2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822803

EmoTracer: A User-independent Wearable Emotion Tracer with Multi-source Physiological Sensors Based on Few-shot Learning

Wenting Kuang*, Shuo Jin*, Danyang Wang*, Yuda Zheng*, Yongpan Zou*, Kaishun Wu‡
*College of Computer Science and Software engineering, Shenzhen University
‡Information Hub, Hong Kong University of Science and Technology (Guangzhou)
{kuangwenting2023, jinshuo2022, 2021150056}@email.szu.edu.cn,
Yolanda_swam@163.com, yongpan@szu.edu.cn, wuks@hkust-gz.edu.cn

***Abstract*—The rising prevalence of mood disorders, including depression and anxiety, underscores the critical state of physical and mental health issues in contemporary society. Automatic emotion recognition technology emerges as a potential tool for monitoring mental health disorders, and offering valuable guidance through human-computer interfaces. However, existing technologies grapple with limitations, such as emotional concealment, potential privacy leakage, and specific device location restrictions. To overcome these limitations, we propose a mobile wearable emotion recognition system called EmoTracer incorporating ubiquitous multi-source sensors to measure physiology signals, and address challenges such as the intricate signal-emotion relationship, sparse data processing, and notable disparities among different subjects. We implement a real-time prototype and carry out comprehensive experiments to evaluate its performance. For six basic emotions, the results show that EmoTracer can achieve 95.6% accuracy in intra-subject emotion classification and 80.7% accuracy with 5 shots in cross-subject emotion classification.**

***Index Terms*—emotion recognition, wearable device, physiological signal, cross-subject**

I. Introduction

During the COVID-19 pandemic, mental health became a focal of research, with depression ranked as the 10th most studied topic [1]. According to the data reported by the World Health Organization (WHO) in 2021 [2], an alarming 280 million people worldwide are grappling with depressive disorders, impacting approximately 3.8% of the global population. Emotions like joy and happiness enhance well-being, while sadness, pain, and anxiety can negatively impact mental and physical health, even leading to harmful behaviors. To sum up, accurate and timely assessment of emotions is paramount.

In recent years, the field of affective computing has gained substantial attention. Researchers [3]–[5] commonly employ image and audio processing techniques to analyze facial expressions, body language, or acoustic signals, aiming to discern emotional states intuitively. However, these methods are challenged by privacy leakage results from camera or microphone recordings, and susceptibility to external environmental interference. Additionally, the analysis of physiological signals has been shown to closely relate to emotional changes [6]. These physiological signals are spontaneous reactions regulated by the nervous system that are difficult for individuals to deliberately manipulate, thus enhancing the accuracy and objectivity of emotion recognition. Studies [7]–[9] have utilized physiological signals like electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), and blood volume pulse (BVP). Despite the high accuracy achieved by these methodologies, they are restricted by substantial costs and operational complexity because of the utilization of professional devices. Other physiological signals like respiratory (RSP), skin temperature (SKT) , and galvanic skin response (GSR) can be captured by ubiquitous sensors in a movable manner [10], but their emotion recognition accuracy is often hindered by the lower sampling rates and resolution compared to professional devices.

Based on the above issues, we propose a wearable emotion recognition system called EmoTracer. For emotion assessment, we leverage physiological signals as the foundations. This is because emotional changes can cause variations in adrenaline levels, adjustments in blood circulation supply and changes in respiratory rate, altering physiological signals like heart rate, blood oxygen, galvanic skin and temperature [11]–[14]. Importantly, they can be continuously and conveniently tracked by ubiquitous sensors, providing the possibility of real-time and convenient emotion intervention. For privacy protection and ethical consideration, our system handles sensitive information, including emotional states and physiological responses, with the utmost care. We guarantee that all users grant their explicit informed consent before data collection, and designs ID settings like U1 instead of collecting personal information such as names, ages, or occupations. Any user-provided information is treated with data masking and is not used for commercial purposes. Nevertheless, addressing emotion recognition through physiological signals is not straightforward. We need to tackle the following challenges. First, to capture subtle underlying interactions between physiological signals and emotions, robust data preprocessing and neural network design are essential. Second, sensitivity to conditions and individual difference makes physiological signal samples scarce, risking overfitting and hindering neural network generalization. Third, the diversity emotional and physiological

responses among individual complicates cross-subject emotion categorization.

In response to the aforementioned challenges, we draw from Long Shot-Term Memory (LSTM), proposing LSTM Multi-source Fusion Network (LMFN) that design proprietary feature extractors for physiological signals and then splices them along the feature dimension into the fusion feature extractor. This approach is targeted at minimizing redundant information across different sources. Moreover, we combine a few-shot learning for cross-subject emotion recognition model, to address challenges such as sparse sample size and low accuracy in cross-subject emotion recognition tasks. The results show that our system optimize the performance of intra-subject and cross-subject classification for six emotions.

In summary, the contributions of our research are summarized as follows:

- We design an intelligent wristband embedded with multiple ubiquitous sensors and a mobile application to track emotional states, offering a portable and unobtrusive tool that seamlessly integrated into daily life.
- We propose LMFN to uncover the intricate correlations between diverse sources, achieving an accuracy of 95.6% in intra-subject emotion recognition.
- We combine a few-shot learning on cross-subject emotion recognition, allowing new users to provide sparse shots without retraining the model. The evaluation results indicate that its accuracies of 80.7% with 5 shots per class, meeting the real-life requirement of emotion recognition.

