Learning Cross-Modal Factors from Multimodal Physiological Signals for Emotion Recognition

Yuichi Ishikawa[1][0000−0003−3000−9113], Nao Kobayashi[2], Yasushi Naruse[3], Yugo Nakamura[1][0000−0002−8834−5323], Shigemi Ishida[4][0000−0003−0166−3984], Tsunenori Mine[1][0000−0002−7462−8074], and Yutaka Arakawa[1][0000−0002−7156−9160]

[1] Kyushu University, Fukuoka, Japan `ishikawa_yuichi@yahoo.co.jp`
[2] KDDI Research Inc., Saitama Japan
[3] Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, Hyougo, Japan
[4] Future University Hakodate, Hokkaido, Japan

Abstract. Understanding user emotion is essential for Human-AI Interaction (HAI). Thus far, many approaches have been studied to recognize emotion from signals of various physiological modalities such as cardiac activity and skin conductance. However, little attention has been paid to the fact that physiological signals are influenced by and reflect various factors that have little or no association with emotion. While emotion is a cross-modal factor that triggers responses across multiple physiological modalities, features used in existing approaches also reflect modality-specific factors that affect only a single modality and have little association with emotion. To address this, we propose an approach to extract features that exclusively reflect cross-modal factors from multimodal physiological signals. Our approach introduces a multilayer RNN with two types of layers: multiple *Modality-Specific Layers* (*MSL*s) for modeling physiological activity in individual modalities and a single *Cross-Modal Layer* (*CML*) for modeling the process by which emotion affects physiological activity. By having all MSLs update their hidden states using the CML hidden states, our RNN causes the CML to learn cross-modal factors. Using real physiological signals, we confirmed that the features extracted by our RNN reflected emotions to a significantly greater extent than the features of existing approaches.

Keywords: EEG · ECG · GSR · LSTM · Multilayer RNN.

1 Introduction

Understanding user emotions is extremely important for various human-AI interaction (HAI) scenarios including goal and non-goal oriented dialogue [6,8], user-adapted content creation [1], and content recommendation [2]. While most researchers collect ground truth of emotions by explicitly asking users what their emotions are, it is impractical to do so in real-world scenarios because doing so

interferes with users and degrades the user experience. Therefore, there has been a great demand for recognizing user emotions from data that users generate.

Among various types of user-generated data, we focus on users' physiological signals such as electroencephalogram (EEG), electrocardiogram (ECG), and galvanic skin response (GSR). Using wearable devices (e.g., watches, earphones), these signals can be collected in a less constrained context compared to other types of data such as texts, vocal tone, and facial expressions, which are available only when users write or say something or stay in front of a camera. In addition, unlike these data, physiological signals provide robust signs of emotion even when users exhibit their social masks to hide their true emotions [3].

Thus far, researchers have studied many approaches to recognize emotion from physiological signals and have confirmed their significant utility for emotion recognition [6]. However, little attention has been paid to the fact that physiological signals are influenced not only by emotion but also by various factors that have little or no association with emotion. Among them are factors that influence only a single physiological modality, i.e., a modality-specific factors. For example, heart muscle strength influences ECG signals, but has little influence on modalities other than cardiac activity such as brain activity and skin conductance. In contrast, emotion is a cross-modal factor, which triggers responses across multiple physiological modalities, e.g., anger increases heart rate and skin conductance level. Others are long-term factors such as body size and gender. These factors also influence physiological activity, but they are very different from emotion in a sense that they change very slowly or do not change, whereas emotion changes over short periods of time, i.e., a short-term factor.

As such, while emotion is a cross-modal and short-term factor, physiological signals are also influenced by and reflect factors that are modality-specific and/or long-term. Although they have little utility for emotion recognition, existing approaches extract and use features without distinguishing these factors, instead mixing them into the features. We posit this has degraded emotion recognition.

In light of the above, we propose an approach to extract features that exclusively reflect cross-modal and short-term factors. To achieve this, our approach distinguishes factors reflected in physiological signals along two axes: long- or short-term and modality-specific or cross-modal, and learns four types of factors that are distinct from each other. By adopting RNN, our approach separately models long- and short-term factors.

What is novel is that to model modality-specific and cross-modal factors, we introduce a multilayer RNN that consists of two types of layers: multiple *Modality-Specific Layers* (*MSL*s) that model physiological activity in individual modalities; and a single *Cross-Modal Layer* (*CML*) that learns cross-modal factors, among which is emotion. Our RNN takes sequences of multimodal physiological signals as input (e.g., ECG and GSR signals). Each MSL takes physiological signals of its corresponding modality (e.g., MSL1 takes ECG signals, MSL2 takes GSR signals) and reflects physiological states in its hidden state. When updating the hidden state, the MSL uses not only its own hidden state but also the CML's hidden state. Since this is done in all the MSLs, it makes

the CML's hidden state affect physiological state in all the modalities the MSLs correspond to. In effect, therefore, this enables the CML to learn factors that affect physiological activities across multiple modalities, i.e., cross-modal factors.

