A Real-Time Emotion-Aware System Based on Wireless Body Area Network for IoMT Applications

Chang Li, Yingchi Mao, Qian Huang *Member, IEEE*, Weiliang Xie, Xiaoming He *Member, IEEE*, Jie Wu *Fellow, IEEE*

***Abstract*—The Internet of Medical Things (IoMT) stimulates the development of intelligent medical applications. As mental disorders become a global problem, emotion recognition has received widespread attention, as it can contribute to more comprehensive mental health monitoring and psychological assessment. Physiological signal-based emotion-aware monitoring is a particularly promising application due to its non-invasive and objective data collection. Recently, multi-modal emotion recognition has been enhanced with Wireless Body Area Network (WBAN) access to IoMT, where wireless medical sensors are interconnected and abundant signals are acquired conveniently. However, how to synthesize these multi-source physiological signals to facilitate emotion recognition is a challenging problem due to their heterogeneity and interference. To solve this problem, we propose a real-time Differential Multi-modal Transformer (Diff-MT), where the main components are the Differential Hyper-information Extraction (DHE) module, the Multi-modal Global Cross-attention Encoder (MGCE) and the Difference-augmented Feature Fusion (DFF). Ultimately, we endow the system with emotional awareness and distribute the state to IoMT devices. Extensive experiments demonstrate that the proposed Diff-MT exhibits superior performance compared to existing methods on the WESAD and DEAP datasets and is appropriate for IoMT-based healthcare.**

***Index Terms*—Internet of Things, Emotion recognition, Physiological signal analysis, Deep learning**

I. Introduction

THE Internet of Things (IoT) is the primary promoter of intelligent applications, integrating holistic sensing [18], reliable transmission [26], and intelligent processing [13]. As the embodiment of IoT in healthcare, the Internet of Medical Things (IoMT) facilitates many intelligent medical applications, such as humanoid guide robots, auxiliary diagnosis, and intelligent monitoring [41]. Mental health problems are one aspect of health that is often overlooked, such as depression and anorexia, affecting millions of people around the world [42]. Therefore, emotion perception is vital for IoMT-based healthcare applications. Sentiment monitoring helps prevent, diagnose, and cure mental disorders. Additionally, emotional feedback assists with the creation of personalized services and more friendly human-computer interaction.

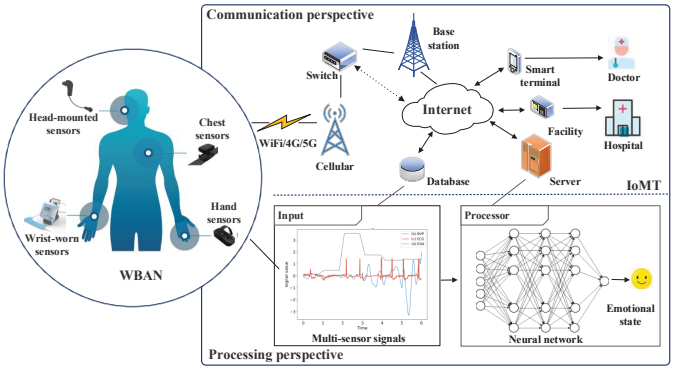


Fig. 1. The schema of the proposed IoMT-based emotion-aware system. Sensors connected to the WBAN collect a variety of physiological signals and transmit them into the database of the IoMT through wireless communication technology. By deploying the proposed Diff-MT model on the server, these signals will be processed and analyzed, and the deduced emotional states will be delivered to the terminal devices to assist intelligent medical applications.

Human emotion is a comprehensive cue that manifests our consciousness, behavior, and health. Human emotions can be manifested through both physiological and non-physiological signals. Non-physiological signals refer to facial expressions, voice tone, body posture, etc. Compared with these behavioral hints, which may be disguised, physiological signals such as Electroencephalogram (EEG), Electrocardiogram (ECG), Respiration (RESP), Blood Volume Pulse (BVP), and Electrodermal Activity (EDA) are more objective and reliable. In addition, the physiological signal has the natural advantage of low data volume, which is more suitable for sustainable and efficient intelligent monitoring. Therefore, emotion recognition based on physiological signals has received extensive attention and is seen as a promising route.

Most existing methods utilize a single physiological signal for emotion recognition [5], [8], [20], [35], rendering them

Chang Li, Yingchi Mao, Qian Huang, and Weiliang Xie is with the College of Computer Science and Software Engineering, Hohai University, Nanjing, Jiangsu, China, 211100. (E-mail: {lichang, yingchimao, huangqian, xieweiliang}@hhu.edu.cn)

Xiaoming He is with the college of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China, 210003. (E-mail: hexiaoming@njupt.edu.cn)

Jie Wu is with the Department of Computer and Information Sciences, Temple University, SERC 362, 1925 N. 12th Street, Philadelphia, PA 19122. (E-mail: jiewu@temple.edu)

This research was funded by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under grant KYCX23_0753, the Fundamental Research Funds for the Central Universities under grant B230205027, the Key Research and Development Program of China under grant 2022YFC3005401, the Key Research and Development Program of China, Yunnan Province under grant 202203AA080009, the 14th Five-Year Plan for Educational Science of Jiangsu Province under grant D/2021/01/39, and the Jiangsu Higher Education Reform Research Project under grant 2021JSJG143.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3458976

