Self-assessed Emotion Classification from Acoustic and Physiological Features within Small-group Conversation

Woan-Shiuan Chien, Huang-Cheng Chou, Chi-Chun Lee
wschien@gapp.nthu.edu.tw,hc.chou@gapp.nthu.edu.tw,cclee@ee.nthu.edu.tw
Department of Electrical Engineering, National Tsing Hua University
MOST Joint Research Center for AI Technology and All Vista Healthcare
Taiwan

ABSTRACT

Individual (personalized) self-assessed emotion recognition has recently received more attention, such as Human-Centered Artificial Intelligence (AI). In most previous studies, researchers utilized the physiological changes and reactions in the body evoked by multimedia stimuli, e.g., video or music, to build a model for recognizing individuals' emotions. However, this elicitation approach is less impractical in the human-human interaction because the conversation is dynamic. In this paper, we firstly investigate the individual emotion recognition task under three-person small group conversations. While predicting personalized emotions from physiological signals is well-studied, few studies focus on emotion classification (e.g., happiness and sadness). Most prior works only focus on binary dimensional emotion recognition or regression, such as valence and arousal. Hence, we formulate the individual emotion recognition task into an individual-level emotion classification. In the proposed method, we consider the physiological changes in each individual's body and acoustic turn-taking dynamics during group conversations for predicting individual emotions. Meanwhile, we assume that the emotional states of humans might be affected by the expressive behaviors of other members during group conversations. Also, we hypothesize that people have a higher probability of feeling specific emotions under the related emotional atmosphere. Therefore, we design an ad-hoc technique by simply summing up the Self-assessed emotional annotations of all group members as the group emotional atmosphere (climate) to help the model predict individuals' emotions. We propose a Multi-modal Multi-label Emotion based on Transformer BLSTM at Group Emotional Atmosphere Network (MMETBGEAN) that explicitly considers individual changes and dynamic interaction via physiological and acoustic features during a group conversation integrates group emotional atmosphere information for recognizing individuals' multi-label emotions. We assess the proposed framework on our recently collected extensive Mandarin Chinese collective task group database, NTUBA. The results show that the method outperforms the existing approaches on multi-modal multi-label emotion classification on this database.

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8471-1/21/10…$15.00
https://doi.org/10.1145/3461615.3485411

CCS CONCEPTS

• **Computing methodologies** → **Temporal reasoning**; • **Information systems** → *Multimedia information systems*; • **Human-centered computing** → Social network analysis.

KEYWORDS

Multi-modal interaction, multi-label emotion classification, group emotion regression, group interaction

ACM Reference Format:
Woan-Shiuan Chien, Huang-Cheng Chou, Chi-Chun Lee. 2021. Self-assessed Emotion Classification from Acoustic and Physiological Features within Small-group Conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3461615.3485411

1 INTRODUCTION

Predicting emotions is a critical technology in human-centered applications and human-computer interactions. There is a wide range of essential cues to be modeled in emotion recognition, such as facial expression, speech, spoken language, gesture, and postures, and physiological signals [11]. Physiological signals dominated by the central nervous system (e.g., Electroencephalography (EEG)) and peripheral nervous system (e.g., Electrocardiography (ECG) and Photoplethysmography (PPG)) are generally separate of humans' will and not easily restrained [12, 60] compared to other signals that can be controlled willingly. Hence, physiological signals may supply more dependable cues for individuals' emotions than other expressive behavioral cues, such as visual and audio cues.

While several approaches have been proposed to predict individual self-assessed emotions from the physiological signals, most previous studies utilized the laboratory elicited/induced corpus [41], and this limits its potential for practical applications in the real world. More specifically, these databases are recorded and collected by triggering subjects with selected/intended emotional stimuli (e.g., music, videos, images, movies, to name a few), and this fixed and one-time stimulus scenario is not easy to be applied and adapted in the human interaction situation, like during group discussion, because of difficulty on modeling turn-taking dynamics. We can imagine that the stimulus (expressive behaviors of interlocutors/partners) could be changed over time in the group discussion. Hence, the prior studies are not easy to handle this situation. Besides, with consideration of privacy protection, it is advantageous to estimate emotions without observable facts. Therefore, in this paper, we use the dynamic turn-taking stimulus from speech (expressive behaviors) and the changes in the physiological signals from PPG

over time during the small group conversation to train the proposed framework. Then, we evaluate the proposed framework on a new large Mandarin Chinese collective task group database, whose participants spontaneously interact with each other in small groups, so the database is closer to the real-world scenario than other public databases (e.g., DEAP [24], MAHNOB-HCI [42], ASCERTAIN [43], and AMIGOS [30]) for training the emotion detection model. Furthermore, most previous studies on physiological-based emotion recognition [36, 40, 49] regard the emotion classification as a single label task. However, emotional feelings are naturally subjective, and people often interpret the same situation differently based on different emotional experiences. For example, when humans are asked to describe perceived emotion, they usually need more than one emotion category to show their emotion perception, which might be due to the ambiguous boundary between emotion categories [39] and cause the existence of multiple emotional labels. Therefore, in this paper, we formulate the emotion classification task as a multi-label classification task.

To handle the above two issues, dynamic stimulus and multiple self-assessed emotions, we introduce a new corpus that collects the entire conversations between many three-person groups. These subjects are asked to finish a shopping task with a limited budget together in 30 minutes, so they have to discuss it with each other actively. This database is very suitable for our research because the corpus carefully records audio recordings and physiological signals during group discussion and collects self-assessed emotion perception on each emotion category in the 7-point Likert scale. The database allows us to investigate and model dynamic stimulus and changes of physiological signals over time for predicting individual self-assessed multi-label emotion classification. Moreover, the study [3] said that the in-group mutual interactions can affect the group members' emotions, and the group emotional atmosphere is also related to the composition of individuals' emotions. In addition, Qiao-Tasserit et al. [37] showed that negative (comparing neutral) short videos raised participants' tendency to categorize unclear faces as fearful during some minutes. Instead, positive clips (comparing neutral) influenced categorization toward happiness only for those movies perceived as most absorbing. That is, the research [37] reveals that absorption of emotions significantly affects how people perceive facial expressions.

Therefore, inspired by the studies mentioned above, we hypothesize that people are more likely to feel specific emotions under the related group emotional atmosphere. We sum up the self-assessed emotion scores on the emotion category of group members as the group emotional atmosphere. We propose a novel graph-level Graph Neural Network to regress this group's emotional atmosphere scores jointly trained with the physiological changes of individuals in the body and their vocal characteristics of speech. Finally, we propose a Multi-modal Multi-label Emotion based on Transformer BLSTM at Group Emotional Atmosphere Network (MMETBGEAN) that explicitly considers individual changes within acoustic and physiological features during a group conversation and incorporates group emotional atmosphere information for recognizing individual self-assessed multi-label emotions. The contributions of this paper can be summarized as follows:

- We introduce a new small-group multi-modal collective task corpus, utilized for building individual self-assessed emotion classification.
- The proposed model can model the dynamic stimulus from acoustic features and the changes of physiological signals in the body to recognize individual self-assessed multi-label emotions.
- We design a graph-level Graph Neural Network to regress multi-target group emotional atmosphere scores with the group-level physiological and acoustic descriptors.