To evaluate our approach, we recruited participants and measured their EEG, ECG, and GSR signals while presenting them with musical pieces and movie clips (i.e., stimuli). We trained our RNN by these signals and, using the CML's hidden states, evaluated how accurately we could recognize emotions that the participants reported after each stimulus.

Our main contributions are as follows. 1) We propose a multilayer RNN that separates the RNN layer to learn factors that affect physiological activities across multiple modalities from the other layers designated to model modality-specific physiological activities. This enables our approach to extract features that exclusively reflect a cross-modal nature of emotion, which existing research has not focused on. 2) Using real physiological data, we demonstrate our RNN extracts features that reflect emotion to a greater extent than existing approaches.

2 Related Work

Similar to our approach, many existing approaches recognize emotion from multimodal physiological signals. Subramanian et al. [11] and Miranda et al. [7] used ECG, GSR, and EEG signals. Using feature extraction techniques that are widely used for each modality, they extracted features from each modality (*physiological features*; e.g., standard deviation of heartbeat intervals from ECG signals, mean skin conductance level from GSR signals). They then concatenated these physiological features and fed them into a classifier (i.e., early fusion). However, modality-specific factors reflected in the physiological features could not be removed by simple concatenation, thus limiting recognition accuracy. In addition, short- and long-term factors were not distinguished in the features. While they also tested late fusion, in which they combined recognition results in individual modalities to derive final results, the same issues remained because they used the same physiological features as in the early fusion, whose modality-specific factors hindered emotion recognition in each modality.

There are also multimodal approaches that adopt deep learning techniques. However, they have the same issues. Liu et al. [5] and Yin et al. [12] used deep autoencoders to learn shared representations of physiological features of multiple modalities (e.g., EEG and Electrooculogram) and recognized emotions by feeding the shared representations into classifiers. They trained the autoencoders so that the physiological features of each modality could be reproduced from the shared representations. This made the shared representations reflect not only cross-modal factors but also modality-specific factors. In addition, the use of the autoencoders did not help to distinguish between short- and long-term factors.

On the other hand, the approach proposed by Li et al. [4] can extract features that exclusively reflect short-term factors. Using the dataset built in [7], they fed time-series sequences of physiological features into LSTM, whose hidden states were then fed into an attention network. These steps enabled them to focus on
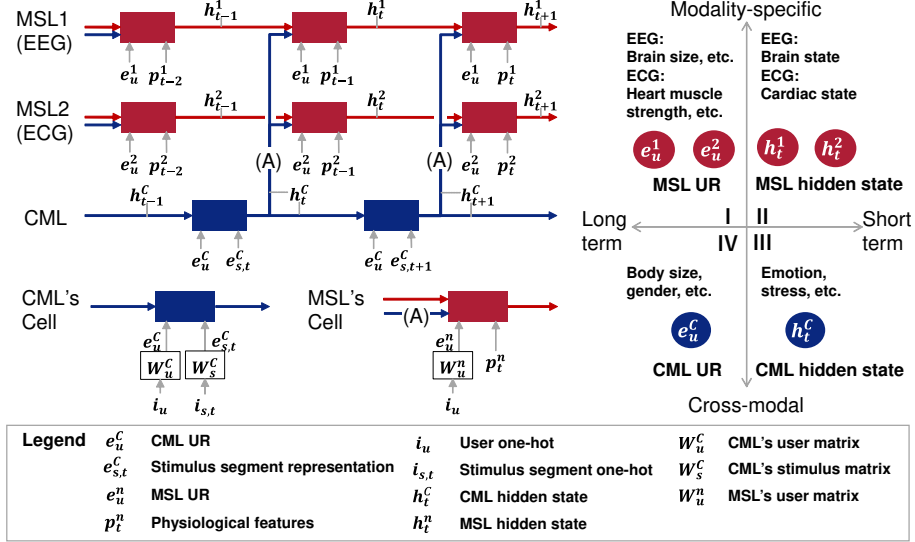
Fig. 1. An example of our multi-layered LSTM.

emotionally salient parts of the sequences, from which they extracted the hidden states and fed them into a multilayer perceptron (MLP) to recognize emotions. However, they performed these steps in each physiological modality and derived final results by combining the results of individual modalities (i.e., late fusion). Therefore, as in [11, 7], emotion recognition in individual modalities was hindered by modality-specific factors, which also degraded the final recognition results.

3   Proposed Approach

In contrast to the existing approaches, our RNN explicitly distinguishes the four types of factors that influence physiological activity. Fig. 1 exemplifies our RNN (left) and shows how the four types, I−IV, are mapped to its variables (right). Modality-specific factors, I and II, are modeled by the MSLs. Each MSL corresponds to a single modality, e.g., MSL1 to EEG, MSL2 to ECG. It takes sequences of 1) physiological features of the corresponding modality, which are extracted in the same way as existing approaches (e.g., [7, 11]), and 2) one-hot vectors of user ID, by which a user representation (UR) is retrieved from the user matrix. Since the physiological features fed to the MSL are limited to the corresponding modality, its URs and hidden states reflect factors specific to this modality (I and II). In addition, while the hidden states are updated sequentially, the user matrix (set of URs) stays the same. This causes the MSL URs to reflect long-term factors (I) and its hidden states to reflect short-term factors (II).