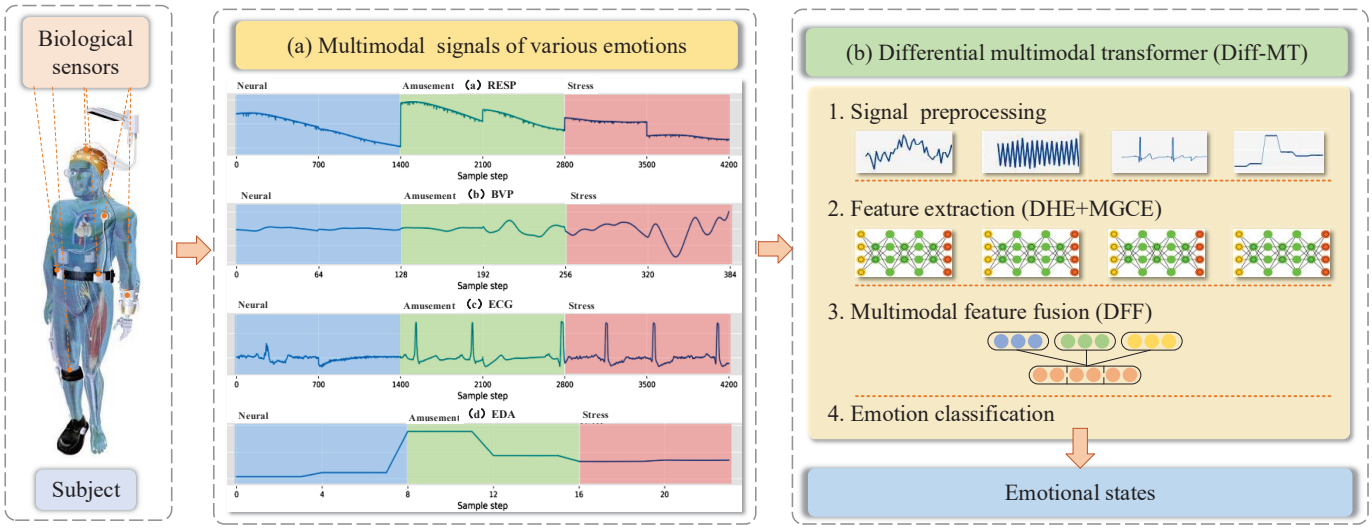


Fig. 2. The flowchart of the proposed Diff-MT. It comprises DHE, MGCE, and DFF modules for signal processing, feature extraction, and multi-modal feature fusion, respectively. It acquires multi-modal physiological signals via biological sensors connected by WBAN. After emotion classification, Diff-MT delivers human affective states to cloud services, contributing to healthcare applications.

one-sided and limited scenarios. Mental disorders are always accompanied by abnormalities of multiple physiological indicators, so leveraging multiple sensors is a more comprehensive and robust strategy. Furthermore, by employing a single signal, we cannot exclude external interference. With the Wireless Body Area Network (WBAN) connected to the IoMT ecosystem, obtaining multiple signals by connected wireless sensors in/around the human body is more convenient. However, an efficient emotion-aware system oriented to IoMT applications based on these WBANs has yet to be fully developed.

For the purpose of designing such an application, we investigate emotion-aware schema from multiple sensor signals in WBAN, as depicted in Fig. 1. We regard signals collected from various sensors as multi-modal data to transform the problem into multi-modal signal analysis. That is, given multiple sensor signals, the corresponding emotion categories need to be predicted and output by the system. With the accumulation of signals, developing a neural network to process and analyze these data from WBAN is reasonable. However, due to the heterogeneity of multi-modal physiological signals, this work mainly encounters the following challenges.

First, time synchronization and feature alignment are intractable. Different medical sensors have varying sampling frequencies and channels, resulting in different lengths and distributions of acquired signals even during the same interval. Taking the WESAD [31] dataset as an example, the sampling frequency for the RESP signal is 700Hz, whereas for the EDA signal, it is only 4Hz. Second, to limit the effects of biased or irrelevant information in feature fusion, it is necessary to mine latent and long-term dependencies across modalities. However, there are significant differences in how these signals respond to emotional changes. For instance, when experiencing stress, the BVP signal exhibits pronounced fluctuations (see Fig. 2(a)). In contrast, the fluctuations in the EDA signal are subtle in the stress state but are evident during the amusement state.

To surmount the above issues, we introduce a deep learning framework named Differential Multi-modal Transformer (Diff-MT) to process multi-sensor physiological signals for emotion recognition. As shown in Fig. 2(b), it processes signals through a Differential Hyper-information Extraction (DHE) module, extracting features using a Multi-modal Global Cross-attention Encoder (MGCE) module, fusing multi-modal features through a Difference-augmented Feature Fusion (DFF) strategy. After emotion classification, Diff-MT can deliver human emotional states and assist in intelligent healthcare applications and cloud services. Overall, the contributions of our work are summarized as follows.

- We propose an emotion-aware scheme based on IoMT to supplement human-centered medical applications. At the communication level, we leverage WBAN to achieve radio efficient data transmission between medical sensors and other devices in IoMT. At the processing level, we realize multi-modal signal feature fusion and sentiment analysis.
- We introduce a novel deep learning framework, Diff-MT, for multi-modal physiological signal processing, feature extraction, feature fusion, and emotion recognition. DiFF-MT mainly includes DHE, MGCE, and DFF modules and is compatible with widespread signal-based approaches.
- Our method achieves promising accuracy and real-time performance on two public datasets, WESAD and DEAP. The experimental results demonstrate that it is appropriate for IoMT-based services.

The rest of this article is organized as follows. Section II reviews the emotion-aware technology based on physiological signals. Section III introduces the implementation of Diff-MT. Section IV reports experimental results, and Section V analyzes the effectiveness of components contained in Diff-MT. Then, we discuss the applications and limitations in Section VI. Finally, we summarize this work in Section VII.

II. Related Work

A. Emotion Recognition Based on Physiological Signals

Compared with non-physiological signals, such as facial expressions, voice tone, and body posture, which are emotional manifestations that may be disguised, physiological signals have the advantages of objectivity and reliability. Therefore, emotion recognition methods based on physiological signals have received widespread attention, especially in fields such as healthcare that require objective emotional evaluations. Physiological signals are mainly collected by non-intrusive or intrusive sensors. Wearable sensors, which are non-intrusive, are widely used in healthcare applications because of their comfort and portability. The collected physiological signals like EEG, ECG, RESP, EDA, BVP, Electrooculogram (EOG), Electromyogram (EMG), Galvanic Skin Response (GSR), temperature, photoplethysmogram, and heart rate variability are commonly utilized for emotion analysis due to their objectivity and low-power dissipation. In the early stage, emotion analysis methods mainly adopted a single modality. Choi et al. [5] constructed the deep learning model Attention-LRCN to extract temporal features and reduce the effect of noises in photoplethysmogram signals through an attention module. Gao et al. [20] proposed an EEG-based method using coincidence filtering and simulated the information extraction pattern of artificial-features-based methods to design Convolutional Neural Networks (CNNs). However, relying on single-modal indicators to reflect sentiment may result in the problem of one-sidedness and contingency. In addition, once this signal is externally disturbed, it will greatly mislead the judgment of the emotion recognition system. Considering the complementarity among various signals, we are motivated to employ multi-modal signals to provide a more comprehensive and precise scheme.