2 RELATED WORKS

2.1 Multi-label Emotion Classification

Multi-label emotion classification seizes more and more scholars' attention because the boundaries between categorical emotions are ambiguous, and some emotions are hard to split clearly, such as sad and frustrated, excited and happy. The study [10] designed a multi-label focal loss to incorporate emotion correlation information into model training for detecting all associated emotions expressed in a given piece of text. Additionally, the multi-label classification problem has been transformed into a sequence generation task to detect multiple emotions in the images [10]. Also, Ju et al. [21], and Zhang et al. [57] have proposed models to simultaneously model label-to-label and modality-to-label dependency within multi-modal scenarios, including text, audio, and video cues. However, all data of most (if not all) databases observers annotate the previous studies utilized on multi-label emotion classification, and there are very few databases consisting of multi-label emotion annotations from subjects' self-assessed feelings.

Different from the studies mentioned earlier, we regard the self-assessed multiple emotions as the learning targets. To the best of our knowledge, this paper is the first attempt to research self-assessed multi-label emotion classification in group conversations using multi-modal features.

2.2 Multi-modal Emotion Recognition

Emotion convey is a complex expression process, and it can be in multiple channels, such as speech, spoken language, and facial expression. More and more computational studies on emotion recognition exploit multiple signals to capture more emotional cues by building multi-modal emotion recognition models. More specifically, most recent research on multi-modal emotion recognition is mainly based on multi-modal fusion frameworks. The study [44] utilized the addition of a parallel stream to the bidirectional language model by integrating acoustic information into contextualized lexical embeddings for emotion recognition. Mittal1 et al. [34] used human gaits and faces cues, various socio-dynamic and situational context information to build emotion detection models. Also, Mittal et al. [33] presented a learning-based method for emotion recognition by combining face, text, and speech cues. Very recently, Zhang et al. [58] have proposed a heterogeneous hierarchical message-passing network to effectively model feature-to-label, modality-to-label, and label-to-label dependency using audio, text and video modalities for multi-label label emotion classification. However, physiology plays a more critical role when we attempt to estimate human self-assessed emotions. The above studies only

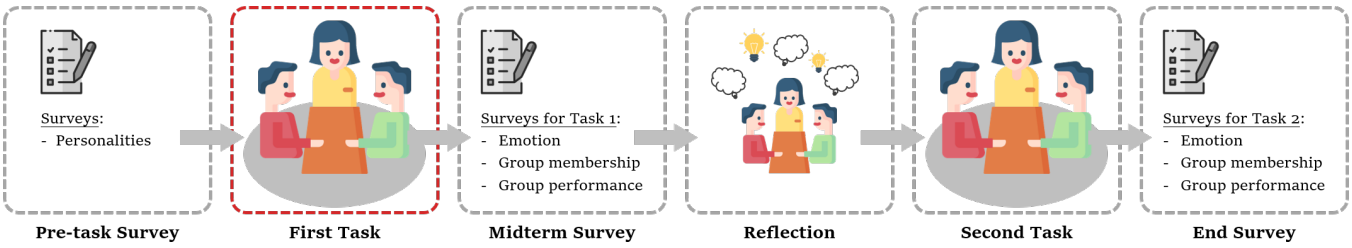


Figure 1: The procedures of collecting NTUBA database [4].

considered and exploited expressive behavior cues outside the body and lacked physiological changes in the body. Nevertheless, people may show no changes in visible activity or expressive behaviors even though they manifest changes in autonomic nervous system activity [13]. On the other hand, the conventional studies on emotion recognition using physiology only focus on single-modality by combining visual stimulus [5], or multiple physiological signals [2].

Different from the above literature, we use emotional cues from expressive acoustics and capture the physiological changes in the body for training the proposed model. We use a graph-level prediction Graph Neural Network (GNN) to model the mutual interactions between group members over time with the physiological and acoustic features.

2.3 Self-assessed Emotion Recognition from Physiology

With the rapid development of wearable devices, the physiology from wearable devices acquiring essential human signals has been the other alternative to analyze individual self-assessed emotions. Very recently, Komuro1 et al. [25] has proposed a customized emotion recognition model (wireless sensors) for individuals according to collected thermography signals and surrounding indoor data, such as temperature and light intensity. However, their emotional ground truths (happy, stress, relax, and sad) come from the third party system, NEC Emotion Analysis Solution [1], not from the participants. Also, Luo et al. [28] has presented a semi-supervised joint domain adaption solution to minimize the cross-domain distribution discrepancies between the multiple source subjects and the target individuals based on the recordings of Electroencephalogram (EEG) traces of subjects on an individual emotion recognition task (valance and arousal). However, their data is collected by inducing participants with one-minute video clips, and their method is hard to be adapted in the group interaction scenario.

To date, most existing works on individual emotion recognition from physiology [56, 59–61] still utilized the induced data, such as DEAP [24], MAHNOB-HCI [42], ASCERTAIN [43], and AMIGOS [30]. While these databases contain many various physiological signals, including EEG, ECG, galvanic skin response (GSR), and Electrodermal activity (EDA), the emotional labels are only dimensional emotion states (not categorical emotion states). Very few studies build a physiology-based individual self-assessed emotion recognition model with PPG signals.

Unlike the works mentioned above, we use the recordings of physiological signals of participants during group interaction in our recently collected corpus, whose scenario is closer to the real-world group conversation. To the best of our knowledge, we are one of the first to predict individual self-assessed multi-label emotions with PPG signals during group discussion comparing the previous physiology-based self-assessed emotion recognition based on PPG signals.

3 PROBLEM FORMULATION

The main goal of this paper is to recognize individual self-assessed multi-label emotions. We briefly explain the definition of this task and its meaning as below.

3.1 Definition of Individual Self-assessed Emotion

We follow the definition of previous works [8, 55]. The individual self-assessed emotion is an emotion perception annotated by subjects in this paper. Unlike most prior studies on emotion recognition, our learning targets come from subjects themselves, not others. Also, the setting of ground truth is the same as the studies [60], but the big difference from them is the types of emotion representation. We use categorical emotions, but they use dimensional emotions (valence and arousal).

