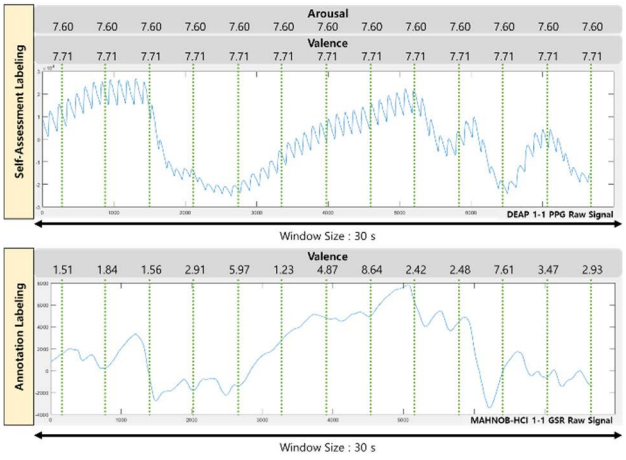

FIGURE 4. Distribution of data labels of the same 30-s length; (Top) Self-Assessment of the DEAP dataset; and (Bottom) annotation of the MAHNOB-HCI dataset.

B. MERTI-APPS DATABASE

In this study, we performed multimodal emotion classification based on PPG and GSR signals using the DEAP dataset and the Asian multi-modal database MERTI-Apps [42], which is detailed in Table 2. Asian multimodal data, MERTI-Apps, extracted data on 15 emotion-evoking videos from 62 subjects (28 males, 34 females). Self-evaluation and observer annotation labeling were conducted by five pre-trained observers. The data obtained were measured for Experiment 1 (EMG, EOG, PPG, GSR), Experiment 2 (EEG, EMG, EOG), and Experiment 3 (EEG, EOG, PPG, GSR). The acquired emotion-induced video-based labeling scale ranged from −100 to 100.

IV. PROPOSED METHOD

Conventional bio-signal-based emotion recognition and classification methods generally involve data pre-processing and extracting signal features from the time-frequency domain.

TABLE 2. Public data (MERTI-Apps).

Participants	62 (28 male, 34 female)
Video	40 videos (60 second)
Experiment 1	
Bio-signals	EOG, EMG, GSR, PPG
Emotional response to videos	5 videos (Sad 1, Happy1, Angry 2, Scared 1)
Labeling	Annotation Labeling (Arousal, Valence) Self-assessment Labeling (Arousal, Valence)
Experiment 2	
Bio-signals	EEG, EMG, EOG
Emotional response to videos	5 videos (Sad 1, Happy1, Angry 1, Scared 1, Neutral 1)
Labeling	Annotation Labeling (Arousal, Valence) Self-assessment Labeling (Arousal, Valence)
Experiment 3	
Bio-signals	EEG, EOG, PPG, GSR
Emotional response to videos	5 videos (Sad 1, Happy2, Angry 1, Scared 1)
Labeling	Self-assessment Labeling (Arousal, Valence)

However, conventional methods using additional handcrafted feature extraction are too complex to be applied to a real-time system. In this paper, we propose a real-time emotion classification method using PPG and GSR signals for user convenience and a labeling method to reflect the bio-signal characteristics. The proposed method is divided into three main categories: data pre-processing with window slicing and prevention of waveform duplication and loss, training data label setting, and weight-reduced model learning for emotion classification. Without additional handcrafting, the window sliced by one pulse unit of PPG and GSR signals is mainly used as input for a multimodal neural network model.

As the basic model in emotion classification, a 1D convolutional autoencoder model consists of an input (encoder) and output (decoder) of the same structure.

A. BIO-SIGNALS FOR USER CONVENIENCE AND USABILITY

To improve smooth interaction between humans and machines to which bio-signals are applied, convenience, and usability in daily life, the number, attachment position, and accessibility of the acquisition electrodes should be considered. Among the representative bio-signals, up to 64 electrodes were acquired, as illustrated in Fig. 5(a). The EEG signal shows high emotion classification accuracy based on information and data collected from many electrodes. However, it is not easy to use in real time or daily life. ECG signals were acquired up to six electrodes, as shown in Figure 5(b). Although high-performance emotion classification is possible owing to direct heart change-based sensitivity and regularity, it is inconvenient to attach directly to the chest near the heart and have low access to the external environment. EMG signals can be acquired using various body parts and a small number of electrodes. However, because some

muscle signals contain movements that humans can control, it is difficult to see them as signals suitable for inner emotion classification.

Therefore, in contrast to other signals requiring large numbers of electrodes, this study used a PPG signal with information on heart rate and blood flow and a GSR signal with physical details such as body temperature change and sweat emission with two electrodes. The two bio-signals can be measured and acquired at the user's fingertips, as shown in Figure 5(c), improving convenience and accessibility in daily life.

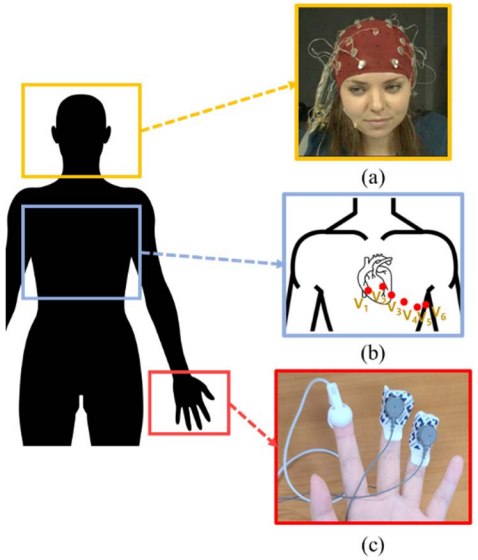

FIGURE 5. Actual acquisition method and location according to bio-signal: (a) EEG [39]; (b) ECG; (c) PPG, GSR.

B. DATA PRE-PROCESSING OF PPG AND GSR SIGNALS

Raw PPG signal data are more convenient and usable than ECG signals, but the noise of high-frequency components, such as power source noise and low-frequency components caused by capillaries, must be removed. In this study, the Butterworth filter, which is a simple linear frequency-domain filter, was applied. The dynamic noise was reduced through high-order polynomial fitting, baseline fluctuation noise of the data, and moving average filter. The same pre-processing was repeated for the GSR signal.