B. Multi-modal Emotion Recognition

With the development of sensor devices, a variety of physiological signals can be collected conveniently. Afterward, multi-modal physiological datasets of human affective are available, such as DEAP [16] and WESAD [31], allowing emotion recognition based on multi-modal sensor signals to flourish. Souza et al. [7] proposed a novel pipeline for identifying stress sequences through Recurrent Neural Network (RNN) with Gaussian noise layers. The five modalities, including ECG, EDA, RESP, EMG and temperature, are fed into the pipeline. This paradigm of early fusion ignores the cross-modal information correlation between signals. To employ the complementarity of multi-modal information, Rashid et al. [29] proposed a late fusion method, namely, SELF-CARE. They determined the noise context using EMG or motion acceleration of a subject and performed Kalman filter-based late fusion methods after feature extraction for classification. To enhance multi-modal learning, many researchers have attempted to mid-level feature fusion. Bhatti et al. [3] presented an attentive cross-modal connection between CNNs to share intermediate representations among ECG and EDA signals. However, these practices still have limitations in effectively capturing the high-level correlation between multi-modal features, which decreases their performance. In addition, such asynchronous signal processing and redundant neural networks aggravate the computational and time consumption of previous methods, which are challenging to meet real applications. In this work, we design a lightweight deep learning framework

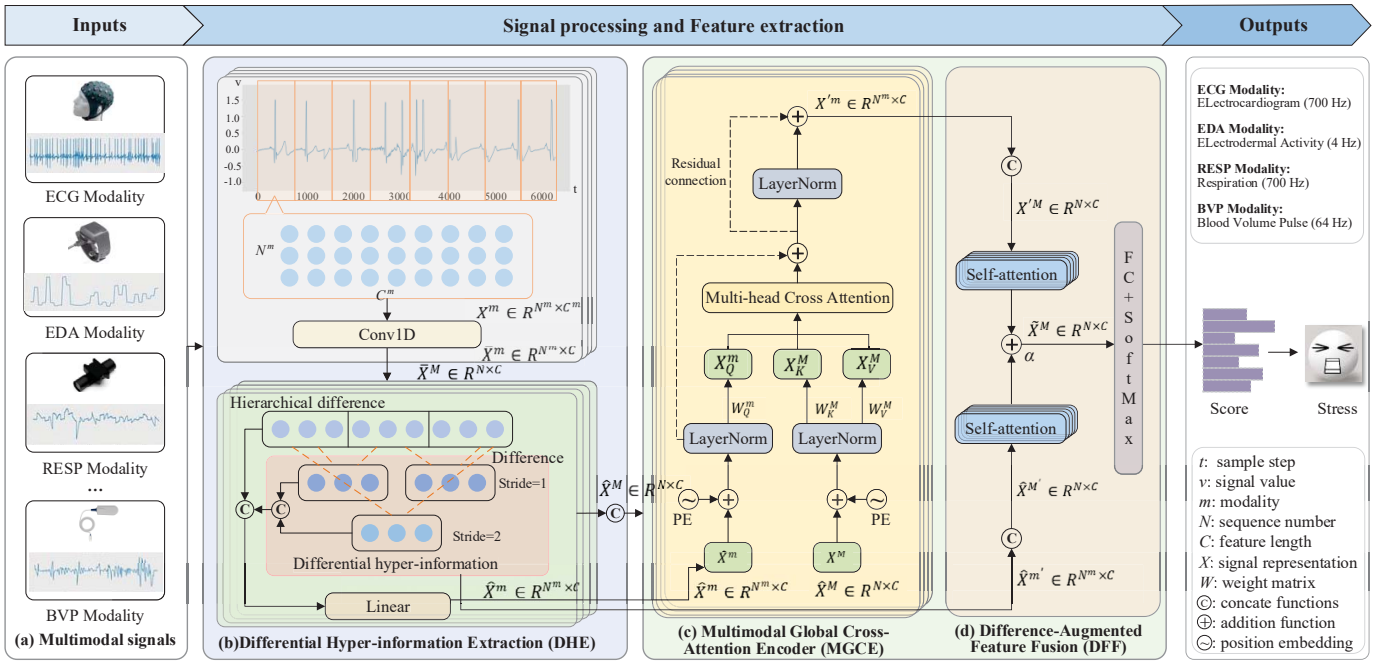


Fig. 3. The framework of the proposed Diff-MT. Diff-MT takes multi-modal physiological signals (a) as the input. For instance, the inputs are ECG, EDA, RESP, and BVP signals on the WESAD dataset. In Diff-MT, each modality is encoded simultaneously using a multi-branch structure that contains embedded DHE (b) and MGCE (c) modules. Next, representations of the multi-modal signals are fused through the DFF module (d). Finally, fully connected layers and Softmax are employed for emotion recognition.

and utilize wireless body area networks to efficiently process multiple signals and secure data transmission, making it suitable for IoMT-based intelligent healthcare applications.

III. Methods

In this section, we present the construction of Diff-MT depicted in Fig. 3. Diff-MT is a multi-branch neural network embedded with DHE, MGCE, and DFF modules. Among them, the DHE module aligns the multi-modal physiological signals and generates differential hyper-information hierarchically. The MGCE module identifies the potential relationship between each modality and global modality features. The DFF module is responsible for fusing high-level multi-modal features and feeding them to the classifier for emotion recognition.

A. Differential Hyper-Information Extraction

Multi-sensor signal alignment. The physiological signals collected by multiple sensors are heterogeneous due to the varying settings of signal sources, such as sampling frequency and signal length. As a result, the received signals are asynchronous, have different lengths, and contain different data channels. Thus, a preprocessing scheme is necessary before signal analysis. We propose the DHE module to align these multi-modal signals and generate differential hyper-information hierarchically, as depicted in Fig. 3(b). The multi-modal signal alignment process consists of three steps. To achieve time synchronization, we segment multi-source signals into samples with a duration of 1s according to the sampling frequencies. After that, samples record the emotional state of the same duration, facilitating synchronous transmission and second-level response. To achieve channel alignment, we employ 1D convolutions to project various signals into a feature space of the same dimension in an end-to-end manner. Then, the aligned multi-modal features can be conveniently cascaded to obtain preliminary multi-modal fusion features. Specifically, given a sample $X^m \in R^{N^m \times C^m}$, where N^m represents the sampling steps of the corresponding sensor m and C^m is the number of channels, the corresponding preprocessed sequence $\bar{X}^m$ can be obtained by the following equation

$$\bar{X}^m = \mathrm{Conv1D}\left(K, X^m\right), \bar{X}^m \in R^{N^m \times C} \tag{1}$$

where C is the number of channels after feature alignment. Conv1D is a temporal convolution with kernel size K. On this basis, a multi-modal sequence $\bar{X}^M \in R^{N \times C}$ can be obtained by cascading. To illustrate the technical details of multi-modal signal alignment, we provide an example in Fig. 4. Finally, we adaptively realize multi-modal signal alignment, laying the foundation for mining the cross-modal complementarity and real-time emotional feedback.