4 METHODOLOGY

4.1 NTUBA Database

In this study, we used the National Taiwan University Business Administration (NTUBA) database [4] collected by the College of Management of National Taiwan University (NTU). The database is a Mandarin multi-modal corpus, which includes audio, video, and physiology recordings. The collection atmosphere was to explore the relationship between group behaviors and group performances. Each group was assigned a shopping task by following [51] of diverse scenarios where they were prompted to discuss with each other and concluded the best solution in a limited 30 minutes. All participants have signed informed consent and been fully informed of all experimental procedures under the approved ethical guidelines (IRB approved). There were 80 three-person groups, who mainly were undergraduate students at NTU. This corpus is particularly useful for our study. Most group databases lack physiological recordings or emotional annotations, while the NTUBA database recorded the whole three-person group conversations and their

physiological and audio recordings, allowing us to study in-group individual multi-modal multi-label emotion classification.

The collecting processes have six sessions in total shown in Figure 1. Researchers inquired about prior familiarity between group members and instructed subjects first to fill out the self-reported questionnaires. Then, the first task began and lasted for 30 minutes. Afterward, the participants completed a midpoint survey about the perceived group cohesion and performance. Furthermore, they were asked to reflect on what they had just completed the task and discuss for 10 minutes how to perform better at the second round, and then the second task began and lasted for 30 minutes. The NTUBA carefully and simultaneously recorded audio, video, and physiological signals during sessions. Finally, an endpoint survey was self-assessed reported. In this paper, we only used the data in the first task, which would have less influence on intervention (e.g., familiarity with each other) in this particular interaction setting.

In this article, we only utilized audio modality and one type of physiological signal, Photoplethysmography (PPG), recorded using the wrist-worn E4 sensor with a 64Hz sample rate. Also, the subjects were inquired to annotate their subjective perceptions at the end of each task in the degree of the group's emotion. Specifically, participants responded to 18 emotion categories with the 7-level Likert scale score (1 = "strongly disagree" and 7 = "strongly agree"). To conduct preliminary research, we adopted 8 emotions (*Anger*, *Fear*, *Happiness*, *Annoyed*, *Excited*, *Nervous*, *Frustrated*, *Sadness*) in this paper. We use level 5 as a threshold to decide which emotion is valid among the eight emotions. For example, if he/she gives level 6 on fear and level 4 on anger, it means he/she has fear but no anger, which is the ground truth for the individual multi-label emotion classification task.

On the other hand, we generate the ground truth by a simple aggregation on the group multi-label emotion regression task, which preserves all original scales from group members. Table 1 summarizes the statistics of the data samples with multiple emotion labels and the number of samples in each emotion category. More detailed label pre-processing is below:

Individual self-assessed emotion label: We transform original emotional values from questionnaires into binary class by low class (scores from 1 to 4) and high class (scores from 5 to 7).

Group emotional atmosphere score: We aggregate the scores of members in each group as a single group-level score.

Table 1: The statistics on the NTUBA dataset.

Multi-label	Number		Emotion	Number	
	Audio	PPG		Audio	PPG
none	8	5	Anger	6	4
one	11	7	Fear	29	16
two	36	23	Happiness	127	78
three	67	40	Annoyed	125	80
four	53	35	Excited	118	77
five	14	8	Nervous	147	92
six	2	1	Frustrated	25	14
seven	1	1	Sadness	8	5
eight	0	0	-	-	-
Total	**192**	**120**	Total	**585**	**366**

Table 2: A Table shows an overview of low-level physiological descriptors extracted from NeuroKit and HeartPy.

Modality	Low-Level Descriptors
PPG(35)	RMSSD, meanNN, sdNN, cvNN, CVSD, medianNN, CD, madNN, mcvNN, pNN50, pNN20, DFA_1, ULF, VLF, LF, HF, VHF, LFn, HFn, LF/HF, LF/P, HF/P, Triang, Sample_Entropy, Entropy_Spectral_HF, Entropy_SVD, Total_Power, FD_Petrosian, FD_Higushi, Shannon_h, Shannon, Fisher_Info, Entropy_Multiscale_AUC, Entropy_Spectral_LF, Entropy_Spectral_VLF

4.2 Multi-modal Features Extraction

4.2.1 ***Physiological Descriptor.*** We first pre-process individual physiology data with a low-pass filter cut-off at 60Hz on PPG signals and then use several standard low-level physiological descriptors (LLDs) listed in Table 2 to extract 35-dimensional features by the NeuroKit [29] and HeartPy [46]. Furthermore, a standard z-normalization is used participant-wise on each feature dimension to alleviate the variance coming from individual differences.

4.2.2 ***Audio Descriptor.*** We follow [7] to extract 988-dimensional acoustic features using "emobase.config" in the openSMILE toolkit [14] because they have utilized this feature set to develop a speech emotion recognition model in Mandarin Chinese corpus. It contains 988-dimensional acoustic features, which are further normalized for each speaker using z-score normalization. The further detailed information, please refer to [14].

4.3 Task Definition

We define the following notations to describe two multi-label emotion recognition tasks, including classification and regression. Given the label space with L emotion labels $L = \{emo^1, ..., emo^L\}$, and the multi-modal feature X^A (from audio) and X^P (from physiology) containing the timestamp of length T. Two tasks aim to assign a subset y consists of L' labels in the emotion label space L, e.g., $\{y_1, ..., y_{L'}\}$. Each data sample in two tasks could have multiple labels (one or more), but the labels are binary ($y \in 0, 1$) and a positive real number ($y \in \mathbb{R}^+$) in multi-label emotion classification and regression respectively. We assume that Z is the raw output of the neural network and define the model predictions and data sample target as y^p and y^t in the following loss equation, respectively.

4.3.1 ***Objective Function for Individual Multi-label Emotion Classification.*** The y^p can be estimated by passing Z in the Sigmoid activation function (σ) [32], and the loss can be computed as below.

$$y^p = \sigma(Z) = \frac{1}{1 + exp(-Z)}. \tag{1}$$

$$\begin{aligned} Loss_I(y^p, y^t) = -\frac{1}{L} * \sum_{1}^{N} y^t[i] * log((1 + exp(-y^p[i]))^{-1}) \\ + y^t[i]) * log(\frac{exp(-y^p[i])}{(1 + exp(-y^p[i]))}). \end{aligned} \tag{2}$$

4.3.2 ***Objective Function for Group Multi-label Emotion Regression.*** We follow the study [50] to use a loss function based