1) BUTTERWORTH FILTER [42]

The Butterworth filter is a frequency filter with a flat pass-band amplitude spectrum and a smooth transition band. The simple linear frequency filter is expressed as follows: the higher the order N, the smaller is the transition band and the sharper is the transition.

- The basic formula for filter approximation is

$$|H(j\omega)| = \frac{K}{\sqrt{1+\varepsilon^2 f(\omega^2)}}. \tag{1}$$

- Butterworth filter approximation function ($f\left(\omega^2\right) = \omega^{2n}$):

$$|H(j\omega)| = \frac{K}{\sqrt{1+\varepsilon^2\omega^{2n}}}. \tag{2}$$

- Normalized Butterworth filter approximation function:

$$(K=1, \varepsilon=1, \omega_c=1), \quad |H(j\omega)| = \frac{K}{\sqrt{1+\omega^{2n}}}. \tag{3}$$

2) POLYNOMIAL CURVE FITTING [43]

Higher-order polynomial fittings are commonly used to address outliers and errors in the data after biometric signal acquisition. The correlation between the input and output variables can be analyzed and predicted to reduce the error rates and obtain the most appropriate data. The higher-order polynomial fitting is expressed as follows:

- Nth-order polynomial equation:

$$y(x, \omega) = \sum_{j=0}^{N} \omega_j x^j. \tag{4}$$

3) MOVING AVERAGE FILTER [44]

A general moving filter is the average of the sum of the data measured to the present time. If the measured value changes continuously with time, then a moving filter is not appropriate. Therefore, a moving average filter is used as a method to remove noise from the data and average the dynamic data. The average of the most recent N values is used instead of the average of all the data. The moving-average filter is expressed as follows:

- Moving average filter for N data:

$$\bar{x}_k = \frac{x_{k-n+1} + x_{k-n+2} + \ldots + x_k}{n} = \bar{x}_{k-1} + \frac{x_k - x_{k-n}}{n}. \tag{5}$$

4) PRE-PROCESSED SIGNAL

Figure 6 shows the pre-processing results of the PPG and GSR signals. After bio-signal pre-processing, the 10-s-long PPG signal reference peak and average heart rate information were collected and stored for accurate window slicing.

C. WINDOW SLICING FOR TRAINING DATA

GSR signals containing information on body temperature changes and sweat emissions generally have an irregular form. However, the blood flow rate that changes according to the heartbeat always has a distinct regularity in the PPG signal. To properly divide into cycles, such as heart rate and blood pressure fluctuations according to emotional changes, signal division in short waveform units is performed based on regular PPG signals. However, the acquired PPG signals should be divided into precise waveform units considering personal characteristics because regular cycles, such as blood flow and heartbeat according to the experimenter, vary widely from person to person. When unilaterally dividing according to sample size, the peak value of the PPG signal and the bottom data overlap and loss occurs based on the window size and sampling rate, as shown in Figure 7.

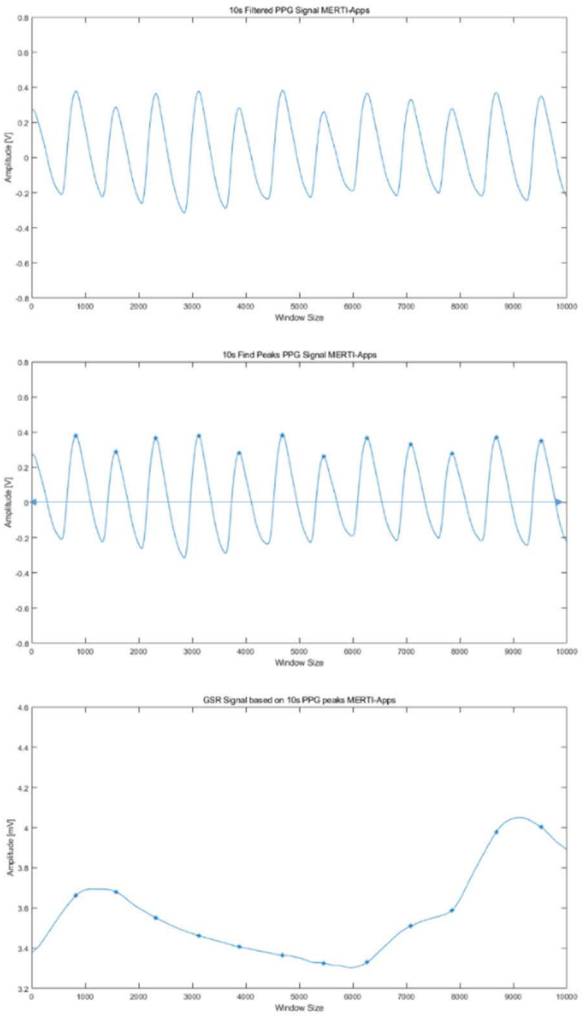

FIGURE 6. 10s-long PPG, GSR signal pre-processing result: (top) pre-processed PPG signal; (middle) peak-to-peak of PPG signal; and (bottom) GSR signal of the same section as the peak value of PPG signal.

In this study, the zero-zone was set and divided through the PPG signal reference peak value and heartbeat information stored in the pre-processing process, in consideration of personal characteristics after completion of pre-processing. To accurately separate the PPG and GSR signals, the sample size was divided equally based on the peak value of the PPG signals stored in advance. As a result, the sample size of the PPG and GSR signals measured at 1 kHz was maintained at 1,100. In addition, the heart rate was used to treat it as a zero zone, except for one waveform section, to prevent duplication. Figures 8 shows that PPG and GSR signals appear without redundancy and loss as 1.1 seconds after the final treatment is completed.