Hierarchical sequence differencing. Physiological signals are prone to unintended external interference or noise from sensors, compromising accuracy and reliability. To emphasize valid data and fully exploit the dynamic patterns of physiological signals, we extract differential hyper-information, as illustrated in Fig. 3(b). We first conduct a sliding window operation on the aligned features to yield clipped features. Assuming that the H^0 layer has n windows, and the window size w is C/n. Then, spaced window sampling and differencing are performed hierarchically according to predefined strides. The differences obtained by various strides describe the multi-range dynamics of emotional signals in different temporal spans. Integrating these differences leads to more discriminative sentiment patterns, i.e., hierarchical hyper-information

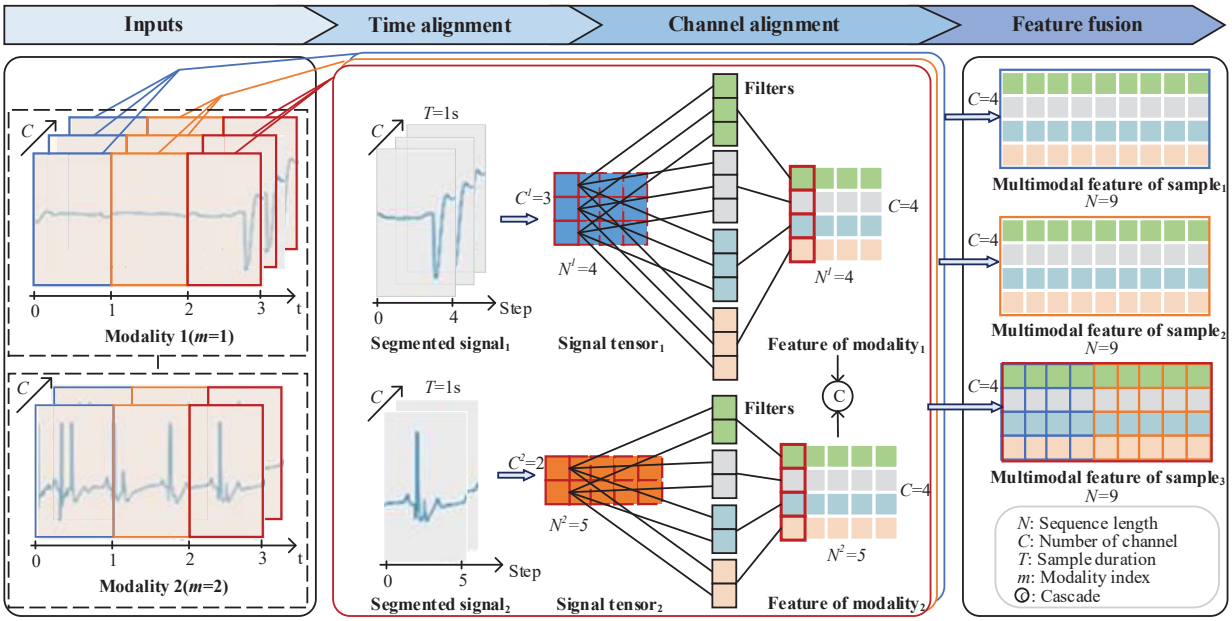


Fig. 4. An example of multi-modal signal alignment. We illustrate the alignment process for two different modal signals in three steps: temporal alignment, channel alignment, and feature fusion. The input is two signal samples of duration 3s, where $X^1 \in R^{4\times3}$and $X^2 \in R^{5\times2}$. After segmentation and alignment, it transforms into three multi-modal samples with the same shape $\bar{X}^M \in R^{9\times4}$.

$\hat{X}^{m'}$. Mathematically, given the DHE module with H layers, the differential hyper-information can be obtained by Eq. (2).

$$\hat{X}^{m'} = \text{Concat}_{h=1}^{H}\left(\text{Concat}_{i=0}^{n-\mathcal{S}(h)}\left(\bar{X}_{h,i+\mathcal{S}(h)}^{m} - \bar{X}_{h,i}^{m}\right)\right) \quad (2)$$

where the differencing stride set $\mathcal{S} \subseteq \{1, 2, \ldots, h\}$, and $|\mathcal{S}| \leqslant h \leqslant n$. $\mathcal{S}(h)$ is the differencing step of the h_{th} layer. $\hat{X}_{h,i}^{m} \in R^{N^m \times w}$ is the representation of the i_{th} window in the h_{th} layer. We employ the cascade function Concat$(\cdot)$ to combine the differences of each layer and integrate them to obtain the hierarchical hyper-information, which effectively reflects the global fluctuation of the physiological signals and reveals their temporal dynamics in a refined manner. Moreover, it can reduce the impact of perturbations and increase the system's robustness. By default, H is set to 3, and sliding windows are non-overlapping. Furthermore, we fuse the hyper-information with original signals to avoid any loss of information as follows.

$$\hat{X}^{m} = \varphi\left(\text{Concat}\left(\hat{X}^{m'}, \bar{X}^{m}\right)\right) \quad (3)$$

where $\varphi(\cdot)$ represents linear transformation. Then, the multi-modal feature $\hat{X}^{M} \in R^{N \times C}$ enhanced with hierarchical hyper-information is obtained by cascading $\hat{X}^{m}$.