II. Related Work

A. Emotion recognition based on external traits

Visual technology is widely studied in emotion recognition, where Convolutional Neural Networks (CNNs) stand out as pivotal for image processing. For instance, Mehendale et al. [15] introduce a dual-component CNN to isolate facial features from their background and distills expression vectors. And Cui et al. [16] employs CNN classifier to deduce emotional states based on human posture and form. These CNNs excel at learning spatial details but struggle with dynamic changes in image, leading us to consider the temporal aspect, crucial for fully expressing human emotions. Subsequently, researchers turn to the acquisition of vocal signals. The Recurrent Neural Networks (RNNs) and their variants, such as LSTM and Gated Recurrent Units (GRUs), are skilled in analyzing sequential signals. Wang et al. [17] proposed a dual-sequence LSTM to capture temporal and frame-level emotional cues. However, this model might be less effective in data-scarce scenarios, potentially reducing prediction accuracy. Furthermore, it raises privacy concerns due to reliance on camera or audio data, and can be challenged by individuals concealing their true emotion.

B. Emotion recognition based on internal indicators

The generation and fluctuation of internal indicators like physiological signals can not be deliberately controlled by individuals, gaining traction in affective computing. Technology of multi-modal data fusion have marked a significant trend. Physiological signals, uncontrollable by individuals, are , where multi-modal data fusion have marked a significant trend. The study [18] apply stacked autoencoders to combine EEG, pulse, and blood oxygen signals, achieving an increasing accuracy from 53.8% to 73.5% in emotion recognition. The 1D-CNN-SVM emotion recognition model [19], exemplifies the power of multi-level fusion, realizing a 7% enhancement in accuracy through the integration of EEG with peripheral physiology signals. Advanced method [20], based on CNN and LSTM, have achieved high accuracies with individual signals, but decision-level fusion through majority voting can push these to an impressive 99.0%. Despite these advantages, challenges remain, including the potential for lost information in the final feature extraction layer and the influence of individual physiological signal variability. Additionally, the significant variability in physiological signals across individuals presents a challenge for model tuning.

III. System Design

A. Overview

Considering the limited API interfaces of commercial wearable devices, which restrict deep analysis and personalized data processing, we design EmoTracer , a wearable emotion recognition system. This system integrates multiple ubiquitous sensors into wearable device, aiming to capture physiological signals like heart rate, blood oxygen level, galvanic skin and temperature. When monitoring physical states, EmoTracer explores the correlation between physiological signals and emotion, thereby achieving emotion recognition. This process can be further divided into four parts: data collection, data preprocessing, multi-source fusion, and emotion recognition models. In the emotion recognition task, EmoTracer can effectively identify six types of emotions: neutral, sad, happy, angry, disgusted, fear, to deliver meaningful emotion recognition feedback to users.

B. Data Preprocessing

Upon receiving continuous data at the software point, we identify emotion-inducing segments using timestamps and segment signals into 1-second windows. To effectively extract features from the data, we utilize different preprocessing methods on the different characteristics of physiological signal.

1) Heart Rate Signal: Over extended periods, data collection might experience trend drifts due to factors such as system bias or varying ambient light intensity. Such drifts might appear as a slow-changing trend in the photoplethysmogram signal, inconsistent with normal heart rate characteristics. To mitigate this, Detrended Fluctuation Analysis (DFA) is employed to filter out the trend component in the raw PPG signal, using a time window length of 20. Further signal processing utilizes a Butterworth bandpass filter to remove high-frequency noise and motion artifacts, with a pass-band cutoff frequency of 4 Hz and a stop-band cutoff frequency of 0.05 Hz. The signal is then normalized for consistency.

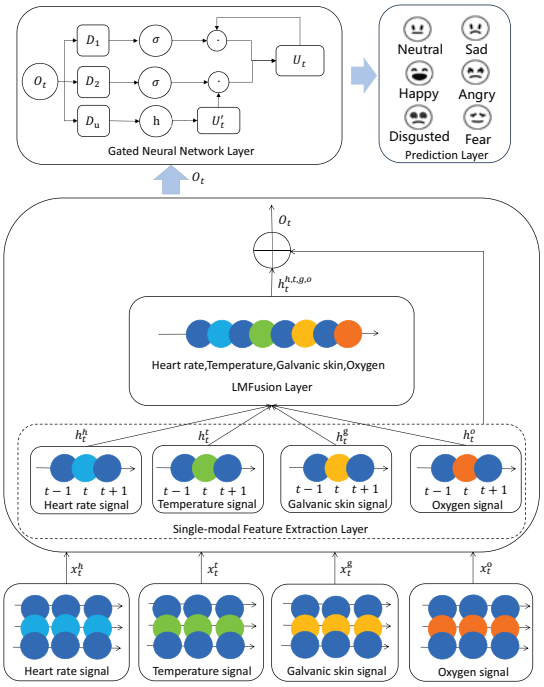


Fig. 1. LMFN network structure.

2) Blood Oxygen Signal: To minimize the data drift caused by skin surface factors such as stains and sweat, we apply a short-time Fourier transform on the collected red light signal and infrared light signal. Based on the signal distribution characteristics, we utilize a Butterworth bandpass filter to truncate frequencies below 0.5 Hz, with a pass-band cutoff frequency of 4 Hz and a stop-band cutoff frequency of 0.1 Hz. To eliminate outliers, soft normalization is applied using the 5th and 95th percentiles.