- Zero-zones setting to prevent duplication:

$$zero\ zone = \frac{\left(\frac{total\ frequency\ sample\ size}{heart\ rate} - sample\ size\right)}{2}. \quad (6)$$

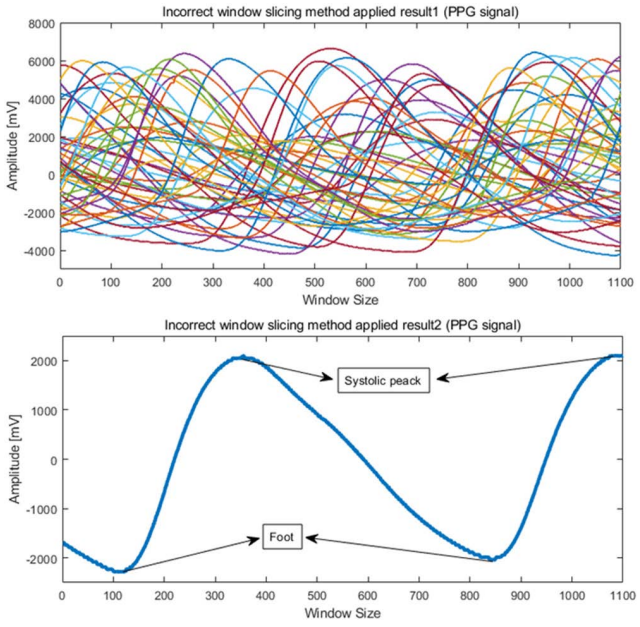

FIGURE 7. PPG signal processed with incorrect window slicing method.

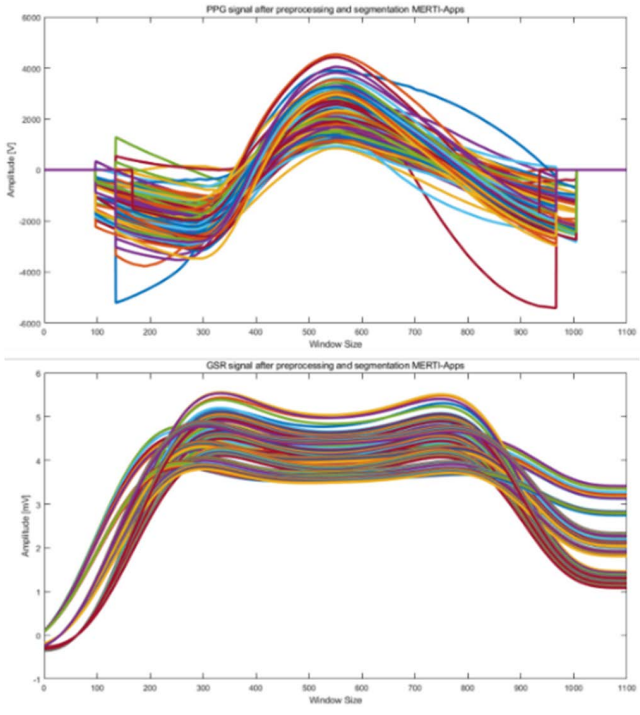

FIGURE 8. Final PPG, GSR signal output (sample size 1100); (Top) Results of the PPG signal; and (bottom) results of GSR signal.

D. LABEL SETTING FOR TRAINING DATA

Existing observer annotation labels show that trained commentators express the experimenter's facial expressions at 0.25 s intervals, and the scale provides labels that vary continuously over time. Therefore, for a short emotional classification in unit of 1.1 seconds, the label of the DEAP and MERTI-Apps database was defined as three classes based on

arousal and valence, as shown in Table 3. The overall range of self-assessment labels in the DEAP dataset was from 1 to 9, and in this study, three classes were defined: positive (7–9), neutral (4–6), and negative (1–3). In addition, the overall range of the observer annotation labels in the MERTI-Apps database was expressed as -100 to 100, in which three classes were defined: positive (20 to 100), neutral (-20 to 20), and negative (-100 to -20).

TABLE 3. 3-class criteria for emotional classification in 1.1 seconds.

Data		DEAP	MERTI-Apps
Labeling		Self-assessment	Annotation
Existing Label Scope		1 to 9	-100 to 100
Emotion classification class	positive	7 to 9	20 to 100
	neutral	4 to 6	-20 to 20
	negative	1 to 3	-100 to -20

Labeling of the extracted waveform unit PPG and GSR signals was performed based on the peak-to-peak values of the PPG signals stored in advance using the section set, as shown in Table 3, and the observer annotation label of MERTI-Apps. In the MERTI-Apps database, a total of 32,000 segments were extracted from the data and pre-processed through five 1∼5 min-long emotion-evoking images for 62 experimenters. In the previously pre-processed and divided 32,000 segments, the label average within the sample size section was used for the label, and data that did not match the experimenter's evaluation result were not used.

The diversity of the final label distribution is confirmed, as shown in Table 4. The self-assessment labeling technique is expressed as unified emotion labeling because the experimenter proceeds after watching all emotion-evoking videos. Therefore, as shown in Table 4, because simple labels based on five emotion-evoking videos are provided to the model, diversity for learning is not provided. The observer annotation labeling technique can provide a model with a wide variety of learning data based on 32,000 segments because the third observer reflects the change in the face according to the emotion change in the label in real time.

TABLE 4. Distribution of Segment Label for ppg, gsr signals.

		Annotation labeling	Self-assessment labeling
Number of data extracted		32000 segments	
Label (arousal)	high	14,489	2
	neutral	11,190	2
	low	6,321	1
Label (valence)	positive	6,716	1
	neutral	5,289	1
	negative	14,227	3

E. THE MODEL SETTING FOR EMOTION CLASSIFICATION

1) MULTIMODAL 1D CONVOLUTIONAL AUTOENCODER

A 1D convolutional autoencoder was used to perform feature extraction and emotion classification based on the pulse units of the PPG and GSR signals. The model consisted of two convolution layers and two pooling layers, and the encoder and decoder had the same structure. The extracted data were transmitted to the fully connected layer through the concatenation layer. The information on each layer is shown in Figure 9, and Adam (learning rate: 0.0008) was used as the optimizer. By setting the batch size to 16, the repetition learning was executed 100 times. 64% of the total data were learned as training data, 16% as validation data, and 20% as performance validation data.