Algorithm 1: Diff-MT

Input: multi-modal physiological sensor signals $X = \{X_1, X_2, \ldots, X_M\}$
Output: the recognition result of emotional state $\hat{y}$

Initialize
for $i \leftarrow 1$ *to* M **do**
  `// multi-modal signal alignment`
  align channels by Eq.(1) to get $\bar{X}^i$
  `// hierarchical sequence differencing`
  segment $\bar{X}^i$ according to the window w
  calculate the difference $\hat{X}^{i'}$ by Eq.(2)
  perform Eq.(3) to yield hyper-information $\hat{X}^i$
  `// differential hyper-information fusion`
  multi-modal differences $\hat{X}^{M'} \leftarrow$ **Concat**$\left(\hat{X}^{i'}\right)$
  multi-modal hyper-information $\hat{X}^{M} \leftarrow$ **Concat**$\left(\hat{X}^{i}\right)$
end
`// multi-modal global cross-attention encoding`
for $j \leftarrow 1$ *to* M **do**
  embedding position by Eq.(4)
  get global crossmodal feature X'^{j} by Eq.(5)
  global crossmodal feature $X'^{M} \leftarrow$ **Concat**$\left(X'^{j}\right)$
end
`// difference-augmented feature fusion`
calculate multi-modal emotion feature $\tilde{X}^{M}$ by Eq.(6)
`// classification`
generate the output $\hat{y}$ by Eq.(7)
return $\hat{y}$

B. Multi-modal Global Cross-Attention Encoder

By fusing multi-modal emotion signals, we can capitalize on the complementarities between them and stimulate the development of emotion analysis techniques. However, it is challenging to do so due to the multiplicity and heterogeneity of multi-modal signals from various sensors. Providing a latent adaptation across modalities is an excellent way to fuse cross-modal information. Thus, we propose the MGCE module presented in Fig. 3(c) to reveal the global relationship between a specific modality and other modalities. We focus on mid-level feature fusion to address the issues of modal bias and information redundancy. The MGCE module contains positional embedding, global cross-attention, and self-attention.

Positional embedding. Biological signals are essentially time series data where timestamps carry significant information. To ensure that the encoded emotional features emphasize temporal information, we introduce positional embedding [33]. In mathematical terms,

$$\begin{aligned} \boldsymbol{P}[i, 2j] &= \sin\left(\frac{i}{10000^{\frac{2j}{c}}}\right) \quad i \in \{0, 1, \ldots, N\} \\ \boldsymbol{P}[i, 2j+1] &= \cos\left(\frac{i}{10000^{\frac{2j}{C}}}\right) \quad j \in \left\{0, 1, \ldots, \lfloor \frac{C}{2} \rfloor\right\} \end{aligned} \quad (4)$$

where i and j are the indexes of feature tensors. Leveraging sine-cosine encoding, positional embedding generates a matrix with unique locations. Then, we add the acquired positional embedding to the extracted features to preserve temporal semantics in signal processing.

Global cross-attention. Cross-attention is a good practice for cross-modal fusion. It generates attention weights for each paired modality by learning their local correlations. Unlike existing approaches, we focus on the global relevance between every modality and all the involved modalities, including itself. To this end, we design a global cross-modal attention mechanism that allows one modality to receive global latent adaptation. Specifically, we first obtain the query Q_m about $\hat{X}^{m}$ and the K_M, V_M for the global feature $\hat{X}^{M}$ by three transformations ${W_Q}^m$, ${W_K}^M$, and ${W_V}^M$, respectively. On this basis, the global cross-attention encoding X'^{m} can be obtained by the function $G(\cdot)$. In particular,

$$\begin{aligned} G\left(\hat{X}^{m}, \hat{X}^{M}\right) &= \text{SoftMax}\left(\frac{Q_m K_M^T}{\sqrt{C}}\right) V_M \\ &= \text{SoftMax}\left(\frac{\hat{X}^{m} W_Q^m \left(\hat{X}^{M} W_K^M\right)^T}{\sqrt{C}}\right) \hat{X}^{M} W_V^M. \end{aligned} \quad (5)$$

We set up a three-head attention to obtain cross-modal correlations across different feature spaces and construct the MGCE module, as shown in Fig. 3(c). The MGCE is more efficient than the pair-wise relational modeling approach, especially for processing more than two modalities. We design an M-branch network structure with M branches stacked with L MGCE layers to perform the parallel encoding task, making Diff-MT more time-friendly (see Section IV). In this work, we set M = 4 and L = 1 by default. We also facilitate training stability through residual connections and layer normalization. Then, the global cross-attention encoded feature X'^{m} is obtained through forward propagation, embedding high-level relevance across modalities adaptively.

C. Difference-Augmented Feature Fusion

Different physiological signals reflect human emotional states from various perspectives. To fully utilize the complementary nature of multi-modal signal features, we design the

DFF module. As depicted in Fig. 3(d), DFF fuses the hierarchical hyper-information with the weights of relevant modalities. First, the features X'^m are cascaded to obtain the global multi-modal signal representation X'^M. Second, we cascade the differential hyper-information $\hat{X}^{m'}$ (see Section III.A) as the initialized differential multi-modal features $\hat{X}^{M'}$. Then, we pass these two features through multiple self-attention layers to generate the adaptive cross-modal representation and the higher-order differential information, respectively. Finally, they are weight-fused to improve the discriminative properties of the multi-modal features. Mathematically, the final cross-modal features can be obtained by the following equation.

$$\tilde{X}^M = (1-\alpha)\cdot \mathrm{Attn}\left(\hat{X}^{M'}\right) + \alpha \cdot \mathrm{Attn}\left(X'^M\right) \tag{6}$$

where $\mathrm{Attn}(\cdot)$ represents the self-attention module with a default of 5 layers. The learnable parameter α regulates the weight of differential hyper-information during cross-modal fusion. The extracted emotional features from multi-modal emotional signals will go through the classification head, resulting in the emotion state $\hat{y}$, as shown in Eq. (7).

$$\hat{y} = Softmax\left(FC\left(GAP\left(\tilde{X}^M\right)\right)\right) \tag{7}$$

where GAP is the global average pooling and FC is the fully connected layer. The Diff-MT model can effectively capture and identify discriminative emotion patterns reflected by multi-modal physiological signals, enhancing performance on sensor-based emotion recognition tasks. Concretely, the forward propagation of Diff-MT is shown in Algorithm 1.

IV. Experiments

A. Datasets

DEAP dataset. The DEAP [16] dataset involves 32 participants who watched 40 distinct music video clips designed to evoke different emotional states. The participants assess their emotional states of arousal, valence, liking, and dominance on a scale of 1 to 9 according to the self-assessment manikin [4]. The collected physiological signals include EOG, EEG, EMG, and GSR. In this work, we utilize the official preprocessed datasets and exclude the baseline recording. We regard every 1s signal as a sample, resulting in 76800 samples. We classify the scale ratings (1-9) into three levels: negative (1-3), neutral (4-6), and positive (7-9) for non-binary classification. We also categorize negative and positive emotions with a threshold of 5 for binary classification. We set aside samples of one subject as the test set and the other as the training set. After performing leave-one-subject-out trials, we take the average result as the final.