On the other hand, cross-modal factors, III and IV, are modeled by the CML. As shown by link (A) in the figure, the CML sends its hidden states to the MSLs.

Table 1. List of notations

Notation	Meaning	Subscript	Superscript
N_u, N_s	# of users/SS	u: # of users, s: # of stimulus segments (SS)	
$\boldsymbol{i_u} \in \mathbb{R}^{N_u}, \boldsymbol{i_{s,t}} \in \mathbb{R}^{N_s}$	One hot vector of …	u: user ID, s, t: SS ID of t-th action	-
$d_{ue}^C, d_{ue}^n, d_{se}^C, d_h^C, d_h^n, d_p^n$	# of dimensions of …	ue: UR, se: SS Representation (SR), h: hidden state, p: physiological features	
$\boldsymbol{W_u^C} \in \mathbb{R}^{d_{ue}^C \times N_u}, \; \boldsymbol{W_u^n} \in \mathbb{R}^{d_{ue}^n \times N_u}$	User matrix	u: user matrix	
$\boldsymbol{W_s^C} \in \mathbb{R}^{d_{se}^C \times N_s}$	SS matrix	s: SS matrix	C: CML n: MSL n
$\boldsymbol{e_u^C} \in \mathbb{R}^{d_{ue}^C}, \; \boldsymbol{e_u^n} \in \mathbb{R}^{d_{ue}^n}, \; \boldsymbol{e_{s,t}^C} \in \mathbb{R}^{d_{se}^C}$	UR/SR	u: UR, s, t: SR of t-th action	
$\boldsymbol{h_t^C} \in \mathbb{R}^{d_h^C}, \boldsymbol{h_t^n} \in \mathbb{R}^{d_h^n}$	Hidden state	t: hidden state after t-th action	
$\boldsymbol{p_t^n} \in \mathbb{R}^{d_p^n}$	Physiological features	t: physiological features in t-th action	

The MSL cell uses these CML hidden states together with its input (the physiological features and URs) and its previous hidden states to update its hidden states. Because updated MSL hidden states are used to predict the physiological features at the next timeslot, it can be regarded as representing physiological state. Updating such MSL hidden states using the CML hidden states means that the CML hidden states affect physiological activity of individual modalities. Because all the MSLs update their hidden states in this way, the CML learns factors that affect physiological activity across multiple modalities, i.e., cross-modal factors (III and IV). As in the MSL, the CML also reflect short-term factors (III) in its hidden states and long-term factors (IV) in its URs, but the difference being they are cross-modal.

In addition to modeling I∼IV, our RNN also models the process by which individual physiological differences moderate the relationship between emotion and physiological activity. For example, users with different heart muscle strength would have different ECG signals even when their emotions are the same. Our RNN models such moderating effect of individual differences by updating the MSL hidden states (reflecting physiological state) using both the CML hidden states (emotion) and the MSL URs (individual differences, e.g., heart muscle strength). This also differentiates our RNN from the existing approaches discussed in Section2, all of which do not consider this moderating effect.

The next section describes in detail the hidden state updating in our RNN and its training process. See Table 1 for the notations and their descriptions.

3.1 Updating the Hidden States

Input to the CML and MSL are formatted as

$$\text{CML: } data_a^C = [x_{a,1}^C, x_{a,2}^C, ..., x_{a,t}^C, ..., x_{a,T}^C] \text{ and} \tag{1}$$

$$\text{MSL } n: data_a^n = [x_{a,1}^n, x_{a,2}^n, ..., x_{a,t}^n, ..., x_{a,T}^n], \tag{2}$$

where $x_{a,t}^C = (\boldsymbol{i_u}, \boldsymbol{i_{s,t}})$ denotes user a's t-th action (e.g., viewed t-th segment of a movie clip M); and $x_{a,t}^n = (\boldsymbol{i_u}, \boldsymbol{p_t^n})$ denotes his physiological features extracted from signals during t-th action. Once $x_{a,t}^C$ is input to the CML, it first retrieves

a UR and stimulus segment (SS) representation (SR) from the user and SS matrices, i.e., $e_u^C = W_u^C i_u$ and $e_{s,t}^C = W_s^C i_{s,t}$, respectively. It then updates its hidden state h_t^C as follows:

$$h_t^C = f^C(\, h_{t-1}^C,\ e_u^C,\ e_{s,t}^C \,), \tag{3}$$

where f^C is a function implemented by LSTM. See the supplementary material at https://osf.io/mj3nr/ for detail.