2) LIGHTWEIGHT MODEL OBTAINED USING KNOWLEDGE DISTILLATION

A knowledge distillation technique was applied to two bio-signal-based 1D convolutional autoencoder models (teachers) in an environment where real-time programs and hardware performance are limited. The knowledge distillation technique used in this study is a method for distributing and using the knowledge of a pre-trained complex deep neural network (teacher model) to a simple deep neural network (student model). To transfer knowledge to the student model, an output value softer than the actual output value was used by adding a hyperparameter (temperature) to the output value of the neural network through the softmax function of the teacher model. In addition, a loss function is constructed by comparing the teacher model's soft label and the student model's soft predicted value through the distillation loss value so that the same output of the student and teacher models is derived [45]. Knowledge distillation is expressed as follows:

- Probability value of the class through the last softmax layer:

$$q_i = \frac{exp(z_i)}{\sum_{j=1}^{k} exp(z_j)} \tag{7}$$

- Set soft and hard result values using temperature coefficient:

$$q_i = \frac{exp(z_i/temperature)}{\sum_{j=1}^{k} exp(z_j/temperature)} \tag{8}$$

- Distillation loss value (L:loss function, S:student model, T:teacher model, L_CE: cross-entropy loss, L_KD:distillation loss, θ:train parameters, τ:temperature)

$$L = \sum_{(x,y)\in D} L_{KD}\left(S\left(x, \theta_S, \tau\right), T\left(x, \theta_T, \tau\right)\right) + \lambda L_{CE}(\hat{y_s}, y) \tag{9}$$

As described above, knowledge of the existing teacher model was transferred to the 1D CNN model (student) with a simple structure, as shown in Figure 10. It consisted of convolutional and pooling layers, and an Adam optimizer (learning rate: 0.0005) was used. By setting the batch size

to 16, the repetition learning was executed 200 times. The final student model reduced the computation time by compressing the network and reducing the number of parameters while maintaining accuracy.

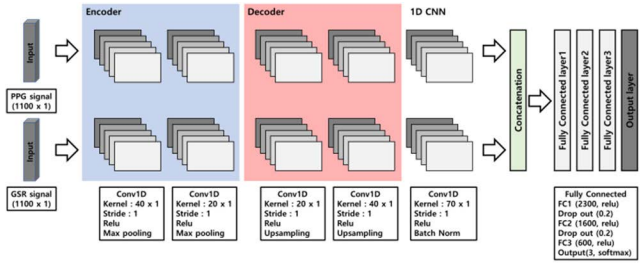

FIGURE 9. Proposed multimodal 1D convolutional autoencoder diagram.

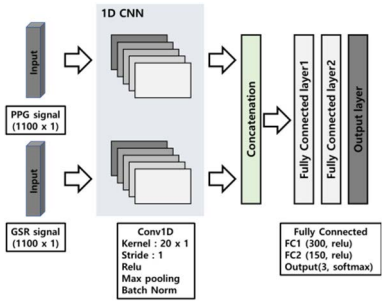

FIGURE 10. Configuration diagram of student model (1D CNN) used in the paper.

V. EXPERIMENTAL RESULT

This section presents the experimental environment, bio-signal-based emotion classification model composition, and experimental results. Figure 11 shows the overall composition of the aforementioned contents and results.

A. EXPERIMENTAL ENVIRONMENT

In this study, the MERTI-Apps and DEAP datasets, including both PPG and GSR signals, were used. The data used in this study were defined as 1100 samples(1kHz) on average through the pre-processing process mentioned earlier, peak of the signal, and prevention of duplication and loss by dividing the PPG and GSR signals into one pulse unit through average heart rate detection. Typical PC configurations used for model learning include Intel Corei7-9700, RAM: 64GB, OS: Windows 10, and GPU: Nvidia GeForce RTX 2080Ti. In addition, different hardware environments were configured to analyze the computational time required for the pre-trained final model.

B. COMPARISON OF BIO-SIGNAL-BASED PRECEDENT STUDIES AND RESULTS

To apply real-time bio-signals to an emotion recognition system, pre-processing and feature extraction are required in the same way as the learning data of the model. The feature extraction process has limitations such as additional computation time and low utilization. Therefore, this study compared the data-processing time and accuracy of the handcrafted feature extraction method and the proposed method used in the Asian multi-modal data MERTI-Apps-based precedent study [8], [11]. For comparison of experimental results, the same dataset consisting of 1,100 samples for PPG data and 12 channels for EEG signal was used. In addition, time, frequency, and time-frequency domain characteristics were extracted from the data in the same way as the existing research method [27], [46]. For the 1-channel PPG signal, a scalogram that could check time change according to the frequency band was used [47].

TABLE 5. Performance of equipment used in the experiment.

Category	CPU	GPU	Memory
Nvidia Jetson AGX Xavier	8-core ARM v8.2 64-Bit	512-core volta with tensor cores	32GB 256-bit LPDDR4x
High-performance computer	Single Intel Xeon Silver 4214R 2.4GHz	NVIDIA Quadro RTX 6000 D6 24GB	64GB
Normal computer	Intel Core i7-9700	NVIDIA GeForce RTX 2080Ti	32GB

To compare the performance of previous studies [46], [47], feature extraction from the EEG signals acquired from 12 channels was performed in the time domain for changes over time and the frequency domain to understand the spatial resolution of the EEG signal. Four regions were divided using the PSD characteristics: slow alpha (8–10 Hz), alpha (8–13 Hz), beta (13–29 Hz), and gamma (30–50 Hz). Moreover, the average and absolute values were used. In the time-frequency domain, the EEG signal was decomposed over time through a discrete wavelet transform, and then their average and absolute values were extracted in the frequency range. The PPG signal was converted into a scalogram, from which noise was removed through data pre-processing. It was used through MATLAB's Toolbox, and a filter tank separated the band in the range of 0.3–5 Hz, and an octave with 48 sections was set.