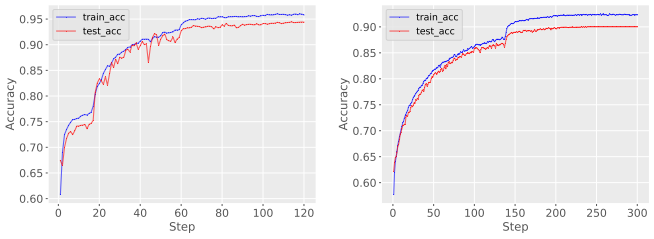
WESAD dataset. The WESAD [31] dataset contains EDA, BVP, EMG, ECG, and RESP signals collected by RespiBAN and Empatica E4 placed on the chest and wrist. The WESAD dataset is compiled from 15 participants, 12 males and 3 females. During testing, participants try to close their eyes to elicit a neutral state. For the stress state, participants engage in speaking about their traits in front of panels. To produce the amusement state, participants watched funny videos. In this paper, we treat per-second signals as a sample for a total of 27287 samples, and perform binary (stress vs. non-stress) and non-binary (stress, neutral, and amusement) classification. The final experimental result is the average of 10-fold cross-validation. We report the average accuracy and F1-score metrics of the experiments.

B. Training Details

We utilize cross-entropy as the loss function and SoftMax as the classifier. To avoid overfitting, we adopt data shuffling and dropout tricks. We also employ the adaptive scheduling strategy, i.e., when the loss function on the validation set does not decrease for successive ten epochs, the learning rate is multiplied by 0.1. For the WESAD dataset, we train Diff-MT with 100 epochs, and the learning rate is initialized to 0.002. For the DEAP dataset, the epoch is set to 300, and the initial learning rate is 0.005. The batch size for WESAD dataset is set to 512, while for the DEAP dataset it is set to 1024. Fig. 5 shows the training curve of Diff-MT for non-binary emotion recognition.

C. Results and Comparations with SOTAs

Cross-validation results. The final result of Diff-MT is the average of multiple cross-validation results by default. In order to disclose more experimental details, we further report the experimental results of cross-validation. Fig. 6 illustrates the 10-fold cross-validation results of Diff-MT on WESAD dataset (non-binary classification). We obtain high performance on each fold of validation data, where the optimal accuracy is as high as 96.19%, and the optimal F1-score is as high as 95.62%. Although physiological signals are highly objective, there is still some degree of individual variation. This intra-class variation makes the emotion recognition task challenging. Fig. 7-8 illustrates the results of leave-one-subject-out testing on the DEAP dataset (non-binary classification). As we can see, our method is generally robust and achieves promising average



(a) Training curve on WESAD (b) Training curve on DEAP

Fig. 5. The training curve of Diff-MT. We reported the first-fold and leaving-the-first-subject validation on the WESAD and DEAP datasets, respectively.

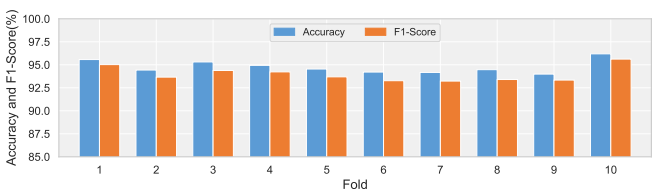


Fig. 6. Results of 10-fold cross validation on the WESAD dataset.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3458976

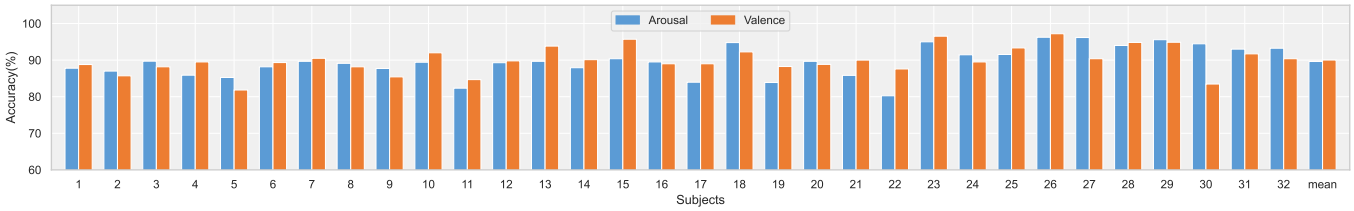


Fig. 7. Mean accuracy of leave-one-subject-out trails on the DEAP dataset.

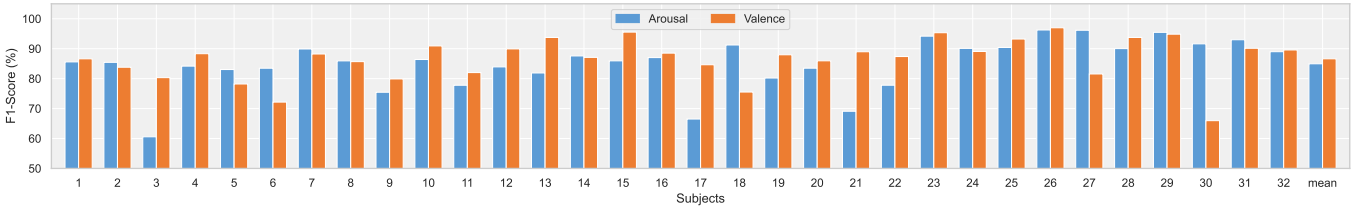


Fig. 8. Mean F1-Score of leave-one-subject-out trails on the DEAP dataset.