3) Galvanic Skin Signal: We regard the average galvanic skin signal under neutral emotional state as a reference and normalize the data from other emotional states accordingly, to mitigate skin differences among users. Considering that the primary frequency band of galvanic skin signals is below 0.2 Hz, we utilize a second-order Butterworth filter with a cutoff frequency of 0.3 Hz to minimize artifact interference.

4) Temperature Signal: To mitigate interference from external temperature on temperature signals, we employ a three-point moving average algorithm with weights (0.8, 0.1, 0.1). Furthermore, to achieve noise reduction and smoothing, we utilize a Savitzky-Golay filter with a window size of 99 and a polynomial fitting order of 1, which preserve the shape of data across varying time series effectively.

C. Multi-source Fusion

Existing multi-source fusion techniques often prioritize the fusion of the highest-level source information, leading to the neglect of the correlation between higher and lower layers. In order to better achieve the fusion of multi-source physiological signals, we design the LMFN. The LMFN consists of a single-source extraction layer and a multi-source fusion layer, whose structure is shown in Fig. 1.

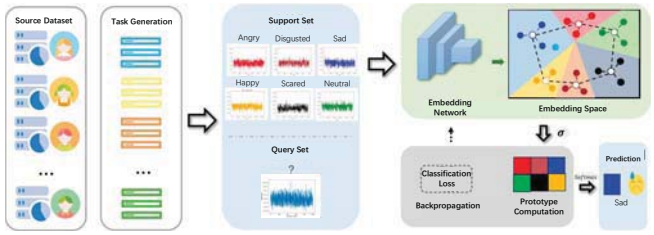


Fig. 2. Prototype network design.

1) single-source extraction layer: We take the feature input X of each mode at time t as the combination of C at previous time $t-1$ and X at current time t for the four basic physiological signals. In this way, the information will be gradually propagated and updated through time steps. The feature input X at time t is calculated by the Eq(1).

$$h_t^k = x_t^k + c_{t-1}^k \tag{1}$$

where $k = h, b, g, t$, represents the heart rate, blood oxygen, galvanic skin and temperature signal, respectively. x_t^k represents the feature input of physiological signal at time t. c_{t-1}^k represents the memory unit information at time $t-1$. Ultimately, we get the output value h_t^k, which integrates both the previous historical information and the current information.

2) multi-source fusion layer: In the multi-source fusion layer, we not only fuse the single-source signals, but also further integrate the amalgamation of multi-source signals. It is mainly because that the fusion of single-source signals may provide a partial information of the data, accompanied by inherent limitations such as missing information, noise interference, or data biases. By integrating information from multi-source signals, we can compensate for the shortcomings of each source and overcome the incompleteness of data, resulting in improved fusion quality and accuracy. Consequently, the output of multi-source fusion layer fuses both single-source and multi-source signals, which is calculated by Eq(2).

$$O_t = \sum_{k \epsilon N} h_t^k \tag{2}$$

where $N = h, b, g, t, (h, b, g, t)$, O_t is the fusion of single-source and multi-source signals. Furthermore, we make the O_t as input to the gated memory network.

D. Intra-subject Emotion Recognition

In a straightforward manner, we employ the output of the LMFN network to perform intra-subject emotion recognition. Through feature extraction within and between sources, we observe that the final recognition results are correlated with the outputs of each signal and the gating memory network. We can calculate the final recognition results by Eq(3).

$$P = D(h_t^{\mathrm{h}} + h_t^{\mathrm{t}} + h_t^{\mathrm{g}} + h_t^{\mathrm{o}} + u_t) \tag{3}$$

By passing through neural network $D()$, we obtain probability values P for 6 types of emotion, and the emotion category

result is determined based on the magnitude of the probability values. In addition, we apply cross-entropy as the loss function to continuously optimize the model.

E. Cross-subject Emotion Recognition

According to the temporal features of physiological signals, we design a few-shot cross-subject emotion recognition model inspired by the prototype network. This model not only achieves fast and accurate classification of few shots but also addresses the issues of over-fitting and limited information exchange between distant features that are common in the prototype network. The network design is shown in Fig. 2.

1) Task Generation: To create varied and authentic personal scenarios based on a given source dataset, we draw on the task generation strategy of MetaSense [21] to decompose the batch data into multiple sub-tasks, which are divided into support set and query set during each training process. The support set forms the prototypes, while the query set refines the position of the prototypes. In each sub-task, following the setting of n-way k-shot, N classes are selected from the source domain data, and k shots are chosen as the support set S from each selected class. For the support set S, a total of $N * k$ shots are selected. From the remaining shots of each class in the N classes, b shots are chosen as the query set Q, resulting in a total of $N * b$ shots. In this research, the shots come from a single user. Random sampling will be employed within the shots of user, ensuring an equal number of shots across all categories for this task. Specially, we generate 300 tasks for each user data, and select 2 tasks from each batch. For six basic emotions, the support set is constructed by the 6-way k-shot setting, with k shots selected for each class. The remaining shots for each class are chosen as the query set.