After updating the hidden state, the CML sends it to all MSLs via link (A), which is done every time the CML updates its hidden state. When the MSL receives h_t^C, it retrieves a UR from its user matrix ($e_u^n = W_u^n i_u$) and uses them together with input physiological features (p_{t-1}^n) to update its hidden state h_t^n as follows:

$$h_t^n = f^n(\, h_{t-1}^n,\ h_t^C,\ e_u^n,\ p_{t-1}^n \,), \tag{4}$$

where f^n is a function implemented by LSTM (see the supplementary material).

3.2   Model Training

When $data_a^n$ is fed, each MSL predicts physiological features in each timeslot, e.g., if the input is $data_a^n = [x_{a,t}^n, x_{a,t+1}^n, ..., x_{a,T-1}^n]$, the output is $[\hat{p}_{t+1}^n, \hat{p}_{t+2}^n, ..., \hat{p}_T^n]$. The predicted physiological features are compared with the actual features to calculate the loss that is used to learn the parameters of the MSL and CML cells and the user and SS matrices (W_u^n, W_u^C, and W_s^C). Fig. 2 shows how the prediction and loss calculation are performed. The MSL predicts physiological features using its hidden state as follows: $\hat{p}_{t+1}^n = f_{MLP}^n(h_{t+1}^n)$, where f_{MLP}^n is an

Fig. 2. Loss calculation

MLP with ReLu activation. Then, the MSL calculates the residual sum of squares between actual and predicted physiological features as the loss.
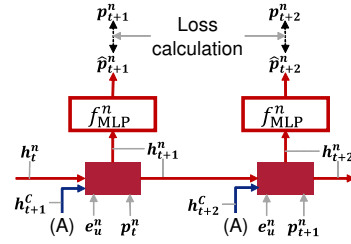
4   Experiment

We built datasets and evaluated the extent to which the CML hidden states reflect emotion. We performed the following three steps: (1) Feature extraction - from the physiological features stored in our datasets, we extracted another set of features for emotion recognition (*emotion features*). In our RNN, the CML hidden states were used as the emotion features; (2) feature selection - we then performed LASSO regression to select the emotion features; and (3) linear regression - using the selected emotion features, we built models to predict emotions and evaluated their model fit and prediction accuracy.

We performed (1)-(3) for our RNN and three approaches to compare. The first approach, which was implemented following [11, 7], did not distinguish between

the four types of factors at all when extracting the emotion features (hereafter "baseline"). The second one distinguished between long and short-term factors but not between cross-modal and modality-specific factors as in [4]; and the third one distinguished the four types of factors, but did not model the moderating effect of individual physiological differences. The last two were implemented by removing key features from our RNN (will be explained in 4.4 Ablation Study).

We built two different datasets and performed (1)-(3) for each dataset. In addition, because combinations of physiological modalities available in real-world scenario would be different depending on the devices users wear, we performed (1)-(3) for all possible modality combinations available in our datasets. That is, A) EEG+ECG+GSR, B) EEG+ECG, C) EEG+GSR, and D) ECG+GSR.

4.1 Dataset

Due to page limitations, only a brief summary of the datasets is described below. See the supplementary material (https://osf.io/mj3nr/) for detail. Although several datasets are publicly available today (e.g., [7, 11]), we built and used our own datasets. One reason is because the contacts of these datasets did not respond to our requests. The other is because they used only videos as stimuli when collecting physiological signals. Because music is another popular type of stimulus that would be played more often especially while working, studying, etc., we considered evaluation should be done for both music and video.

We built Music and Movie datasets by conducting data collection experiments, in which 54 and 52 (out of 54) subjects participated, respectively. They were presented with multiple stimuli, each of which was 60 seconds long, while their EEG, ECG, and GSR signals were measured. In total, 2,336 and 2,119 trials were performed for the music and movie datasets, respectively (one trial denotes one subject listening to/viewing one stimulus). After listening to/viewing each stimulus, they reported emotions according to the six dimensions whose scores ranged 0-15, (a) sad-happy, (b) nervous-relaxed, (c) fear-relieved, (d) lethargic-excited, (e) depressed-delighted, and (f) angry-serene. Although Russel's circumplex model [10] has been widely used to determine emotion, we did not use it because it is not easy for lay participants to report "arousal" and "valence" defined in the model. We selected the six dimensions so that the participants can easily report their emotions and the dimensions cover the Russel's circumplex as much as possible.

After collecting the physiological signals, we extracted the physiological features from the raw signals by feature extraction techniques that are widely used for each modality as in [11, 7]. We extracted two types of features: window and stimulus features, which are summarized in Table 2. For the window features, we applied sliding window to the raw signals measured during one stimulus and extracted features from each window. We set the window size to ten seconds and used two different slide sizes, three and five seconds. That is, we had 17 and 11 windows for each stimulus, respectively. The stimulus features were extracted from entire signals measured during a stimulus. We stored the physiological features in the datasets after performing z-standardization for each dimension.

Table 2. Physiological features. Bold numbers denote dimension.