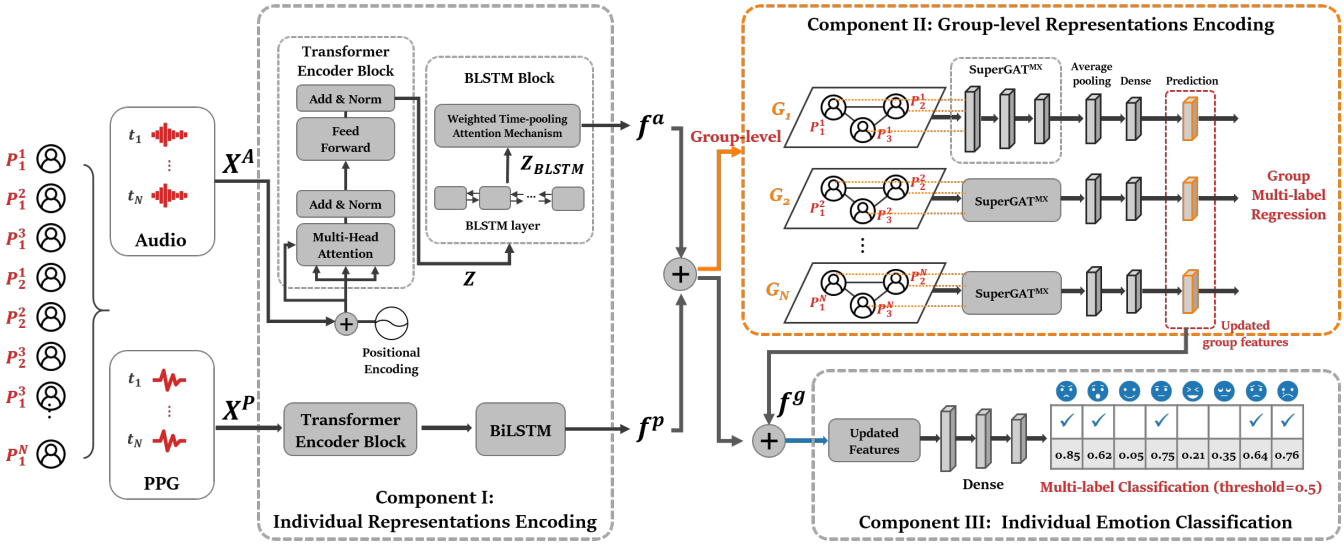


Figure 2: The proposed Multi-modal Multi-label Emotion based on Transformer BLSTM at Group Emotional Atmosphere Network (MMETBGEAN) for individual multi-label emotion classification. First, we define each person as P_j^i, which represents the person is from the j^{th} group with the i^{th} group member. The acoustic and physiological features are retrieved as inputs. We employ a standard Transformer encoder and one BLSTM block to describe the individual time-series information. Then, we build a group-level graph to encode the group-level representations by the group multi-label emotion regression. Finally, the outputs of component I (f^a and f^p) and component II (f^g) are concatenated as the inputs to generate the final output predictions.

on concordance correlation coefficient (CCC) [26] for group multi-label emotion regression in this paper. The y^p can be estimated by passing Z in the ReLU activation function (R) [18], and the loss can be computed as below.

$$y^p = R(Z) = max(0, Z). \tag{3}$$

$$Loss_G(y^p, y^t) = 1 - \frac{\sigma_{y^p y^t}}{\sigma_{y^p}\sigma_{y^t}} * \frac{2\sigma_{y^p y^t}}{\sigma_{y^p}^2 + \sigma_{y^t}^2 + (\mu_{y^p} - \mu_{y^t})^2}. \tag{4}$$

4.4 Computational Framework

We propose a novel Multi-modal Multi-label Emotion based on Transformer BLSTM at Group Emotional Atmosphere Network (MMETBGEAN) for recognizing individual self-assessed multi-label emotions shown in Figure 2, which contains three components and two tasks given acoustic and PPG features as inputs. We briefly explain and introduce each component as below. We first retrieve all the inputs from the people, and each person is defined as P_j^i, which represents the person is from the j^{th} group with the i^{th} seat. Component II and III describe the two tasks, group multi-label emotion regression, and individual multi-label emotion classification, respectively. The objective function optimizes all trainable parameters as below:

$$Loss = Loss_I + Loss_G. \tag{5}$$

4.4.1 ***Component I: Individual Representations Encoding.*** To model an individual's time-series changes in the PPG and acoustic features during group conversations, we are inspired by the TRANS-BLSTM-1 model in [20] to design a model shown in Figure 2 (Component I) based on an encoder part of standard Transformer [47] and a Bidirectional Long Short Term Memory (BLSTM) layer. We hypothesize that the Transformer and BLSTM layers may be complementary to produce a better joint model. More specifically, Component I contains an encoder of standard Transformer, one BLSTM layer with a weighted time-pooling attention mechanism proposed by [31], and then one feedforward layer. Different from [20], we add the BLSTM layer and one feedforward layer (BLSTM-DNN) after the entire encoder of the standard Transformer, not to replace the feedforward layer with a BLSTM layer. Additionally, we add a weighted time-pooling attention mechanism on the output of BLSTM, which has been proposed and used for speech emotion recognition [6, 31], for better performance. There are two Transformer BLSTMs in Component I given acoustic and PPG features as inputs separately, and the outputs of two Transformer BLSTMs are denoted by f^a and f^p in the paper, respectively.

Since we employ an encoder of standard Transformer to build Component I, we summarize its particular mechanism. The encoder maps an input representations $X = (x_1, ..., x_n)$ to a sequence of continuous representations $Z = (z_1, ..., z_n)$. The shape size of X is (batch size, the length of timestamp, the number dimension of input feature). The self-attention is suitable for us to model the changes in physiology and acoustics according to varying times, and the attention mechanism can compute the relations between timestamps. We introduce a self-attention operation to focus on these timestamps for employing relevant features. Given **X** as input of the encoder, we follow [47] to compute "Scaled Dot-Product Attention" (contains queries and keys of dimension d_k, and values

of dimension d_v) representation $\mathbf{H}$ in the following. The definition of Q, K, and V for query, key, and value is in [47].

$$\mathbf{H} = softmax(\frac{\left(\mathbf{X}\mathbf{W}^Q(\mathbf{X}\mathbf{W}^K)^T\right)}{\sqrt{d_k}})\mathbf{C}\mathbf{W}^V. \quad (6)$$

In the proposed model, we deploy the multi-head attention version, which can be computed as:

$$\mathbf{H_j} = softmax(\frac{\left(\mathbf{W}_j^Q\mathbf{X}(\mathbf{W}_j^K\mathbf{X})^T\right)}{\sqrt{d_k}})\mathbf{W}_j^V\mathbf{X}, \quad (7)$$

$$\mathbf{Z} = Concat(\mathbf{H_j}, \ ..., \ \mathbf{H_h})\mathbf{W}^O, \quad (8)$$

where $W^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W^K \in \mathbb{R}^{d_{model} \times d_k}$, $W^V \in \mathbb{R}^{d_{model} \times d_v}$, $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ are learn-able parameter matrices. Additionally, h and d_{model} means the number of heads and the number dimensions of outputs of an encoder, respectively. Besides, we add the standard "positional encodings" in [47] to the input embeddings by summing. Since we believe that emotional sections are in order, we inject positional information about the relative order in the sequence, which means timestamp in this paper.