2) Embedded network design: In our design, the embedding network is optimized by the LMFN network structure, which can extract comprehensive and valuable information efficiently. Moreover, in order to prevent zigzagging gradient directions during back-propagation, we employ the LeakyReLU activation function as embedded network kernel. After processing each channel's physiological signal for multi-source fusion based on LSTM, the data is then fed into two merged convolution layers (Merged Conv1D1, Merged Conv1D2) to further explore the inter-modality correlations. Additionally, to mitigate over-fitting, the model is augmented with a global average pooling layer and a Dropout layer, facilitating propagation of effective information between modalities in the subsequent module for feature extraction. Furthermore, the Batch normalization is incorporated after each convolution layer to stabilize the network training and markedly reduce the computational cost of the model.

3) Training process: We employ the Euclidean distance as a metric to calculate the distance between query samples and prototypes of different classes, enabling us to determine the class to which a sample belongs. To address the influence of outliers on the class mean in few-shot data, we incorporate the Inverse Multiquadric (IMQ) function into the distance calculation, assigning weights to each shot point accordingly.

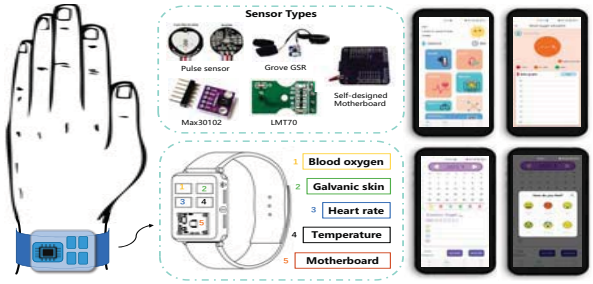


Fig. 3. The hardware and software of EmoTracer.

There is a severe problem in which minimal vectors possess non-zero gradients during the parameter update process during each training episode for few-shot tasks. Therefore, LazyAdam optimizer is adopted in this research, which permits a more tempered gradient update processing for sparse variables, alleviating the severe impact of over-fitting in few-shot models for physiological signals.

IV. Implementation And Experiment

A. Implementation

For hardware configuration, We design a wristband device for detecting physiological signals. The embedded sensors consist of photoplethysmography (PPG), blood oxygen saturation, galvanic skin response and temperature sensors, to collect relevant signals based on different sampling rates. We chose the STM32 as our microprocessor, specifically the F103C8T6 Cortex-M3 32-bit MCU from STMicroelectronics. This model supports a maximum clock frequency of 72MHz and provides a voltage of 3.3V to the microcontroller and other sensors. Additionally, we choose the RF-Star EFR32BG22A1 as Bluetooth module based on BLE 5.0, to transmit the packaged data frames to the software point for data processing. For the power supply chip, the selected option is the SY8089 DC-DC buck converter chip, which is characterized by ultra-low electrostatic current. It also incorporates a TP4054 charging circuit based on a Type-C interface and 500mA current supply, enabling power supply functionality.

For software configuration, our EmoTracer combines software modules for managing physiological signals, tracking mood records, performing emotion recognition, and storing personal information. It is integrated into a mobile app tested on the Huawei Mate30Pro smartphone, powered by the HUAWEI KIRIN990 processor, 4680 mAh battery, and 8 GB of RAM. Refer to Fig. 3 for a visual representation of the hardware and software system.

V. Evaluation

A. Experiment

We recruit 30 participants aged between 17 and 56 years from our campus to collect real emotional shots by watching video clips. Before the experiment, all participants agree to sign the Institutional Review Board (IRB) agreement [22], and are given a 1-minute buffer to regulate their emotions to ensure reaching a calm state. Aiming to acquire diverse

emotional inductions, we select audiovisual stimuli from multiple databases, including DEAP [23], FilmStim [24], IADS [25], SEED [26], and EMDB [27], complemented by internet-sourced clips reflecting a range of emotions such as entertainment, horror, and anger. There are 173 audiovisual materials utilized across the experiments, which include 4, 54, 30, 15, and 50 clips respectively from aforementioned databases and 20 online video clips as experimental stimuli materials. The experiment takes place in a typical office environment with noise levels ranging from 40 to 55 dB and lasts for 2 to 3 hours for each participants. Participants are equipped with our designed wristband device on their right hands, and sit quietly to view the audiovisual stimuli. Simultaneously, they activate the device and press the 'Start Test' button on main screen of app. As each clip concluded, participants cease the test, following which they rate their emotional experiences using the SAM scale from 1 to 9 for pleasure, arousal, and dominance and evaluate the familiarity and likability on a scale from 1 to 5 in their self-reports. Finally, we collect 11 GB of physiological data.

It should be mentioned that we encounter challenges with device instability that results in noisy data during the initial experiment. We exclude the affected participants and make necessary corrections to our hardware system. And some participants do not strictly adhere to the experimental procedure, such as engaging in large movement during the testing process or not watching the videos attentively, which could interfere with the collection of physiological signals. To ensure the validity of our experimental results, we conduct a further selection of participants, ultimately choosing 18 individuals for the emotion recognition experiment Therefore, there are 13,918 available shots collected throughout the entire study.

B. Comparison of Intra-subject Emotion Classification

To access the effectiveness of the LMFN network, we evaluate its performance in intra-subject emotion classification tasks alongside a LSTM network. Specifically, we collect emotion states and physiological data from single participant who is required to watch specific types of videos. Furthermore, we analyze between the collected data and the predicted results from the networks by calculating precision, recall, F1 score and accuracy, to evaluate their performance in identifying, classifying, and predicting emotions of the participants. Table I presents a comparison of the individual subject emotion recognition performance between the LMFN and LSTM network.