Modality	Window features	Stimulus features
EEG	**29**: Power within individual frequency bands ($\delta, \theta, \alpha, \beta$) and composite score ($S$) at each electrode; power asymmetricity of individual frequency bands for one pair of electrodes	**116**: Statistics (mean, min, max, and standard deviation) of 29 window features
ECG	**7**: SDNN, pNN50, RMSSD, and HR statistics (mean, min, max, and standard deviation)	**10**: SDNN, pNN50, RMSSD, and HR statistics (mean, min, max, and standard deviation) and power within HF, LF and power ratio (HF/LF)
GSR	**8**: SCL related features (slope, mean) and SCR related features (peak, skew, kurtosis, max, min, mean)	**11**: SCL related features (slope and mean) and SCR related features (the number of peaks, max and min of peaks, skew, kurtosis, max, min, and mean)

SDNN: Standard Deviation of Normal-to-Normal intervals, pNN50: The proportion of NN50 divided by the total number of NN intervals, RMSSD: Root Mean Square of the Successive Differences, HR: Hear Rate, HF/LF: High/Low Frequency, SCL: Skin Conductance Level, SCR: Skin Conductance Response

4.2   Step1 - Extraction of Emotion Features

Proposed approach Of the two types of the physiological features, we used the window features as the input to our RNN. That is, an input sequence to the CML and MSL n ($data_a^C$ and $data_a^n$) corresponds to a trial. An element of $data_a^C$ (i.e., $\boldsymbol{i_{s,t}}$) and $data_a^n$ (i.e., $\boldsymbol{p_t^n}$) correspond to t-th window of a stimulus and the physiological features extracted from the raw signals in t-th window of the stimulus, respectively. The total number of input sequences was equal to the number of trials, out of which 80% were used for training and 20% for validation. We did not use the stimulus features because $\boldsymbol{i_{s,t}}$ corresponds to a stimulus if we do so and thus the number of input sequences, which is equal to the number of participants, was too small for training our RNN.

The hyper parameters were as follows: slide size of the sliding window = [3sec, 5sec], learning rate = $[5 \times 10^{-4}, 1 \times 10^{-3}]$, dimension of hidden layers of the MSL's MLP (i.e., f_{MLP}^n) = $[(16, 8), (32, 16)]$ (from input to output layer), batch size = $[16, 32]$, and dimension of UR, SR, and hidden state of the CML and MSL = $[8, 16]$. For all possible combinations of the hyper parameters, we conducted training and validation for 100 epochs and extracted the CML hidden states of the validation samples when we observed the minimum validation loss. We repeated this changing training and validation samples so that we could obtain the CML hidden states for all trials. Because the prediction target is emotion after each trial, we used the last CML hidden state of each trial as the emotion features, i.e., if the last element of $data_a^C$ was $\boldsymbol{i_{s,T}}$, we used $\boldsymbol{h_T^C}$.

Baseline Following [7, 11], we first concatenated the physiological features across modalities. This was done for both the stimulus and window features. For example, if the modality combination was A) EEG+ECG+GSR, we built 137 ($116 + 10 + 11$) dimension features from the stimulus features and 748 ($(29 + 7 + 8) \times 17$) dimension features from the window features (if there are 17 windows in a stimulus) for each trial. We then reduced their dimension by performing PCA and extracted top n features in terms of their contribution ratio so that their cumulative contribution ratio is maximum below a threshold. We used these features

as the emotion features. We set three different thresholds, 0.85, 0.90, and 0.95. In the following, S and W denote the emotion features extracted from the stimulus and window features, respectively. Because Miranda et al. [7] reported that recognition by unimodal features outperformed multimodal features, we also extracted S and W for each physiological modality.

4.3   Step 2 and 3 - Feature Selection and Linear Regression

After extracting the emotion features, we performed feature selection and linear regression. These were done for each of the six emotion dimensions.

We performed the feature selection because dimension of the emotion features of the baseline was large relative to the sample size (i.e., the number of trials). For fair comparison, this was done for both the baseline and our approach. We first finetuned the LASSO parameter λ, which controls the strength of the imposed regularization based on the number of selected features. Over a set of λ values, we sought the value that output the most accurate prediction (i.e., minimum mean squared error between the actual and predicted emotion scores) performing five-fold cross-validation multiple times. Second, we conducted the LASSO regression again using the value of λ determined in the previous step and selected features for which the regression coefficients were not zero.

After the feature selection, we performed two types of linear regression. One is model fit evaluation using all samples. The other is prediction evaluation by performing five-fold cross validation.

4.4   Ablation Study

To determine the effectiveness of the key features of our RNN, we evaluated its variants without the key features, which are shown in Fig. 3. One is a single layer RNN (AB1) that takes concatenated multimodal physiological features ($p_{s,t}$ in the figure) as input and the other is a multilayer RNN without the MSL URs (AB2). We extracted their hidden states (the CML hidden states in AB2) as the emotion features and evaluated them in the same way as our RNN.