A general PC environment was used to set the same experimental environment and data size, and the results were confirmed according to the experiment, as shown in Table 6. It took 4.91 s to extract features from the EEG signal and an average of 2.34 s for the PPG signal to be pre-processed and converted into a scalogram. The method without additional handcrafting and based on PPG and GSR signals proposed in this paper performs pre-processing and feature extraction within an average time period of 0.17 s, which was faster than the data processing time of existing precedent studies. In addition, the bio-signal should consider the minimum window size to proceed with the emotion classification. The time required for the final data process is the sum of the minimum waiting times for emotion recognition, data pre-processing,

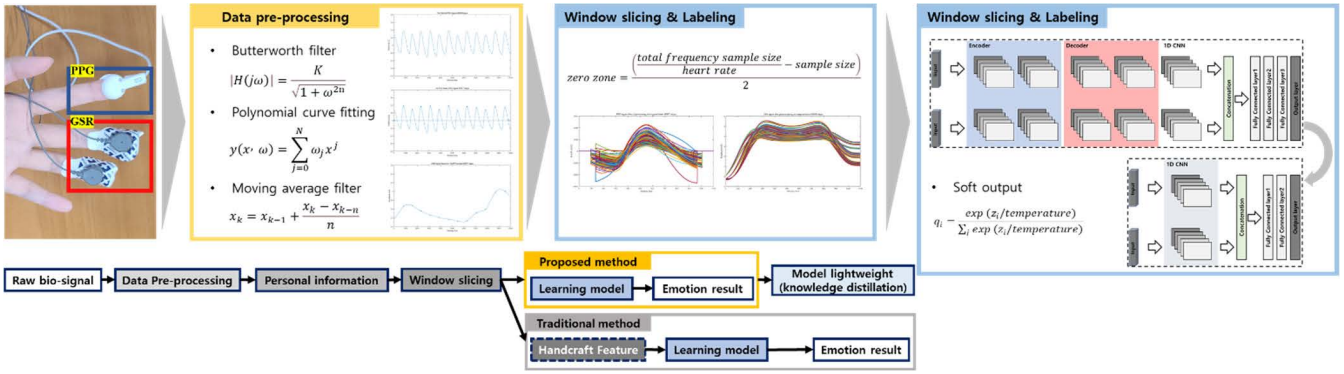

FIGURE 11. Proposed overall configuration diagram for emotion classification based on bio-signals (PPG, GSR).

feature extraction, and model calculation time. Therefore, when comparing the results excluding the model calculation time, the EEG signal required the most time at approximately 64.91s. The existing scalogram (PPG) method [47] is approximately 3.44 s faster than the power density feature extraction (EEG) method [46] but slower than the proposed method. In addition, the proposed method has twice the sample size as the scalogram feature extraction method, but requires a short time of approximately 1.27 s. The accuracy of emotion recognition and classification of the proposed method was 80.89% in the arousal domain, which was the highest among the three methods, and 81,25% in the valence domain, which was the second highest among the three methods. However, the difference in accuracy between the scalogram method and proposed method was approximately 1% in the valence domain, indicating the same performance.

This confirms that the proposed method can improve the processing time and performance by optimizing data pre-processing and window size compared with previous studies on emotion recognition and classification based on MERTI-Apps and Asian multimodal data.

TABLE 6. Comparison of bio-signal-based precedent studies and the time required for the proposed method.

Category	EEG [46]	PPG [47]	PPG, GSR
window size	60 s	1.1 s	each 1.1 s
feature extraction	handcraft features (time, frequency, time-frequency domains)	scalogram	sliced and filtered raw data
pre-processing & feature extraction time	4.91 s	2.34 s	0.17 s
total time required	64.91 s	3.44 s	1.27 s
deep learning model	bidirectional LSTM	2D CNN	1D convolutional autoencoder
emotional accuracy arousal/valence	78.00% / 81.00%	76.09% / 82.33%	80.89% / 81.25%

C. EMOTION CLASSIFICATION RESULT OF THE PROPOSED MULTIMODAL MODEL

Using the Asian multimodal data MERTI-Apps and the open dataset DEAP, the accuracy, recall, precision, and F1 score were derived by comparing the performance of the emotion classification model in the arousal and valence domains. In general, the accuracy, which is a representative indicator of a model's performance, is the proportion of correctly classified data to the total data. Accuracy can be a misleading indicator depending on the data used. Therefore, we used two additional performance indices: recall, which represents the ratio of data that the model classifies as true to true data, and precision, which represents the ratio of data that is true to data that the model classifies as true. Furthermore, recall and precision have a tradeoff relationship. The F1 score is an index that can identify the data imbalance between classes, classified as a harmonic mean of recall and precision.

- The proportion of correctly predicted data out of the total data:

(TP: true positive, TN: true negative, FP: false positive, FN: false negative),

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{10}$$

- The proportion of data that predicted factual data:

$$Recall = \frac{TP}{TP + FN}. \tag{11}$$

- The proportion of factual data among data predicted to be true:

$$Precision = \frac{TP}{TP + FP}. \tag{12}$$

- Harmonic means precision and recall.

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}. \tag{13}$$

In MERTI-Apps, emotion classes were divided into two (high, low) and three (high, neutral, low) in the arousal domain and two (positive, negative) and three (positive, neutral, negative) in the valence domain. The mean results for each group were compared. These results are presented in Table 7 and Figure 12. All performance index results of the teacher (1D convolutional autoencoder) model were confirmed to be 74.00% or higher, and Figure 12(a) shows the results of the teacher model for the three-class criteria: accuracy (79.18/74.84%), recall (73.61/72.44%), precision (76.66/74.31%), and F1 score (74.99/73.18%) based on