The results show that the LMFN network significantly outperforms the single LSTM network across all evaluation metrics, which validates the reliability of the intra-subject emotion recognition and the effectiveness of multi-source feature extraction based on LMFN network.

C. Comparison of Cross-Subject Emotion Classification

For comparison of cross-subject emotion classification, we conduct parallel experiments between the LSTM-based multi-source Fusion Cross-subject Emotion Recognition (LMFN-Cross) and the Prototype Network-based Cross-Subject Emotion Recognition (PN-Cross). Utilizing data from 18 subjects, a leave-one-subject-out cross-validation was performed for testing, to validate broad effectiveness of emotion recognition model. Specifically, 1 subject is treated as a new unseen data. This subject intentionally is excluded from both the training and validation process of the model, and only provide few shots during the testing. In addition, the remaining 17 subjects are used to train and validate the model. In PN-Cross, The n-way k-shot task in PN-Cross is set to 6-way 1-shot, with six emotion categories and 1 shot each category provided. Fig. 4 indicates that PN-Cross consistently surpasses LMFN-Cross across all participants in cross-subject emotion recognition. On average, LMFN-Cross achieves an accuracy of 39.7%, while PN-Cross achieves an average accuracy of 72.1%, resulting in an average improvement of 33.4%. These findings demonstrate that adopting a few-shot learning and utilizing the prototype network can effectively enhance the performance of cross-subject emotion recognition tasks.

TABLE I
THE RESULT OF INTRA-SUBJECT EMOTION RECOGNITION

Performance \ method	LSTM	LMFN
Precision	67.3%	88.8%
Recall	66.6%	88.2%
F1-Score	66.9%	88.5%
Accuracy	68.3%	**95.6%**

TABLE II
THE RESULT OF CROSS-SUBJECT EMOTION RECOGNITION.

Performance \ method	DANN-Cross	MN-Cross	PN-Cross
Precision	66.2%	68.5%	74.3%
Recall	65.7%	67.6%	70.6%
F1-Score	63.2%	68.6%	71.8%
Accuracy	65.6%	69.1%	**72.1%**

In addition, this research also encompasses a comparison of diverse cross-domain models with our designed model. Within this comparison, DANN-Cross refers to the classic model on physiological signals for solving cross-subject problems in emotion recognition tasks, and MN-Cross represents the matching network algorithm in metric learning for few-shot learning. From the Table II, it can be demonstrated that the PN-Cross model achieves the best performance in emotion recognition compared to the other models.

D. Comparison of Few-shot Support Set

We evaluate the impact of the number of shots per class within the support set on cross-subject emotion recognition performance utilizing a prototype network. For each of the six emotions, we conduct experiments with 1, 3, 5, and 10 shots as the support set. Each shot corresponds to a fixed time segment, which is divided based on the user-provided physiological signal detection time. The emotion recognition accuracy of all participants for different shot numbers is shown in Fig. 5. As the number of shots in the support set increases, the performance of emotion recognition gradually stabilizes.

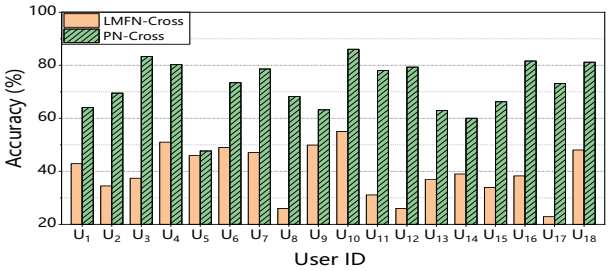


Fig. 4. Accuracies of different classification methods.

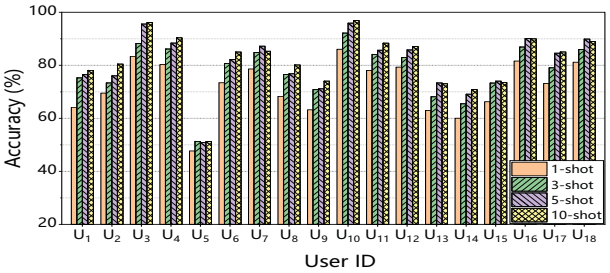


Fig. 5. Accuracies of different numbers of shots.

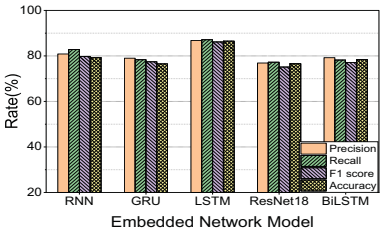


Fig. 6. Performance of different network.

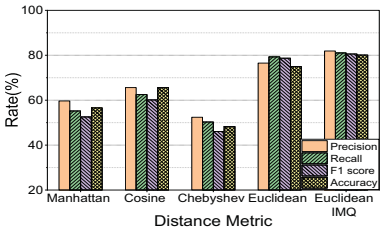


Fig. 7. Performance of different distance metrics.

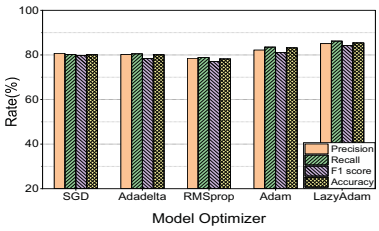
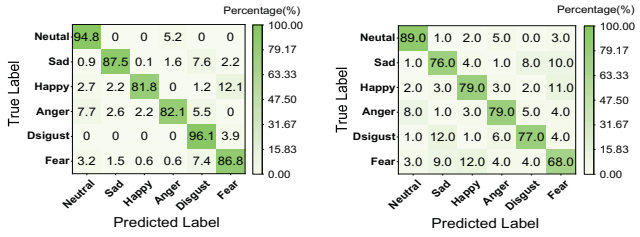


Fig. 8. Performance of different model optimizers.