Now, given the output of an encoder of Transformer Z, the BLSTM layer then generates an output sequence $y = (y_1, \ ..., \ y_t)$. T equals the (maximum) length timestamp of input features (acoustic and physiological features), and t is at each timestamp. We introduce a weighted-pooling with local attention proposed by [31] as below. A softmax function is performed to the results to get a set of final weights for the frames which sum to unity:

$$\alpha_t = \frac{exp(u^T y_t)}{\sum_{i=1}^{T} exp(u^T y_i)}, \quad (9)$$

where u is the attention parameter vector.

The above attention weights are utilized in a weighted average in sequence to get the output representation:

$$Z_{BLSTM} = \sum_{i=1}^{T} \alpha_t y_t. \quad (10)$$

Finally, given Z_{BLSTM}, the feedforward layer then generates an output representation f^a and f^p according to types of input features.

4.4.2 ***Component II: Group-level Representations Encoding.*** To model the group emotional atmosphere, we utilize the self-supervised GAT (SuperGAT) layer in "MX" version [22] to build a graph-level prediction graph neural network (GNN) for regressing group multi-label emotional atmosphere scores. Component II consists of three SuperGAT layers, one global mean pooling layer, one feedforward layer, and one prediction layer with ReLU activation function. The global mean pooling layer averages all nodes in a graph for a prediction over a whole graph. The goal of component II is to regress group emotions by an entire graph instead of single nodes or edges. Each group member is linked to a node, and the edges in the graph are the bonds between group members. In graph regression, an attributed graph is given as an input, and a real-valued output variable is predicted. Each graph represents one group in the NTUBA, and the outputs of component I, f^a and f^p, are concatenated as the nodes' attributes. The input attributed graph is self-loop and unidirectional.

The SuperGAT works under the assumption that two nodes are more relevant than others if two nodes are linked. Since SuperGAT performs well on nodes and links level prediction tasks, we adapt it to apply to this paper's graph perdition task. We clearly describe the **MX** mechanism of SuperGAT in the following.

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, N is the number of nodes and F^l is the number of features at layer l. Graph attention layer takes a set of features $\mathbb{H}^l = \{\mathbf{h}_1^l, \ , \ ..., \ \mathbf{h}_N^l\}$ ($\mathbb{H}^1$ is equal to $Concat(f^a, \ f^p)$ in this paper). To compute $\mathbf{h}_i^{l+1}$, the model multiplies the trainable weight matrix $\mathbf{W}^{l+1} \in \mathbb{R}^{F^{l+1} \times F^l}$ to $\mathbb{H}^l$, linearly combines the features of its first-order neighbors (including itself) $j \in \mathbb{N}_i \cup \{i\}$ by attention coefficients α_{ij}^{l+1}, finally applies a non-linear activation ρ. That is,

$$\mathbf{h}_i^{l+1} = \rho\left(\sum_{j \in \mathbb{N}_i \cup \{i\}} \alpha_{ij}^{l+1}\mathbf{W}^{l+1}\mathbf{h}_j^l\right), \quad (11)$$

where the two types of attention α_{ij} are computed as (the LeakyReLU function is proposed by [54])

$$\alpha_{ij}^{l+1} = \frac{exp(\mathbf{LeakyReLU}(\mathbf{e_{ij}^{l+1}}))}{\sum_{k \in \mathbb{N} \cup \{i\}} exp(\mathbf{LeakyReLU}(e_{ik}^{l+1}))}, \quad (12)$$

$$e_{ij}^{l+1} = (\mathbf{a}^{l+1})^T\left[\mathbf{W}^{l+1}\mathbf{h}_i^l || \mathbf{W}^{l+1}\mathbf{h}_j^l\right] \cdot \sigma\left((\mathbf{W}^{l+1}\mathbf{h}_i^l)^T\mathbf{W}^{l+1}\mathbf{h}_j^l\right), \quad (13)$$

where $\mathbf{a}^{l+1} \in \mathbb{R}^{2F^{l+1}}$ is the coefficients computed by the original GAT [48].

The self-supervised task in the SuperGAT layer is a link prediction using the attention values as input to predict the likelihood ϕ_{ij} (shown in the below) that an edge exists between nodes, and it can softly drop neighbors that are not likely linked while implicitly assigning importance to the remaining nodes.

$$\phi_{ij} \ = \ \sigma\left((\mathbf{W}\mathbf{h}_i)^T\mathbf{W}\mathbf{h}_j\right). \quad (14)$$

Finally, the outputs of the SuperGAT layers are as the inputs to one global mean pooling layer, one feedforward layer, and then the prediction layer with the ReLU activation function. We take the embeddings before the prediction layer, denoted by f^g, for the next component III.

4.4.3 ***Component III: Individual Self-assessed Emotion Classification.*** The outputs of component I (f^a and f^p) and component II (f^g) are simply concatenated as the inputs of component III, which consists of three feedforward layers and then one prediction layer with a sigmoid activation function. Notice that each person in the one group has the same embeddings from component II. Given the inputs ($Concat(f^a, \ f^p, \ f^g)$), the component III then generates the final output predictions.

5 EXPERIMENT

5.1 Experimental Setup

5.1.1 ***Implementation Details.*** The dimension of PPG features (d^p) and acoustic features (d^a) are 35 and 988 respectively. The number of nodes of three feedforward (dense) layers in component III is $[d, \ d/2, \ d/4]$. The model is trained with a fixed-length timestamp defined as a function on the various window and step sizes. We use zero padding to transform the length of the timestamp of each data sample into the same size if the length is less than the maximum timestamp.

Table 3: The summarized results of models of two categories on multi-label emotion classification task; the five metrics are Hamming Loss (HL), multi-label Accuracy (Acc), macro-F_1 measure (F_1), macro-Precision (P), and macro-Recall (R).

Category	Approach	$HL(\downarrow)$	$Acc(\uparrow)$	$F_1(\uparrow)$	$P(\uparrow)$	$R(\uparrow)$
Multi-label	BR [38]	0.185 ± 0.027	0.238 ± 0.042	0.404 ± 0.030	0.339 ± 0.016	0.500 ± 0.021
	CC [27]	0.212 ± 0.011	0.162 ± 0.038	0.377 ± 0.009	0.339 ± 0.007	0.433 ± 0.015
	LP [45]	0.229 ± 0.020	0.152 ± 0.047	0.401 ± 0.034	0.376 ± 0.056	0.444 ± 0.022
Multi-label & Multi-modal	MMET	0.179 ± 0.030	0.365 ± 0.053	0.379 ± 0.046	0.388 ± 0.042	0.420 ± 0.053
	MMETB	0.168 ± 0.021	0.402 ± 0.035	0.422 ± 0.055	0.424 ± 0.028	0.441 ± 0.050
	MMETBGEAN	**0.152 ± 0.019**	**0.458 ± 0.033**	**0.502 ± 0.040**	**0.495 ± 0.021**	**0.513 ± 0.011**

Moreover, several hyper-parameters are chosen by grid-searched. The learning rate is set among [0.005, 0.001] with the learning rate adjusting mechanism, the cosine warm-up scheduler [19], and the number of warm-ups is 20. The number of multi-head attentions is 8, the number of BLSTM nodes is fixed as [2, 4, 8], batch size is fixed as [8, 16], the max epoch is 1000, the drop out is set to 0.2, and optimizer is ADAMAX [23]. The whole framework is implemented using the Pytorch toolkit [35], Pytorch Geometric [16], and PyTorch Lightning [15]; an early-stopping criterion chooses the hyperparameters via monitoring validation loss according to different tasks. Once the training is finished, we select the model with the lowest Hamming Loss as the final model to evaluate the performance on the validation set. Since the NTUBA has no testing set, we use a 5-fold cross-validation evaluation method, and each fold is split based on the group-independent and class-balanced experiments.