Similar to [4], while AB1 can extract emotion features that exclu-

Fig. 3. Variants for ablation study.

sively reflect short-term factors, it cannot distinguish between modality-specific and cross-modal factors, mixing both into the features. While the emotion features of AB2 would exclusively reflect short-term and cross-modal factors, the MSL in AB2 cannot model the moderating effect of individual physiological differences due to lack of the MSL URs.
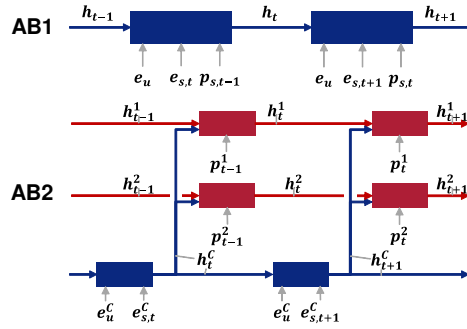
Table 3. Emotion recognition results. A−D represent the modality combinations (ref. section 4). Shaded cells denote results inferior to our RNN (ours) in the same columns. Cells with hatched lines indicate that LASSO selected no emotion feature. Black cells denote the best results for the emotion dimensions. For the baseline, BL and U-BL, the table shows the best results of the three PCA thresholds. U-BL uses a single physiological modality in BL and the table shows the result of the best modality in a corresponding combination (e.g., U-BL in column B show better of EEG and ECG).

Results of Music dataset (n=2,336)

Metric	Model	(a) sad-happy				(b) nervous-relaxed				(c) fear-relieved				(d) lethargic-excited				(e) depressed-delighted				(f) angry-serene			
		A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
AIC	BL(S)	11231	11235	11235	11234	11495	11501	11495	11503	11145	11145	11145	11162	11553	11553	11554	11544	12205	12205	12205	/	10416	10426	10408	10430
AIC	BL(W)	11235	/	/	11234	/	11503	11503	11505	11150	11150	11151	/	11521	11547	11533	11534	12191	12206	12181	12201	10414	10420	/	10426
AIC	U-BL(S)	11237	11237	11237	11237	11499	11499	11499	11503	11145	11145	11145	11160	11551	11551	11554	11551	12205	12205	12205	12211	10415	10415	10415	10428
AIC	U-BL(W)	11234	11236	11234	11234	11503	11503	11503	11504	11151	11151	11151	11154	11543	11549	11543	11543	12198	12201	12198	12198	10416	10416	10416	10426
AIC	AB1	11197	11128	11160	11206	11486	11439	11474	11486	11119	11113	11121	11142	11508	11493	11478	11499	12170	12044	12138	12182	10402	10380	10394	10429
AIC	AB2	11130	11107	11166	11133	11431	11426	11454	11463	11097	11076	11088	11129	11421	11390	11495	11398	12086	12084	12116	12097	10350	10344	10350	10403
AIC	Ours	**11017**	11059	11078	11077	**11363**	11398	11381	11411	11068	11060	**11039**	11094	11346	**11343**	11437	11378	11989	**11961**	12050	12057	**10318**	10376	10337	10346
RMSE	BL(S)	2.232	2.232	2.235	2.233	2.422	2.430	2.423	2.434	2.159	2.159	2.159	2.149	2.392	2.391	2.392	2.401	2.925	2.926	2.926	/	1.856	1.865	1.851	1.860
RMSE	BL(W)	2.232	/	/	2.232	/	2.433	2.433	2.435	2.157	2.157	2.159	/	2.386	2.387	2.394	2.395	2.910	2.923	2.902	2.925	1.854	1.858	/	1.857
RMSE	U-BL(S)	2.231	2.233	2.231	2.231	2.428	2.428	2.428	2.434	2.154	2.154	2.159		2.392	2.392	2.392	2.397	2.923	2.923	2.923	2.928	1.859	1.860	1.859	1.859
RMSE	U-BL(W)	2.229	2.229	2.229	2.229	2.433	2.433	2.433	2.435	2.151	2.156	2.151	2.151	2.388	2.388	2.388	2.395	2.921	2.921	2.921	2.925	1.858	1.858	1.858	1.859
RMSE	AB1	2.218	2.158	2.199	2.217	2.410	2.371	2.396	2.414	2.159	2.150	2.155	2.160	2.389	2.357	2.364	2.376	2.889	2.778	2.850	2.895	1.847	1.841	1.838	1.858
RMSE	AB2	2.185	2.167	2.197	2.196	2.363	2.355	2.384	2.391	2.144	2.125	2.164	2.158	2.339	2.321	2.386	2.321	2.833	2.829	2.834	2.842	1.827	1.822	1.833	1.841
RMSE	Ours	**2.135**	2.136	2.161	2.158	**2.318**	2.342	2.330	2.367	2.135	2.131	**2.121**	2.146	**2.271**	2.286	2.331	2.312	2.752	**2.743**	2.797	2.815	**1.797**	1.808	1.806	1.828

Results of Movie dataset (n=2,119)