TABLE I
THE COMPLEXITY AND PERFORMANCE OF DIFF-MT COMPARED WITH OTHER METHODS

Dataset	Train_n	Test_n	Model	Modality	Acc.(%)	F1(%)	Train_t(s)	Inference_t(ms)	Para.(M)
DEAP	74400	2400	LSTM* [44]	EEG	54.71	50.38	875.13	4.39	4.82
			CNN-LSTM* [45]	EEG	58.02	55.24	328.12	1.84	1.12
			Segade* [10]	EEG	48.76	41.85	106.53	0.67	0.22
			EEG-Net* [43]	EEG	54.14	43.80	109.95	0.21	0.02
			Tseption* [8]	EEG	54.10	44.54	13.60	0.12	0.07
			Diff-MT(ours)	EEG+EOG+EMG+GSR	89.62	84.96	19.24	0.17	0.91
WESAD	24558	2729	LSTM* [44]	ECG	92.45	91.33	470.58	7.70	4.41
			CNN-LSTM* [45]	ECG	94.21	93.54	74.62	4.17	0.19
			Segade* [10]	ECG	80.14	77.38	46.83	1.53	2.02
			EEG-Net* [43]	ECG	58.34	43.43	6.43	0.25	0.01
			Tseption* [8]	RESP+ECG	82.52	79.95	2.49	0.23	0.11
			Diff-MT(ours)	RESP+BVP+ECG+EDA	94.78	93.98	12.22	0.35	0.28

[1] We report the average training time for one epoch and the average inference time for one sample.
[2] These methods marked * are the ones we reproduced in the same experimental settings.

accuracies of 89.62% and 90.02% on the arousal and valence data in the DEAP dataset, respectively.

Complexity and real-time performance. Healthcare applications often demand prompt feedback, so we focus on the model's complexity and real-time performance. To prove the superiority of Diff-MT, we compare its number of parameters, training time, and inference time with other methods, as shown in Table I. We reproduce some signal processing methods in the same experimental environment, including LSTM [44], CNN-LSTM [45], Segad [10], EEG-Net [43], Tseption [8]. We report the three-class accuracy and F1-Score on DEAP (arousal) and WEAD datasets. It can be seen that Diff-MT shows promising real-time performance on the DEAP and WESAD datasets, and each sample's average inference time is only 0.17 ms and 0.35 ms, respectively. Although the single sample inference time of EEG-Net [43] and Tseption [8] is slightly faster than our method, their recognition performance is far behind ours. It is worth noting that although we use four modalities of emotion signals for analysis, the additional data has little impact on the inference speed of our method but dramatically improves the accuracy of emotion recognition. We can conclude that Diff-MT achieves the optimal trade-off between performance and complexity and is qualified to provide a reliable emotion-aware response for IoT-based medical applications.

Comparisons with state-of-the-art methods. In order to demonstrate that the proposed method can effectively recognize emotion, we perform binary and non-binary classification tasks on two publicly available datasets and compare the experimental results with existing methods. The corresponding acurracy and F1-score are shown in Table II and Table III. For the DEAP dataset, Diff-MT achieves 94.83% and 93.95% accuracy for binary classification on the condition of valence and arousal, respectively. For non-binary classification, Diff-MT yields 90.62% and 89.62% with valence and arousal, respectively. It is more advantageous in temporal modeling compared to RNN-based methods [19] and LSTM-based methods [15], [25]. For the WESAD dataset, our method achieves the desired performance with an average accuracy and F1-score of 98.91% and 98.83% for binary classification, respectively. For non-binary classification, Diff-MT obtains the highest accuracy of 94.78% and F1-score of 93.98%. As shown in Table III, Diff-MT outperforms the state-of-the-art methods Attention-LRCN [5] and SELF-CARE [29] on

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DEAP DATASET.

Methods	Valence		Arousal		Year
	F1	Acc	F1	Acc	
Binary Classification					
John et al. [1]	-	73.14	-	73.06	2016
BDAE [23]	-	85.20	-	80.50	2016
Zhang et al. [35]	-	73.06	-	80.78	2017
PCRNN [34]	-	90.80	-	91.03	2018
MMResLSTM [25]	-	92.30	-	92.87	2019
CRNIN [19]	-	91.95	-	93.06	2020
PCC [22]	-	89.49	-	92.86	2020
NeuCube + variance [24]	-	78.00	-	74.00	2020
LF-DfE [15]	-	75.50	-	76.00	2021
DCERNet-SVM [27]	-	88.10	-	-	2022
He et al. [11]	-	64.33	-	63.25	2022
E-EmotiConNet [14]	-	93.09	-	93.69	2022
TSception [8]	62.33	59.14	63.24	61.57	2022
Zhang et al. [36]	-	84.27	-	85.86	2023
Diff-MT (ours)	94.29	94.83	91.86	93.95	-
Non-binary Classification					
Liu et al. [4]	-	53.40	-	51.00	2012
John et al. [1]	-	62.33	-	60.70	2016
Samarth et al. [32]	-	66.79	-	57.58	2017
Zheng et al. [37]	-	69.67	-	-	2019
DCERNet-SVM [27]	-	86.50	-	-	2022
Diff-MT (ours)	86.62	90.02	84.96	89.62	-

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE WESAD DATASET.

Methods	Binary		Non-binary		Year
	F1	Acc	F1	Acc	
Schmidt et al. [31]	-	93.12	-	80.34	2018
Sirat et al. [30]	90.20	94.70	75.80	83.40	2020
Transformer [2]	83.30	91.10	-	-	2021
SMA [17]	97.74	97.75	82.95	88.28	2021
StressNAS [12]	-	92.87	-	83.43	2021
AttX [16]	91.11	92.08	-	-	2021
sTree [40]	95.10	95.80	-	-	2021
U-Net [38]	-	-	-	91.14	2021
Lisowska et al. [21]	83.70	-	-	65.30	2021
Bhatti et al. [3]	91.10	92.80	-	-	2021
H-CNN [28]	86.18	88.56	64.15	75.21	2021
Garg et al. [9]	83.34	84.17	65.73	67.56	2021
MoStress [7]			-	86.00	2022
Choi et al. [6]	95.47	97.11	-	-	2022
SELF-CARE [29]	92.93	94.12	71.97	86.34	2023
Deep CNN-CBAM [39]	97.10	97.50	-	-	2023
Attention-LRCN [5]	98.13	98.44	76.24	88.21	2023
Diff-MT (ours)	98.83	98.91	93.98	94.78	-

both binary and non-binary classification tasks. In general, our method substantially outperforms existing approaches and is able to provide more reliable emotional feedback.