5.1.2 ***Evaluation Metrics.*** We follow the study [53] to present five evaluation metrics to measure the performances of all approaches on multi-label emotion detection tasks in this paper. The five metrics have been universal used in some multi-label classification problems [17, 21, 52], which are Hamming Loss (HL), multi-label Accuracy (Acc), macro-F_1 measure (F_1), macro-Precision (P), and macro-Recall (R). It is noticed that smaller HL corresponds to better classification quality, while larger Acc, P, R, and F_1 measure represents better classification quality.

5.2 Model Comparison

For a penetrating comparison, we conduct various baseline approaches in two categories. Since some baseline approaches are unable to model time-series information, we compute 15 statistical functionals[1] on the timestamp dimension to extract session-level unique features.

The baseline models in the first category ("Multi-label") have been used to treat the multi-label classification task without the ability to premeditate the dependence of multi-modality. For these approaches, the multi-modal inputs are concatenated as their inputs. We briefly introduce each baseline model as below. (1) **BR**[2] [38], Binary Relevance, which ignores the correlations between labels by transforming the multi-label task into multiple single-label binary classification problem, (2) **CC**[2] [27], Classifier Chains, which converts the multi-label task into a chain of binary classification problem and takes high-order label correlations into account, (3) **LP**[2][45], Label Powerset, which creates one multi-class classifier for each combination of labels proven in the training set.

The baseline model in the second category ("Multi-label & Multi-modal") can model temporal context by multi-modal jointly training strategy for multi-label classification. In this paper, we hypothesize that the BLSTM layer and the group emotion can improve the model performance, so we do the ablation study to remove these particular components compared with the proposed model. Therefore, the first baseline model is (4) **MMET**, which removes the BLSTM block and group constraints. The other one is (5) **MMETB**, which removes the group constraints representations.

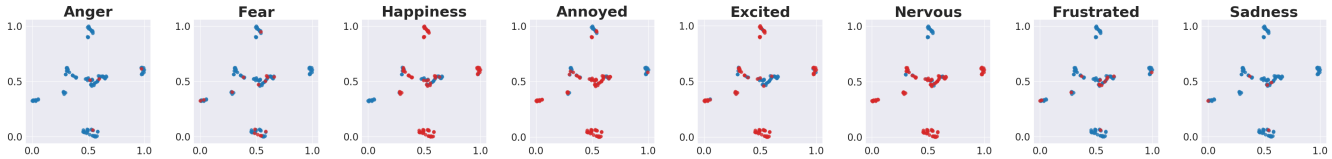
5.3 Emotion Classification Results

Table 3 sums up the experimental results. The proposed model, **MMETBGEAN**, surpasses the baseline methods. The model absolutely improves 9.3%, 12.3%, 10.7% and 9.3% for Acc, F_1, P, and R respectively than the **MMET**. Also, there are several observations. First of all, the proposed approach exists a large discrepancy across modalities, and these discrepancy results could come from either subject differences or group composition; however, they deteriorate multi-label emotion recognition performance when using **BR**, **CC**, **LP** without any multi-modality jointly learning techniques. This result suggests that the physiological and acoustic features do not directly embed emotional information because the intricate dependency between two modalities requires a sophisticated algorithm to learn and model. The early fusion approaches inescapably lead to performance loss.

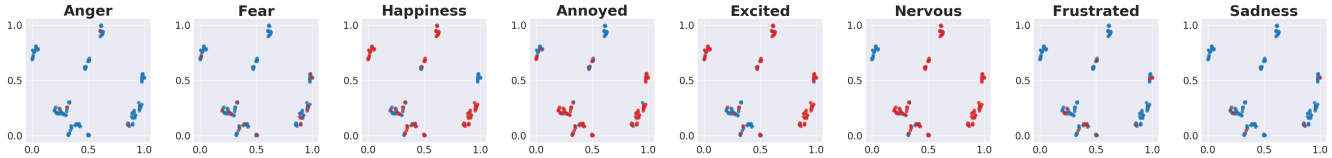
Moreover, we observe that the improvement compared with other baselines is not apparent in this task while our ablated model **MMETB** has been regarded as the most substantial baseline. We explain that the multi-modal multi-label approaches could only minimize the discrepancy of multi-modal data because they mainly focus on mapping the individual feature distribution conditioned on predicted labels. Additionally, **MMETB** is unsuccessful in thinking about the local variations (like conversation dynamics in groups), which is particularly crucial for individual emotion detection during group interactions. In contrast, the proposed **MMETBGEAN** mainly utilizes the group constraint based on a graph neural network (GNN) to model the links of multi-modal representations under in-group individuals, which helps in getting improved results on the individual multi-label emotion classification task during group conversations.

[1] max/min value and respective relative position within input, mean/median value, standard deviation, first percentile, ninety-ninth percentile, the difference between ninety-ninth percentile and first percentile, skewness, kurtosis, quartile 1, quartile 3, and interquartile range

[2] http://scikit.ml/



(a) Testing features extracted by the MMETB model.



(b) Testing features extracted by the proposed MMETBGEAN.

Figure 3: Scatter plot the results by t-SNE for testing features derived from MMETB model and MMETBGEAN model for the two classes of Yes (Red) and No (Blue).

6 ANALYSIS

To understand the potential modulation of group emotional atmosphere representations toward multi-label individual responses, we specifically analyze the multi-faceted on the proposed model. We demonstrate the indispensable multi-modal elements on the classification task, and then we investigate the performance of the regression task based on the different graph convolutional layers. Moreover, we visualize the representations with t-Distributed Stochastic Neighbor Embedding (t-SNE) (Figure 3a and 3b) along with multi-label emotions to show the effectiveness of integrating the group constrained representations.