(a) intra-subject confusion. (b) Cross-subject confusion.

Fig. 9. The confusion metrics of six emotion types in intra-subject emotion recognition(a) and cross-subject emotion recognition(b) respectively.

Considering the limitation of the few-shot physiological signals in our task and the essential goal of stable recognition performance for new users, we choose to utilize $k = 5$ shots for cross-subject emotion recognition in our research.

E. Comparison of Embedded Network Modules

In the design of a few-shot cross-subject emotion recognition system, an appropriate selection of embedded network has a significant impact on the recognition performance. Hence, an evaluative experiment is conducted to compare various sequential models, namely LSTM, RNN, GRU, BiLSTM, and ResNet18. For this experiment, we employ the leave-one-subject-out method to conduct the testing. The n-way k-shot task will be set as a 6-way 5-shot. From the Fig. 6, the result can be observed that LSTM outperforms RNN by 7.1% and GRU by 10% in accuracy. Additionally, in terms of precision, LSTM outperforms RNN by 5.9% and GRU by 7.8%. The superior performance of LSTM can be credited to its design, featuring memory cells and gate mechanisms, effectively learning the temporal features of input sequences and stabilizing gradient propagation over time. In contrast, traditional RNNs often encounter the vanishing gradient problem, hindering their ability to effectively learn long sequences. BiLSTM utilizes two independent LSTM modules to more comprehensively capture sequence context, but it often requires more computational resources and time, which makes it less favorable for this specific few-shot emotion recognition task. Therefore, LSTM demonstrates better emotion recognition performance as the embedded network, and we select it as the foundational framework for the embedded network.

F. Comparison of Distance Metric

In order to determine the optimal distance metric for the prototype network, we conduct the cross-subject emotion recognition experiment by utilizing various distance metrics. For this experiment, we employ the consistent cross-validation method and class-shot settings like comparison of embedded network modules. As shown in Fig. 7, the distance metrics include Manhattan, Cosine, Chebyshev, Euclidean and Euclidean IMQ distance. It can be observed that the Euclidean distance achieves an accuracy of 75.1%, which is 9.5% higher than Cosine distance, 26.8% higher than Chebyshev distance, and 18.5% higher than Manhattan distance. Furthermore, when the IMQ is applied to Euclidean distance, the recognition accuracy improves to 80.1%. In conclusion, we select Euclidean distance with IMQ kernel as the distance metric.

G. Comparison of Model Optimizer

We conduct a comparative analysis of the performance of five optimizer categories: SGD, Adadelta, RMSprop, Adam, and LazyAdam, to assess the influence of different optimizers on the few-shot emotional model. For this experiment, we employ the consistent cross-validation method and class-shot settings like comparison of embedded network modules. As shown in Fig. 8, the emotion recognition accuracy for SGD, Adadelta, RMSprop, Adam, and LazyAdam are 80.1%, 80.1%, 78.2%, 83.2%, and 85.4%, respectively. Among these, LazyAdam performs the best in the overall emotion recognition task. This superior performance is attributed to LazyAdam's exclusive storage of the diagonal elements of the first-order and second-order matrices for each parameter, which effectively diminishes system memory consumption. Consequently, the LazyAdam optimizer emerges as the most apt choice for the emotion recognition task.

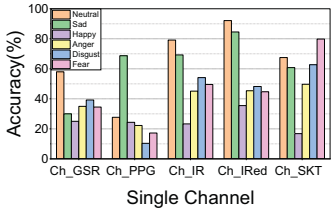


Fig. 10. Performance of different emotions for single channel

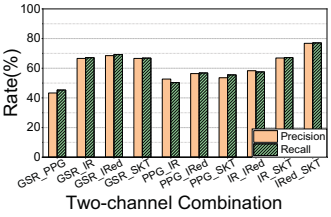


Fig. 11. Performance of two-channel combination.

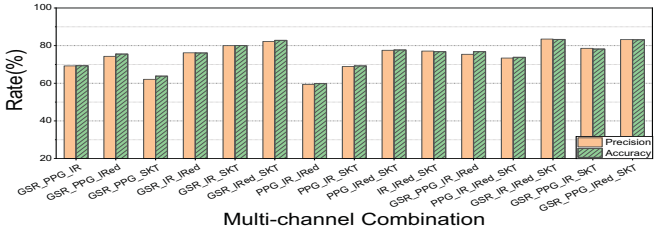


Fig. 12. Performance of multi-channel combination.

H. Confusion of Emotion Types

In an endeavor to discern which emotions are predominantly misclassified in emotion classification tasks, we conducts separately six types of emotion confusion statistics on both the intra-subject and the cross-subject emotion recognition model. We conduct the assessment under the same experimental conditions. As depicted in Fig. 9, there is a notable overlap in classifications between Sad, Disgust and Anger. This confusion is possibly due to the high arousal levels associated with these negative emotions and the overlapping segments in the stimuli. Additionally, the high degree of confusion between Happy and Fear may be due to the fact that both emotions are associated with high levels of arousal.