Metric	Model	(a) sad-happy				(b) nervous-relaxed				(c) fear-relieved				(d) lethargic-excited				(e) depressed-delighted				(f) angry-serene			
		A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
AIC	BL(S)	8782	8782	8782	/	/	/	9227	/	/	9192	9193	/	/	9807	9825	9821	9698	9699	9699	/	9251	9282	9283	9274
AIC	BL(W)	8781	8782	8781	8786	9228	/	9228	9229	9195	/	9194	9197	9812	9799	/	9814	9694	9693	9699	9690	9290	9290	9287	9284
AIC	U-BL(S)	8782	8782	8782	8790	9227	9227	9227	9235	9192	9192	9192	/	9801	9801	9829	9801	9689	9689	9699	9689	9268	9268	9283	9268
AIC	U-BL(W)	8780	8782	8780	8780	9214	9228	9214	9214	9171	9194	9171	9171	9795	9795	9823	9795	9686	9686	9695	9686	9264	9286	9264	9264
AIC	AB1	8778	8777	8763	8776	9182	9223	9214	9211	9133	9186	9135	9164	9764	9774	9821	9781	9673	9668	9695	9670	9197	9210	9225	9161
AIC	AB2	8776	8778	8740	8780	9224	9224	9206	9216	9191	9168	9135	9164	9761	9760	9722	9738	9656	9638	9625	9604	9205	9214	9225	9148
AIC	Ours	8753	8774	**8734**	8748	9194	9197	9203	**9125**	9134	9129	**9105**	9136	9757	**9646**	9757	9715	9651	**9505**	9590	9558	9144	**9119**	9126	9162
RMSE	BL(S)	1.438	1.438	1.438	/	/	/	1.604	/	/	1.607	1.607	/	/	1.800	1.804	1.812	1.736	1.736	1.736	/	1.759	1.772	1.773	1.760
RMSE	BL(W)	1.438	1.438	1.438	1.439	1.605	/	1.605	1.609	1.609	/	1.608	1.616	1.803	1.796	/	1.803	1.735	1.734	1.737	1.738	1.776	1.776	1.775	1.778
RMSE	U-BL(S)	1.438	1.438	1.438	1.444	1.604	1.604	1.604	1.611	1.607	1.607	1.607	/	1.797	1.797	1.813	1.797	1.736	1.736	1.736	1.736	1.759	1.759	1.773	1.759
RMSE	U-BL(W)	1.436	1.438	1.436	1.436	1.605	1.605	1.605	1.613	1.608	1.608	1.608	1.620	1.790	1.790	1.809	1.790	1.733	1.733	1.736	1.733	1.759	1.774	1.759	1.759
RMSE	AB1	1.432	1.432	1.424	1.435	1.601	1.604	1.607	1.617	1.617	1.617	1.615	1.626	1.778	1.794	1.808	1.781	1.739	1.741	1.740	1.732	1.722	1.744	1.743	1.727
RMSE	AB2	1.439	1.436	1.400	1.429	1.607	1.601	1.613	1.615	1.619	1.616	1.607	1.622	1.776	1.763	1.768	1.763	1.744	1.740	1.725	1.723	1.732	1.740	1.746	1.701
RMSE	Ours	1.426	1.427	**1.400**	1.421	1.603	1.594	1.609	**1.586**	1.606	1.602	**1.599**	1.614	1.754	**1.709**	1.764	1.777	1.726	**1.678**	1.723	1.711	1.704	**1.687**	1.691	1.718

5 Results and Discussion

Table 3 shows the results. Due to page limitations, the table shows only the Akaike Information Criterion (AIC; model fit metric; the lower the better) and the Root Mean Square Error (RMSE; prediction accuracy metric). See the supplementary material (https://osf.io/mj3nr/) for the results of other metrics. As the table shows, the regression models of our RNN (ours) outperformed the baseline models (BL, U-BL), AB1, and AB2 in most conditions not limited to specific stimulus types, emotion dimensions, or modality combinations.

Compared to the baseline models, which do not distinguish between the four types of factors at all, ours outperformed them in all conditions of both datasets with only one exception (RMSE of (b)-C in the Movie dataset). The differences are significant according to the relative likelihood (RL) that are calculated from their AICs; $RL = \exp((AIC(ours) - AIC(BL \text{ or } U\text{-}BL))/2)$, where $AIC(M)$ de-

notes the AIC of regression model M. In all conditions, the RLs between ours and the best baseline models are less than 0.05 (see the supplementary material), which means that the likelihood of the best baseline models being closer to the true model than ours is less than 0.05. This indicates that the features extracted by our RNN reflect emotions to a significantly greater extent than the baseline.