6.1 Ablation Study

To demonstrate the indispensable of a multi-modal method for multi-label emotion classification, we show the performance of Single-modal Multi-label Emotion based on Transformer BLSTM at Group Emotional Atmosphere Network (**SMETBGEAN**) approach, which only models a single modality. Table 4 shows the performance of **SMETBGEAN** and **MMETBGEAN** approaches. We observe that **SMETBGEAN** with physiological modality outperforms the counterparts with the auditory modality, which suggests that the physiological modality consists of more emotional information than the audio one on the individual multi-label emotion classification task. Additionally, the proposed **MMETBGEAN** achieves the highest performance, and it shows that the complementary between two modalities. This finding is consistent with our motivation and hypothesis that emotional information is conveyed in different channels.

6.2 Visualization

To illustrate the power of the proposed model, we plot the representations of data encoded with the strongest baseline without the group representation factor – **MMTB** and the proposed method – **MMETBGEAN** by using t-SNE. We display the t-SNE according to each emotion category one by one to explicitly show the effect on different emotions. Figure 3 (a) shows that these representations are indistinguishable under mostly emotions. In contrast, in Figure 3 (b), due to the group constraint, the encoded representations in the emotion "*Happiness*", "*Annoyed*", "*Excited*" and "*Nervous*" are indeed more distinguishable even though the distinction of some emotions are still not obvious. Overall, the group constraint improves recognition power, and it indicates that we should consider the group constraint when modeling group conversations on the individual multi-label emotion classification task.

Table 4: A table summarizes the single-modal model (SMETBGEAN) and multi-modal model (MMETBGEAN) on a multi-label emotion classification task in the NTUBA dataset.

Approach	Modality	*HL*	*Acc*	F_1	*P*	*R*
SMETBGEAN	Audio	0.177	0.380	0.386	0.374	0.408
	PPG	0.169	0.402	0.451	0.389	**0.440**
MMETBGEAN	Audio+PPG	**0.152**	**0.458**	**0.502**	**0.495**	0.433

7 CONCLUSION

The study presents a novel framework with graph-based group constraints for in-group individual multi-label emotion classification, named Multi-modal Multi-label Emotion based on Transformer BLSTM at Group Emotional Atmosphere Network (MMETBGEAN). The MMETBGEAN explicitly considers individual changes within PPG and acoustic features during group conversations and incorporates the group emotion intensity information by the SuperGAT layers for individual multi-label emotion classification. The experiments show that the proposed method evaluated on the NTUBA reaches promising results on the multi-label emotion classification task. To our best knowledge, this is one of the first works to research individual multi-modal multi-label emotion classification in group conversations. However, our work still has a few limitations. For instance, the NTUBA is indeed relatively small. Nevertheless, there is no existing database that contains physiological data collected. In the future work, we will train the models to learn the self-assessed emotion scores (7-point Likert) directly, and we plan to employ the personalities and interaction dynamics under group conversations to deepen group composition analysis [3], such as the effect of compositional personalities on group emotions and conversational temporal dynamics [9] respectively.

REFERENCES

[1] K Abe and S Iwata. 2019. NEC's emotion analysis solution supports work style reform and health management. *NEC Tech. J* 14, 1 (2019), 44–48. https://www.nec.com/en/global/techrep/journal/g19/n01/pdf/190109.pdf

[2] Değer Ayata, Yusuf Yaslan, and Mustafa E Kamasak. 2020. Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. *Journal of Medical and Biological Engineering* (April 2020), 1–9. https://doi.org/10.1007/s40846-019-00505-7

[3] Sigal G Barsade and Donald E Gibson. 1998. Group emotion: A view from top and bottom. (1998), 81–102.

[4] Woan-Shiuan Chien, Huang-Cheng Chou, and Chi-Cun Lee. 2021. Belongingness and Satisfaction Recognition from Physiological Synchrony with A Group-Modulated Attentive BLSTM under Small-group Conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. https://doi.org/10.1145/3461615.3485410

[5] Woan-Shiuan Chien, Hao-Chun Yang, and Chi-Chun Lee. 2020. Cross Corpus Physiological-Based Emotion Recognition Using a Learnable Visual Semantic Graph Convolutional Network. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 2999–3006. https://doi.org/10.1145/3394171.3413552

[6] Huang-Cheng Chou and Chi-Chun Lee. 2019. Every Rating Matters: Joint Learning of Subjective Labels and Individual Annotators for Speech Emotion Classification. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5886–5890. https://doi.org/10.1109/ICASSP.2019.8682170

[7] Huang-Cheng Chou and Chi-Chun Lee. 2020. Learning to Recognize Per-Rater's Emotion Perception Using Co-Rater Training Strategy with Soft and Hard Labels. In *Proc. Interspeech 2020*. 4108–4112. https://doi.org/10.21437/Interspeech.2020-1714

[8] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 292–298. https://doi.org/10.1109/ACII.2017.8273615

[9] Huang-Cheng Chou, Yi-Wen Liu, and Chi-Chun Lee. 2019. JOINT LEARNING OF CONVERSATIONAL TEMPORAL DYNAMICS AND ACOUSTIC FEATURES FOR SPEECH DECEPTION DETECTION IN DIALOG GAMES. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 1044–1050. https://doi.org/10.1109/APSIPAASC47483.2019.9023050

[10] J. Deng and F. Ren. 2020. Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning. *IEEE Transactions on Affective Computing* (2020), 1–1. https://doi.org/10.1109/TAFFC.2020.3034215

[11] Sidney K. D'mello and Jacqueline Kory. 2015. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Comput. Surv.* 47, 3, Article 43 (Feb. 2015), 36 pages. https://doi.org/10.1145/2682899

[12] Maria Egger, Matthias Ley, and Sten Hanke. 2019. Emotion Recognition from Physiological Signal Analysis: A Review. *Electronic Notes in Theoretical Computer Science* 343 (2019), 35–55. https://doi.org/10.1016/j.entcs.2019.04.009 The proceedings of AmI, the 2018 European Conference on Ambient Intelligence.

[13] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384. https://doi.org/10.1037/0003-066X.48.4.384

[14] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor *(MM '10)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1873951.1874246

[15] WA Falcon and .al. 2019. PyTorch Lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning* 3 (2019).