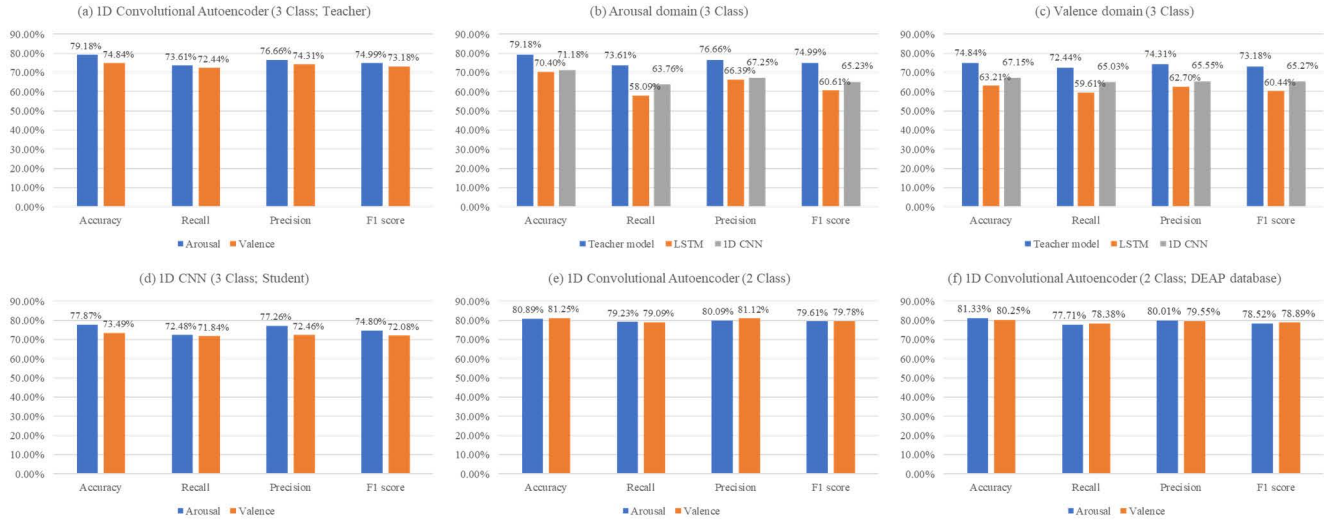

FIGURE 12. PPG, GSR signal-based emotion classification result graph; (a) 1D convolutional autoencoder (three-class, teacher); (b) 1D convolutional autoencoder (two-class); (c) 1D CNN (three-class, student); (d) arousal domain (teacher model, LSTM, 1D CNN); (e) valence domain (teacher model, LSTM, 1D CNN); and (f) result of using the DEAP dataset.

TABLE 7. Research results of emotion recognition and classification based on various bio-signals.

	Bio-signals	window size	Method	class	Accuracy	
					Arousal	Valence
Yoon, 2013 [51]	EEG	-	FFT (fast fourier transform) enhanced feature extraction and classification;	2	70.1%	70.9%
				3	55.2%	55.4%
Ayata, 2018 [28]	PPG, GSR	3~8 s	Random forest, k-NN, Decision tree; DEAP dataset	-	72.06%	71.05%
Mert, 2018 [48]	EEG	60s	Multivariate extension; ANN; DEAP database	2	75.00%	72.87%
Lee, 2019 [14]	PPG	1.1s	1D CNN; DEAP database	2	76.20%	75.30%
Chao, 2019 [49]	EEG	3s	Multiband feature matrix; CapsNet; DEAP database	2	66.73%	68.28%
Xing, 2019 [50]	EEG	60s	SAE+LSTM; DEAP database	2	81.10%	74.38%
Wang, 2019 [52]	EEG	-	STFT (short time fourier transform); NMI-based channel selection	2	73.64%	74.41%
Lee, 2020 [29]	PPG	10s	1D CNN; Selected statistical feature; DEAP database	2	80.90%	82.10%
Raheel, 2021 [53]	EEG, GSR, PPG	60s	Time, Frequency domain; DEAR-MULSEMEDIA database	2	85.18%	76.54%
Hwang, 2022 [46]	EEG, EOG, GSR, PPG	60s	ACO-bidirectional LSTM; MERTI-Apps database	-	78.00%	81.00%
Our	**PPG, GSR**	**1.1s**	1D Convolutional autoencoder; DEAP database	2	81.33%	80.25%
			1D Convolutional autoencoder (teacher model); Annotation labeling; MERTI-Apps database	2	**80.89%**	**81.25%**
				3	**79.18%**	**74.84%**
			1D CNN (student model); Annotation labeling; Knowledge distillation; MERTI-Apps database	3	**77.87%**	**73.49%**
			LSTM; Annotation labeling; MERTI-Apps database	3	70.40%	63.21%
			1D CNN; Annotation labeling; MERTI-Apps database	3	71.18%	67.15%

arousal and valence. Compared with the LSTM and 1D CNN models in Figures 12(b) and 12(c), respectively, it was confirmed that the model showing the highest level of classification performance showed that the teacher model was superior in all performance indicators. LSTM, a comparative model, is suitable for research that includes time series and sequential input/output, but it was confirmed that the performance was low for input data without a separate handcrafted process for feature extraction and PPG and GSR signals in short waveform units. The 1D CNN model, which represents the second-best performance, utilizes the sparsity of parameter sharing and connection to show that it is strong for spatial information, form maintenance, and automatic feature extraction of data but performs lower than the teacher model based on the same size. In addition, in the open dataset DEAP, it was confirmed that the proposed method had a high performance

with a window size of 1.1 s, and an accuracy of 81.33% based on arousal and 80.25% based on valence.

To proceed with real-time emotion classification, the knowledge of the pre-trained complex teacher model showing the highest performance was transferred to the simple structure of the student (1D CNN) model. As shown in Figure 12(d), the results of the performance index of the student model for the three-class criteria showed a maximum error rate of 1.65% based on arousal and 2.49% based on valence. This confirms that the knowledge of the teacher model was correctly transferred to the student model. When the parameters of the existing teacher model used for emotion classification based on PPG and GSR signals and the student model to which knowledge distillation was applied, were compared, it was confirmed that the number of parameters in teacher models was approximately 358M and that of the student model was approximately 6M, which indicates a 98% reduction. In addition, the time required for emotion classification based on 6,400 segments of test data was reduced from 4.72 s to 1.14 s. To compare the performance of the temporal bio-signal-based emotion recognition and classification model, as shown in Table 8, based on 6,400 segments and a 1,100 Hz sample in three hardware environments. Although many computational and hardware (CPU, GPU, etc.) resources are required to use multimodal bio-signal-based emotion classification teacher models, the student model enabled a model weight reduction of 1.87 s on Nvidia Jetson AGX Xavier, 1.14 s on typical PCs, and 0.97s on high-performance PCs, reducing the time spent on all environments by less than half. As a result, applying the pre-learned model to the lightning process confirmed its usability, even in a limited hardware environment, by exhibiting similar performance with reduced computation time.

TABLE 8. Experimental results in different hardware environments.