I. Performance and Combinations of Different Channels.

We evaluate the emotion recognition performance of single channel and multi-channel combinations. The channels are as follows: Ch_GSR is the Galvanic Skin Response sensor, Ch_PPG is the Photoplethysmography sensor, Ch_IR is the Infrared channel of the Blood Oxygen Saturation sensor, Ch_IRed is the red light channel of the SpO_2 sensor, and Ch_SKT is the Skin Temperature sensor. GSR_PPG represents a combination of the Ch_GSR and Ch_PPG channels.

Fig. 10 shows the results of cross-subject emotion classification task with single-channel, revealing diverse expressive capacities for various emotions across different channels. Notably, for neutral and sad emotions, Ch_IRed exhibits robust recognition performance with accuracies of 92.1% and 84.6% respectively. The Ch_SKT channel stands out in recognizing fear and disgust, achieving the highest accuracies of 79.8% and 62.7%, which substantiates the strong correlation between skin temperature and negative emotions. Conversely, for happy emotion, all channels generally have poor recognition performance. This might be attributed to the infrequent induction of happiness during the experiments and the potential overlap in physiological markers with sadness and disgust.

Fig. 11 presents the enhanced two-channel emotion recognition performance, which is significantly improved compared to using single channel. Specifically, the combination of IRed and SKT channels achieves precision, recall, F1-score, and accuracy of 76.8%, 77.1%, 72.5%, and 77.4%, respectively. The union of GSR and SKT channels also enhances the emotion recognition accuracy to 67.9%, marking 27.9% improvement over the solitary use of the GSR channel. These results suggest that the GSR signal captures emotional arousal and intensity, while the SKT signal reflects physiological changes and the persistence of emotions. By combining these two signals, a more comprehensive understanding of various aspects of emotions can be obtained.

Fig. 12 displays the outcomes of emotion classification using both three-channel and four-channel combinations. Multi-channel combinations demonstrate superior emotion recognition performance compared to single channels. This enhancement is attributed to multi-channel combinations compensating for the limitations inherent in the sparse information of single channels. It further validates the effectiveness of the multi-source fusion strategy designed in our research.

TABLE III
THE PERFORMANCE OF RELATED EMOTION RECOGNITION WORKS

Method	Stimuli	Testers	Emotion	Signal	Accuracy
[19]	DEAP	32	HVHA, LVHA LVLA, HVLA	EEG, PHY	intra-subject 93.1%
[28]	CASE MERCA CEAP-360VR	15	1D-2C, 2D-5C	EDA, SKT, BVP	intra-subject 76.62%
[29]	SEED	15	neutral, happy sad	EEG	intra-subject 95.44% cross-subject 86.30%
[30]	SEED-IV SEED	15	neutral, happy sad, fear	EEG	cross-subject 73.92% corss-subject 91.65%
[10]	EMBD	18	neutral, cheer, sad erotic, horror	EDA, HR IBI, SKT	cross-subject 65.6%
Ours	Deap, FilmStim, IADS SEED, EMDB	18	angry, disgusted, sad happy, scared, neutral	HR, GSR SKT, blood oxygen	intra-subject 95.6% cross-subject 80.7%

J. Comparison of Relevant Work

In relevant work, our EmoTracer exhibits remarkable capabilities in emotion recognition, as demonstrated by the results presented in Table III.

For data fusion, unlike previous approaches [19] and [28] that concatenate data, our LMFN method integrates single-source and multi-source signals to better represent features, focusing on long-term dependencies in temporal data. For few-shot learning, [28] employs Siamese network that embeds distances within emotional label samples solely to achieve fine-grained emotion recognition, so it does not emphasize cross-subject test in experiment. Addressing the cross-domain challenge, [29] introduces a multi-source domain adaptation method that aligns the feature distributions of source and target domains with Maximum Mean Discrepancy loss; on the other hand, [30] leverages instance-adaptive graphs for improved generalization across users and sessions. Notably, these methods, primarily focus on EEG signal acquisition in experimental settings, overlooking the practical challenges of adapting to few shots in real-world scenarios. Furthermore, [10] presents a stacked model-based hybrid sensor fusion method for peripheral physiological signals, facilitating user-independent emotion recognition but with suboptimal accuracy. In summary, the innovations of our study in data fusion and few-shot learning, along with its attention to cross-subject adaptability, offer new perspectives and solutions for the field of emotion recognition based on physiological signals.

VI. Conclusion

In this study, we delve deeply into the relationship between emotional fluctuations and physiological indicators, propose a wearable emotion recognition system called EmoTracer , which can capture six basic emotions in intra-subject and cross-subject emotion recognition with high precision. Coupled with user-friendly design, our system stands out as a powerful tool for both clinical and personal use. However, our research focuses solely on the six basic emotions recognition with a sample size of 18 participants, and we acknowledge the complexity, as well as the significant individual difference. The limited range of emotional categories and sample size may not reflect the generalizability of the research findings and the statistical power, increasing the likelihood of drawing incorrect conclusions. Therefore, we aim to explore additional emotional states and participants, offering more rigorous scientific support for the monitoring in the future. Future work may also focus on adding interpretability to the model, an enhancement that could aid healthcare providers and users in better comprehending emotional changes.