[16] Matthias Fey and Jan Eric Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. *CoRR* abs/1903.02428 (2019). http://arxiv.org/abs/1903.02428

[17] Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 4441–4453. https://doi.org/10.18653/v1/2020.coling-main.393

[18] K Fukushima. 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics* 36, 4 (1980), 193—202. https://doi.org/10.1007/bf00344251

[19] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r14EOsCqKX

[20] Zhiheng Huang, Peng Xu, Davis Liang, Ajay Mishra, and Bing Xiang. 2020. TRANS-BLSTM: Transformer with Bidirectional LSTM for Language Understanding. *CoRR* abs/2003.07000 (2020). arXiv:2003.07000 https://arxiv.org/abs/2003.07000

[21] Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-Based Label Set Generation for Multi-Modal Multi-Label Emotion Detection. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 512–520. https://doi.org/10.1145/3394171.3413577

[22] Dongkwan Kim and Alice Oh. 2021. How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Wi5KUNlqWty

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[24] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[25] Nobuyoshi Komuro, Tomoki Hashiguchi, Keita Hirai, and Makoto Ichikawa. 2021. Predicting individual emotion from perception-based non-contact sensor big data. *Scientific reports* 11, 1 (January 2021), 2317. https://doi.org/10.1038/s41598-021-81958-2

[26] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268. http://www.jstor.org/stable/2532051

[27] Oscar Luaces, Jorge Díez, José Barranquero, Juan José Coz, and Antonio Bahamonde. 2012. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence* 1, 4 (2012), 303–313. https://doi.org/10.1007/s13748-012-0030-x

[28] Junhai Luo, Man Wu, Zhiyan Wang, Yanping Chen, and Yang Yang. 2021. Progressive low-rank subspace alignment based on semi-supervised joint domain adaption for personalized emotion recognition. *Neurocomputing* 456 (2021), 312–326. https://doi.org/10.1016/j.neucom.2021.05.064

[29] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing, journal=Behavior Research Methods. (02 Feb 2021). https://doi.org/10.3758/s13428-020-01516-y

[30] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2021. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* 12, 2 (2021), 479–493. https://doi.org/10.1109/TAFFC.2018.2884461

[31] S. Mirsamadi, E. Barsoum, and C. Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2227–2231. https://doi.org/10.1109/ICASSP.2017.7952552

[32] Thomas M. Mitchell. 1997. *Machine Learning* (1 ed.). McGraw-Hill, Inc., USA.

[33] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 02 (Apr. 2020), 1359–1367. https://doi.org/10.1609/aaai.v34i02.5492

[34] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).

[36] P. C. Petrantonakis and L. J. Hadjileontiadis. 2010. Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis. *IEEE Transactions on Affective Computing* 1, 2 (2010), 81–97. https://doi.org/10.1109/T-AFFC.2010.7

[37] Emilie Qiao-Tasserit, Maria Garcia Quesada, Lia Antico, Daphne Bavelier, Patrik Vuilleumier, and Swann Pichon. 2017. Transient emotional events and individual affective traits affect emotion recognition in a perceptual decision-making task. *PLOS ONE* 12, 2 (02 2017), 1–16. https://doi.org/10.1371/journal.pone.0171375

[38] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier Chains for Multi-Label Classification. In *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II* (Bled, Slovenia) *(ECMLPKDD'09)*. Springer-Verlag, Berlin, Heidelberg, 254–269.

[39] Scherer, Klaus R. 1999. Appraisal theory. (1999), 637–663. https://doi.org/10.1002/0470013494.ch30

[40] Dongmin Shin, Dongil Shin, and Dongkyoo Shin. 2017. Development of Emotion Recognition Interface Using Complex EEG/ECG Bio-Signal for Interactive Contents. *Multimedia Tools Appl.* 76, 9 (May 2017), 11449–11470. https://doi.org/10.1007/s11042-016-4203-7

[41] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A Review of Emotion Recognition Using Physiological

Signals. *Sensors* 18, 7 (2018). https://doi.org/10.3390/s18072074

[42] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

[43] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. 2018. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing* 9, 2 (2018), 147–160. https://doi.org/10.1109/TAFFC.2016.2625250

[44] S. Y. Tseng, S. Narayanan, and P. Georgiou. 2021. Multimodal Embeddings From Language Models for Emotion Recognition in the Wild. *IEEE Signal Processing Letters* 28 (2021), 608–612. https://doi.org/10.1109/LSP.2021.3065598

[45] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-Label Classification: An Overview. *Int. J. Data Warehous. Min.* 3, 3 (2007), 1–13. https://doi.org/10.4018/jdwm.2007070101

[46] P Van Gent, H Farah, N Nes, and B van Arem. 2018. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. In *Proceedings of the 6th HUMANIST Conference*. 173–178. http://resolver.tudelft.nl/uuid:5c638e14-d249-4116-aa05-2e566cf3df02

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[48] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJXMpikCZ

[49] Gyanendra K. Verma and Uma Shanker Tiwary. 2014. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage* 102 (2014), 162–172. https://doi.org/10.1016/j.neuroimage.2013.11.007 Multimodal Data Fusion.

[50] Felix Weninger, Fabien Ringeval, Erik Marchi, and Björn Schuller. 2016. Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2196–2202.

[51] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (2010), 686–688. https://doi.org/10.1126/science.1193147 arXiv:https://science.sciencemag.org/content/330/6004/686.full.pdf

[52] Guoqiang Wu and Jun Zhu. 2020. Multi-label classification: do Hamming loss and subset accuracy really conflict with each other?. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 3130–3140. https://proceedings.neurips.cc/paper/2020/file/20479c788fb27378c2c99eadcf207e7f-Paper.pdf

[53] Xi-Zhu Wu and Zhi-Hua Zhou. 2017. A Unified View of Multi-Label Performance Measures. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3780–3788. http://proceedings.mlr.press/v70/wu17a.html

[54] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. *CoRR* abs/1505.00853 (2015). arXiv:1505.00853 http://arxiv.org/abs/1505.00853

[55] H. Yang and C. Lee. 2019. Annotation Matters: A Comprehensive Study on Recognizing Intended Self-reported, and Observed Emotion Labels using Physiology. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7. https://doi.org/10.1109/ACII.2019.8925516

[56] Zhong Yin, Lei Liu, Jianing Chen, Boxi Zhao, and Yongxiong Wang. 2020. Locally robust EEG feature selection for individual-independent emotion recognition. *Expert Systems with Applications* 162 (2020), 113768. https://doi.org/10.1016/j.eswa.2020.113768

[57] Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving Multi-label Emotion Classification by Integrating both General and Domain-specific Knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, Hong Kong, China, 316–321. https://doi.org/10.18653/v1/D19-5541

[58] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing. *Proceedings of the AAAI Conference on Artificial Intelligence* 01 (Feb. 2021). https://www.aaai.org/AAAI21Papers/AAAI-2554.ZhangD.pdf

[59] Wei Zhang, Zhong Yin, Zhanquan Sun, Ying Tian, and Yagang Wang. 2020. Selecting transferrable neurophysiological features for inter-individual emotion recognition via a shared-subspace feature elimination approach. *Computers in Biology and Medicine* 123 (2020), 103875. https://doi.org/10.1016/j.compbiomed.2020.103875

[60] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. 2018. Personality-Aware Personalized Emotion Recognition from Physiological Signals. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm, Sweden) *(IJCAI'18)*. AAAI Press, 1660–1667. https://dl.acm.org/doi/abs/10.5555/3304415.3304651

[61] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer. 2019. Personalized Emotion Recognition by Personality-Aware High-Order Learning of Physiological Signals. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1s, Article 14 (Jan. 2019), 18 pages. https://doi.org/10.1145/3233184