Model	Average processing time (6,400 segments)		
	Nvidia Jetson AGX Xavier	Normal computer	High-performance computer
Teacher	11.41 s	4.72 s	3.21 s
Student	1.87 s	1.14 s	0.97 s

VI. CONCLUSION

In this paper, we proposed a method for emotion classification using multimodal 1D convolutional autoencoder models based on two PPG and GSR signals. As shown in Table 7, research on emotion recognition and classification based on various bio-signals was conducted. In the study of emotion classification based on bio-signals, there is an appropriate window size for input data, and there is a minimum waiting time for feature extraction and regularity of the acquired signal. Therefore, the real-time characteristics decrease owing to the essential waiting time for recognition. Moreover, the accuracy was reduced due to use of a single bio-signal. Yoon *et al.* used FFT (fast Fourier transform analysis) to analyze the signal in the time domain, and applied shape selection based on the Pearson correlation coefficient [51]. However, there was a severe decrease in accuracy as the number of classes increased. Wang *et al.* expressed the EEG signal as a spectrogram in the time-frequency domain using a short-time Fourier transform (STFT) to improve the time interval limit of the existing FFT and used it for emotion recognition [52]. However, an additional handcrafted course was also included. Xing *et al.* extracted correlations and features from EEG signals acquired over many channels and reduced the computational time using SAE-based linear EEG mixing models [53]. However, the window size for emotion recognition is 60 s, which makes it difficult to apply to real-time systems. Lee *et al.* prevented additional time loss by automatically extracting signal characteristics from a single PPG signal in a short waveform unit of 1.1 s using a 1D CNN and showed the possibility of short-term emotion classification in the two classification results [14]. However, labeling integrated with a single signal has limitations in accuracy and stability as the detailed emotion recognition and classification class increases. This study confirmed the possibility of high-level emotion classification without the need for another handcrafted process for extracting the signal characteristics. Using this method, an improved emotion classification performance was confirmed. As a result of the experiment, in DEAP, the proposed method with arousal/valence criteria of 81.33/80.25% showed fast and high performance. In addition, with subdivided labeling, MERTI-Apps showed a high-performance index even with the additional class increase, with arousal/valence criteria of 80.89/81.25% in two classes and 79.18/74.84% in three classes, compared to previous studies. The model lightning work was carried out to fuse the real-time system and additional signals (image, video, etc.). The computational time of the model was confirmed on embedded boards commonly used in robots for actual vehicle and human interaction. By reducing the parameters and sizes of the model, less than half of the computational time of the existing model was confirmed in all environments.

Finally, this paper discussed a study conducted based on autonomic nervous system signals for the accurate recognition of inner human emotions using limited hardware resources. Existing studies have conducted emotion recognition and classification based on two-class criteria. However, in this study, the results for emotion classification were derived using additional class criteria in the arousal and valence domains. This study confirms that although self-assessment labeling is appropriate for emotion recognition and classification between 10 and 60 s, short unit bio-signal emotion recognition and classification between 1 and 5 s can provide diverse learning data for observer annotation labeling settings.

VII. FUTURE WORKS

In future research, we intend to develop a framework for mounting on vehicles and human-robot interaction robots and develop models that can be used with human expressions and

voices. In addition, we plan to study the development of a system that can check whether the vehicle can be controlled through the driver's emotion recognition based on the developed lightweight model, and whether the external and internal emotions match.