The same is true between our RNN and its variants, AB1 and AB2. Out of 24 conditions, ours outperformed them in 23 conditions in the Music dataset and 20 conditions in the Movie dataset for both AIC and RMSE. The RLs between ours and the better of AB1 and AB2 were less than 0.05 in all 23 conditions in the Music dataset and 15 out of 20 conditions in the Movie dataset. These results indicate that the following key features of our RNN, which were not implemented in AB1 and AB2, significantly contributed to causing its emotion features to reflect emotion. That is, the multilayer structure for distinguishing cross-modal and modality-specific factors and the MSL URs for modeling the moderating effect of individual physiological differences.

What is notable in our RNN is that using more modalities does not necessarily make the emotion features (i.e., the CML hidden states) reflect emotion more. As shown in the table, using all three modalities (i.e., A) performed best in only three out of 12 cases (six emotion dimensions × two datasets). This accords with the existing studies [7, 11]. For example, in [11], ECG+GSR outperformed EEG+ECG+GSR for recognizing arousal. The authors considered this would be because EEG did not reflect arousal as well as the other two modalities and would be noise for the recognition.

Although our RNN differs from them in the feature extraction, we consider this is also true for our approach. In our RNN, the CML learns latent common factors that affect all input physiological modalities. While this prevents the CML from learning modality-specific factors, it would be also possible that the CML fails to learn factors that are common to only a subset of input modalities and useful for emotion recognition but do not affect the remaining input modalities. For example, in the Music dataset, C) EEG+GSR outperformed A) EEG+ECG+GSR to recognize c) fear-relieved. We consider using ECG as input would have prevented the CML from learning factors that are common only to EEG and GSR and useful for recognizing this emotion dimension.

In light of the above, as in the existing approaches, it is necessary to compare possible modality combinations to identify the best combination in our approach. Since the best modality combinations are different between emotion dimensions and stimulus types (music and movie), the comparison of modality combinations should be done for each emotion dimension and stimulus type.

6   Conclusions, Limitations and Future Direction

In this paper, we proposed a multilayer RNN to extract features from multimodal physiological signals for emotion recognition. Using a multilayer structure, our RNN models the process by which emotion affects physiological activities across multiple modalities. This enables our RNN to extract features that are cross-

modal, which is one of the characteristics of emotion but has been overlooked in existing studies. The experiments conducted on EEG, ECG, and GSR signals showed that the features extracted by our RNN reflected the participants' emotions to a significantly greater extent than existing approaches.

One limitation is that our RNN only models unidirectional relationship between emotion and physiological activity, i.e., the former affects the latter. According to Roberts et al [9], perception of internal physiological state (known as interoception) would also affect emotion. Modeling this inverse relationship in our RNN would make the features reflect emotion more. This possibility should be explored. Another limitation is that we only examined physiological signals collected while the participants stayed still. In real-world scenarios, however, physiological signals would contain noise caused by body movements. Further studies are warranted to investigate how our RNN performs with such signals.

References

1. Aranha, R.V., Corrêa, C.G., Nunes, F.L.: Adapting software with affective computing: a systematic review. IEEE Transactions on Affective Computing **12**(4), 883–899 (2019)
2. Deng, J.J., Leung, C.H., Milani, A., Chen, L.: Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation. ACM Transactions on Interactive Intelligent Systems (TiiS) **5**(1), 1–36 (2015)
3. Kim, J., André, E.: Emotion recognition based on physiological changes in music listening. IEEE transactions on pattern analysis and machine intelligence **30**(12), 2067–2083 (2008)
4. Li, C., Bao, Z., Li, L., Zhao, Z.: Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. Information Processing & Management **57**(3), 102185 (2020)
5. Liu, W., Zheng, W.L., Lu, B.L.: Emotion recognition using multimodal deep learning. In: ICONIP 2016, Proceedings, Part II 23. pp. 521–529. Springer (2016)
6. Ma, Y., Nguyen, K.L., Xing, F.Z., Cambria, E.: A survey on empathetic dialogue systems. Information Fusion **64**, 50–70 (2020)
7. Miranda-Correa, J.A., Abadi, M.K., Sebe, N., Patras, I.: Amigos: A dataset for affect, personality and mood research on individuals and groups. IEEE Transactions on Affective Computing **12**(2), 479–493 (2018)
8. Peng, W., Hu, Y., Xing, L., Xie, Y., Sun, Y.: Do you know my emotion? emotion-aware strategy recognition towards a persuasive dialogue system. In: ECML PKDD 2022, Proceedings, Part II. pp. 724–739. Springer (2023)
9. Roberts, T.A., Pennebaker, J.W.: Gender differences in perceiving internal state: Toward a his-and-hers model of perceptual cue use. In: Advances in experimental social psychology, vol. 27, pp. 143–175. Elsevier (1995)
10. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980)
11. Subramanian, R., Wache, J., Abadi, M.K., Vieriu, R.L., Winkler, S., Sebe, N.: Ascertain: Emotion and personality recognition using commercial sensors. IEEE Transactions on Affective Computing **9**(2), 147–160 (2016)
12. Yin, Z., Zhao, M., Wang, Y., Yang, J., Zhang, J.: Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Computer methods and programs in biomedicine **140**, 93–110 (2017)