VII. Acknowledgement

This work was supported in part by China NSFC (62172286, U2001207) ; in part by Guangdong NSF (2022A1515011509); in part by the Guangdong Provincial Key Laboratory of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (2023B1212010007); and in part by the Tencent"Rhinoceros Birds"—Scientific Research Fund for Young Researchers of Shenzhen University. Yongpan Zou is the corresponding author.

References

[1] B. X. Tran, G. H. Ha, L. H. Nguyen, G. T. Vu, M. T. Hoang, H. T. Le, C. A. Latkin, C. S. Ho, and R. C. Ho, "Studies of novel coronavirus disease 19 (covid-19) pandemic: a global analysis of literature," *International journal of environmental research and public health*, vol. 17, no. 11, p. 4095, 2020.

[2] A. Zekja, J. Kruja, and K. Gusha, "Epidemiological profile of patients with depression in shkoder, albania," *Psychol Behav Sci Int J*, vol. 18, pp. 1–5, 2021.

[3] S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. Essa, "Eyemotion: Classifying facial expressions in vr using eye-tracking cameras," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1626–1635.

[4] A. V. Atanassov, D. I. Pilev, F. N. Tomova, and V. D. Kuzmanova, "Hybrid system for emotion recognition based on facial expressions and body gesture recognition," in *2021 International Conference Automatics and Informatics (ICAI)*. IEEE, 2021, pp. 135–140.

[5] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

[6] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th international colloquium on signal processing and its applications*. IEEE, 2011, pp. 410–415.

[7] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2020.

[8] S. Wu, X. Xu, L. Shu, and B. Hu, "Estimation of valence of emotion using two frontal eeg channels," in *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2017, pp. 1127–1130.

[9] W. Sato, K. Murata, Y. Uraoka, K. Shibata, S. Yoshikawa, and M. Furuta, "Emotional valence sensing using a wearable facial emg device," *Scientific Reports*, vol. 11, no. 1, p. 5757, 2021.

[10] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," *IEEE sensors Journal*, vol. 19, no. 19, pp. 8402–8412, 2018.

[11] K. Hugdahl, *Psychophysiology: The mind-body perspective*. Harvard University Press, 1995.

[12] J. W. L. Doust and R. A. Schneider, "Studies on the physiology of awareness: the effect of rhythmic sensory bombardment on emotions, blood oxygen saturation and the levels of consciousness," *Journal of mental science*, vol. 98, no. 413, pp. 640–653, 1952.

[13] J.-C. Roy, W. Boucsein, D. C. Fowles, and J. Gruzelier, *Progress in electrodermal research*. Springer Science & Business Media, 2012, vol. 249.

[14] V. Kosonogov, L. De Zorzi, J. Honore, E. S. Martínez-Velázquez, J.-L. Nandrino, J. M. Martinez-Selva, and H. Sequeira, "Facial thermal variations: A new marker of emotional arousal," *PloS one*, vol. 12, no. 9, p. e0183592, 2017.

[15] N. Mehendale, "Facial emotion recognition using convolutional neural networks (ferc)," *SN Applied Sciences*, vol. 2, no. 3, p. 446, 2020.

[16] C. Mingming, F. Jiandong, and Z. Yudong, "Emotion recognition of human body's posture in open environment," in *2020 Chinese Control And Decision Conference (CCDC)*. Ieee, 2020, pp. 3294–3299.

[17] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6474–6478.

[18] P. D. W. Z. P. Z. Yixiang Dai, Xue Wang, "Optimizing emotion recognition with stacked autoencoder for wearable multi-modal sensor networks," *Chinese Journal Of Computers*, 2017 (in Chinese).

[19] Y. Han and Y. Xu, "The research of emotion recognition based on multi-source physiological signals with data fusion," in *ITM Web of Conferences*, vol. 45. EDP Sciences, 2022, p. 01038.

[20] M. N. Dar, M. U. Akram, S. G. Khawaja, and A. N. Pujari, "Cnn and lstm-based emotion charting using physiological signals," *Sensors*, vol. 20, no. 16, p. 4551, 2020.

[21] T. Gong, Y. Kim, J. Shin, and S.-J. Lee, "Metasense: few-shot adaptation to untrained conditions in deep mobile sensing," in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 110–123.

[22] C. Grady, "Institutional review boards: Purpose and challenges," *Chest*, vol. 148, no. 5, pp. 1148–1155, 2015.

[23] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[24] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition and emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[25] M. Bradley and P. Lang, "The international affective digitized sounds (2-nd edition; iads-2): Affective ratings of sounds and instruction manual gainesville," *The Center for Research in Psychophysiology: Gainesville, FL, USA*, 2007.

[26] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

[27] S. Carvalho, J. Leite, S. Galdo-Álvarez, and O. F. Gonçalves, "The emotional movie database (emdb): A self-report and psychophysiological study," *Applied psychophysiology and biofeedback*, vol. 37, pp. 279–294, 2012.

[28] T. Zhang, A. El Ali, A. Hanjalic, and P. Cesar, "Few-shot learning for fine-grained emotion recognition using physiological signals," *IEEE Transactions on Multimedia*, 2022.

[29] T. Song, S. Liu, W. Zheng, Y. Zong, and Z. Cui, "Instance-adaptive graph for eeg emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2701–2708.

[30] W. Guo, G. Xu, and Y. Wang, "Multi-source domain adaptation with spatio-temporal feature extractor for eeg emotion recognition," *Biomedical Signal Processing and Control*, vol. 84, p. 104998, 2023.