REFERENCES

[1] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.
[2] C. M. Jones and T. Troen, "Biometric valence and arousal recognition," in *Proc. 19th Australas. Conf. Comput.-Hum. Interact., Entertaining User Interfaces*, 2007, pp. 191–194.
[3] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2018.
[4] G. F. Wilson and C. A. Russell, "Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 49, no. 6, pp. 1005–1018, Dec. 2007.
[5] D. Novak, M. Mihelj, J. Ziherl, A. Olensek, and M. Munih, "Psychophysiological measurements in a biocooperative feedback loop for upper extremity rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 4, pp. 400–410, Aug. 2011.
[6] W. Wang, J. L. Collinger, A. D. Degenhart, E. C. Tyler-Kabara, A. B. Schwartz, D. W. Moran, D. J. Weber, B. Wodlinger, R. K. Vinjamuri, R. C. Ashmore, J. W. Kelly, and M. L. Boninger, "An electrocorticographic brain interface in an individual with tetraplegia," *PLoS ONE*, vol. 8, no. 2, Feb. 2013, Art. no. e55344.
[7] J. Kilby, K. Prasad, and G. Mawston, "Multi-channel surface electromyography electrodes: A review," *IEEE Sensors J.*, vol. 16, no. 14, pp. 5510–5519, Jul. 2016.
[8] D. Mantini, M. G. Perrucci, C. Del Gratta, G. L. Romani, and M. Corbetta, "Electrophysiological signatures of resting state networks in the human brain," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 32, pp. 13170–13175, Aug. 2007.
[9] A. Topic and M. Russo, "Emotion recognition based on EEG feature maps through deep learning network," *Eng. Sci. Technol., Int. J.*, vol. 24, no. 6, pp. 1442–1454, Dec. 2021.
[10] M. Golgowski and S. Osowski, "Anomaly detection in ECG using wavelet transformation," in *Proc. IEEE 21st Int. Conf. Comput. Problems Electr. Eng. (CPEE)*, Sep. 2020, pp. 1–4.
[11] C. Jing, G. Liu, and M. Hao, "The research on emotion recognition from ECG signal," in *Proc. Int. Conf. Inf. Technol. Comput. Sci.*, vol. 1, Jul. 2009, pp. 497–500.
[12] M. Puurtinen, J. Viik, and J. Hyttinen, "Best electrode locations for a small bipolar ECG device: Signal strength analysis of clinical data," *Ann. Biomed. Eng.*, vol. 37, no. 2, pp. 331–336, Feb. 2009.
[13] R. Rakshit, V. R. Reddy, and P. Deshpande, "Emotion detection and recognition using HRV features derived from photoplethysmogram signals," in *Proc. 2nd Workshop Emotion Represent. Model. Companion Syst.*, Nov. 2016, pp. 1–6.
[14] M. S. Lee, Y. K. Lee, D. S. Pae, M. T. Lim, D. W. Kim, and T. K. Kang, "Fast emotion recognition based on single pulse PPG signal with convolutional neural network," *Appl. Sci.*, vol. 9, no. 16, p. 3355, Aug. 2019.
[15] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.
[16] N. Ganapathy, Y. R. Veeranki, and R. Swaminathan, "Convolutional neural network based emotion classification using electrodermal activity signals and time-frequency features," *Exp. Syst. Appl.*, vol. 159, Nov. 2020, Art. no. 113571.
[17] I. Y. Susanto, T.-Y. Pan, C.-W. Chen, M.-C. Hu, and W.-H. Cheng, "Emotion recognition from galvanic skin response signal based on deep hybrid neural networks," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 341–345.
[18] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.
[19] I. M. Revina and W. R. S. Emmanuel, "A survey on human face expression recognition techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 6, pp. 619–628, 2021.
[20] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 1998.
[21] P. Babajee, G. Suddul, S. Armoogum, and R. Foogooa, "Identifying human emotions from facial expressions with deep learning," in *Proc. Zooming Innov. Consum. Technol. Conf. (ZINC)*, May 2020, pp. 36–39.
[22] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Comput. Speech Lang.*, vol. 25, no. 3, pp. 556–570, Jul. 2011.
[23] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1760–1774, Nov. 2009.
[24] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2375–2385, Dec. 2017.
[25] M. Khan and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019.
[26] G. Hui, N. Jiang, and D. Shao, "Research on multi-modal emotion recognition based on speech, EEG and ECG signals," in *Proc. Int. Conf. Robot. Rehabil. Intell.* Singapore: Springer, 2020, pp. 272–288.
[27] C.-J. Yang, N. Fahier, W.-C. Li, and W.-C. Fang, "A convolution neural network based emotion recognition system using multimodal physiological signals," in *Proc. IEEE Int. Conf. Consum. Electron. Taiwan (ICCE-Taiwan)*, Sep. 2020, pp. 1–2.
[28] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 196–203, May 2018.
[29] M. Lee, Y. K. Lee, M.-T. Lim, and T.-K. Kang, "Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features," *Appl. Sci.*, vol. 10, no. 10, p. 3501, May 2020.
[30] J. L. McClelland and D. E. Rumelhart, *Parallel Distributed Processing*, vol. 2. Cambridge, MA, USA: MIT Press, 1986.
[31] E. Buber and B. Diri, "Performance analysis and CPU vs GPU comparison for deep learning," in *Proc. 6th Int. Conf. Control Eng. Inf. Technol. (CEIT)*, Oct. 2018, pp. 1–6.
[32] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.
[33] T. K. K. Ho and J. Gwak, "Utilizing knowledge distillation in deep learning for classification of chest X-ray abnormalities," *IEEE Access*, vol. 8, pp. 160749–160761, 2020.
[34] J. Panksepp, "Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist," *Perspect. Psychol. Sci.*, vol. 2, no. 3, pp. 281–296, 2007.
[35] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell, "On the importance of both dimensional and discrete models of emotion," *Behav. Sci.*, vol. 7, no. 4, p. 66, Sep. 2017.
[36] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005.
[37] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.
[38] 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. 2011.
[39] 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. 2011.
[40] I. Siegert, R. Bock, B. Vlasenko, D. Philippou-Hubner, and A. Wendemuth, "Appropriate emotional labelling of non-acted speech using basic emotions, Geneva emotion wheel and self assessment manikins," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–6.
[41] L. Fischer, B. Dieter, and B. Frank, "Zur Messung von Emotionen in der angewandten Forschung," *Beiträge zur Wirtschaftspsychologie*, 2002.
[42] J.-H. Maeng, D.-H. Kang, and D.-H. Kim, "Deep learning method for selecting effective models and feature groups in emotion recognition using an Asian multimodal database," *Electronics*, vol. 9, no. 12, p. 1988, Nov. 2020.
[43] I. W. Selesnick and C. S. Burrus, "Generalized digital Butterworth filter design," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1688–1694, Jun. 1998.

[44] F. H. Lesh, "Multi-dimensional least-squares polynomial curve fitting," *Commun. ACM*, vol. 2, no. 9, pp. 29–30, Sep. 1959.
[45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[46] W. Hwang, D. Kang, and D. Kim, "Brain lateralisation feature extraction and ant colony optimisation-bidirectional LSTM network model for emotion recognition," *IET Signal Process.*, vol. 16, no. 1, pp. 45–61, Feb. 2022.
[47] J. Wu, H. Liang, C. Ding, X. Huang, J. Huang, and Q. Peng, "Improving the accuracy in classification of blood pressure from photoplethysmography using continuous wavelet transform and deep learning," *Int. J. Hypertension*, vol. 2021, pp. 1–9, Aug. 2021.
[48] A. Mert and A. Akan, "Emotion recognition from EEG signals by using multivariate empirical mode decomposition," *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 81–89, Feb. 2018.
[49] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, no. 9, p. 2212, May 2019.
[50] X. Xing, Z. Li, T. Xu, L. Shu, B. Hu, and X. Xu, "SAE+LSTM: A new framework for emotion recognition from multi-channel EEG," *Frontiers Neurorobotics*, vol. 13, p. 37, Jun. 2019.
[51] H. J. Yoon and S. Y. Chung, "EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2230–2237, 2013.
[52] Z. Wang, S. Hu, and H. Song, "Channel selection method for EEG emotion recognition using normalized mutual information," *IEEE Access*, vol. 7, pp. 143303–143311, 2019.
[53] A. Raheel, M. Majid, and S. M. Anwar, "DEAR-MULSEMEDIA: Dataset for emotion analysis and recognition in response to multiple sensorial media," *Inf. Fusion*, vol. 65, pp. 37–49, Jan. 2021.

DONG-HYUN KANG received the bachelor's degree in electronic control engineering from Daegu University, South Korea, in 2020. He is currently pursuing the master's degree with the Intelligent Embedded Systems Laboratory, Inha University, South Korea. He is a Research Assistant with the Intelligent Embedded Systems Laboratory, Inha University. His current research interests include human–robot interactions related to signal processing, machine learning, and deep learning. E-mail: kawow123@gmail.com.

DEOK-HWAN KIM (Member, IEEE) received the Ph.D. degree in computer science from the Korean Advanced Institute of Science and Technology (KAIST), in 2003. He has been a Professor with Inha University, South Korea, since 2006. He has authored and/or presented more than 100 publications in major journals and international conferences. His current research interests include storage systems, human–robot interactions, embedded and real time systems, and biomedical systems. E-mail: deokhwan@inha.ac.kr.