Visualization of confusion matrixes. To illustrate how the model recognizes different emotion categories, we visualize the confusion matrix of the classification results. As shown in Fig. 9(a-c), Diff-MT performs outstandingly in binary classification tasks on both datasets. Diff-MT detects emotional stress on the WESAD dataset with an accuracy of 98.57%. The negative emotion recognition accuracies on the DEAP dataset labeled arousal and validity were 93.23% and 94.92%, respectively. In addition, Diff-MT also performs accurately on the non-binary classification task, as described in Fig. 9(d-f), especially on the DEAP dataset, with an accuracy of 97.58% when attempting to recognize negative emotional state in the arousal condition. In addition, on the WESAD dataset, Diff-MT identifies emotions with an accuracy higher than 93% in both labels, demonstrating the potential of Diff-MT for finer-grained emotion analysis. Although Diff-MT has limitations in recognizing neutral samples due to the unbalanced distribution, it can overall recognize human emotional states effectively.

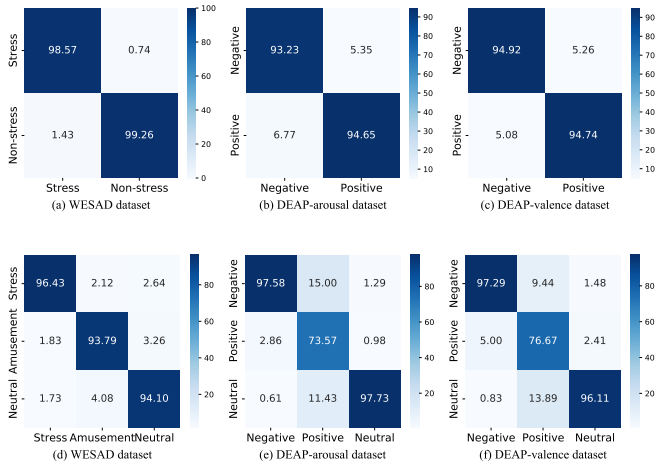


Fig. 9. Confusion matrix for WESAD and DEAP datasets.

V. ABLATION STUDY

This section explores the validity of Diff-MT and its components. We report the three-class classification results on the WESAD dataset with 10-fold cross-validation by default.

The effectiveness of modules in Diff-MT. The proposed emotion-aware system Diff-MT consists of three main modules: DHE, MGCE, and DFF. The DHE serves to preprocess multi-modal signals and generate hierarchical hyper-information. The MGCE includes the position embedding and cross-attention layers, which focus on mining the potential correlation of each modality with global modality features. Finally, the DFF is responsible for extracting high-level multi-modality features. We conduct ablation experiments on these components, and the experimental results are shown in Table IV. It can be seen that all of these modules are practical and improve the model's performance. Adding the DHE module can increase accuracy by 1.64%. Utilizing the MGCE module can improve accuracy by 1.28%. In addition, the PE and the learnable parameter also optimize Diff-MT. The Diff-MT

TABLE IV
COMPARISON OF DIFF-MT AND ITS VARIANTS.

Variants	Accuracy(%)	F1-Score(%)
Diff-MT w/o DHE	93.14	92.45
Diff-MT w/o PE	94.70	93.85
Diff-MT w/o MGCE	93.50	93.12
Diff-MT w/o DFF	93.53	93.25
Diff-MT w/o α	94.10	93.38
Diff-MT (ours)	94.78	93.98

[1] PE is positional embedding in Section III.B.
[2] α is the learnable parameter in Eq. (6).

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3458976

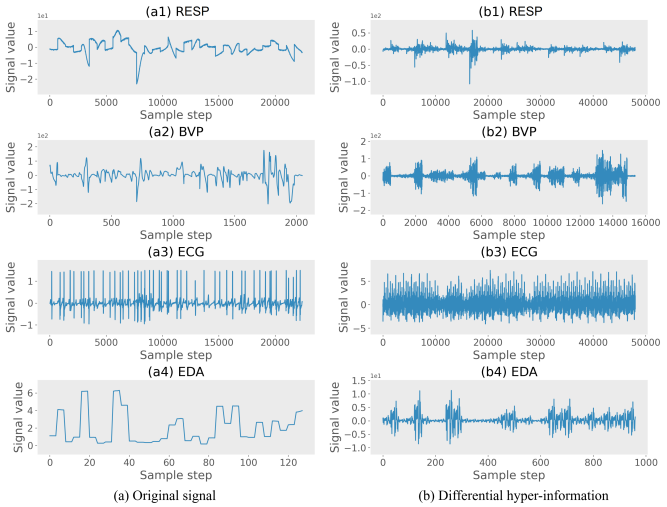


Fig. 10. Visualization of differential hyper-information. Upon comparison of the original signal (a) with its corresponding processed differential hyper-information (b) of the WESAD dataset.

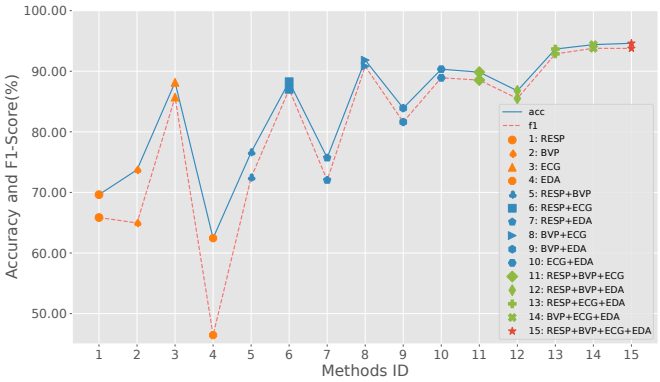


Fig. 11. The results of different modal signals on the WESAD dataset.

model with the DFF module removed loses 1.25% accuracy and 0.73% F1-Score, indicating that the differential hyper-information is essential to the Diff-MT model. Overall, our proposed modules are practical and contribute to the success of Diff-MT. Moreover, these modules are compatible and can be ported to other models to enhance the accuracy of emotion recognition.

Influence of Differential Hyper-information. Table IV demonstrates the importance of the DHE module, which is responsible for extracting differential hyper-information. This components greatly benefits the proposed Diff-MT in two ways. First, it enhances the registration of how physiological signals change with emotion by hierarchically capturing higher-order information. Second, the hyper-information is fused with multi-modal features after cross-attention encoding, elevating the overall recognition accuracy. To more intuitively illustrate the positive impact of differential hyper-information on physiological signal modeling, we provide a visualization of the encoded multi-modal differential features in Fig. 10. It can be seen that this module is compatible and can be applied to different modalities of emotional signals. As expected, the generated differential hyper-information offers more details and emphasizes signal changes while retaining global trends, enabling the model to capture more distinctive multi-modal emotion features.

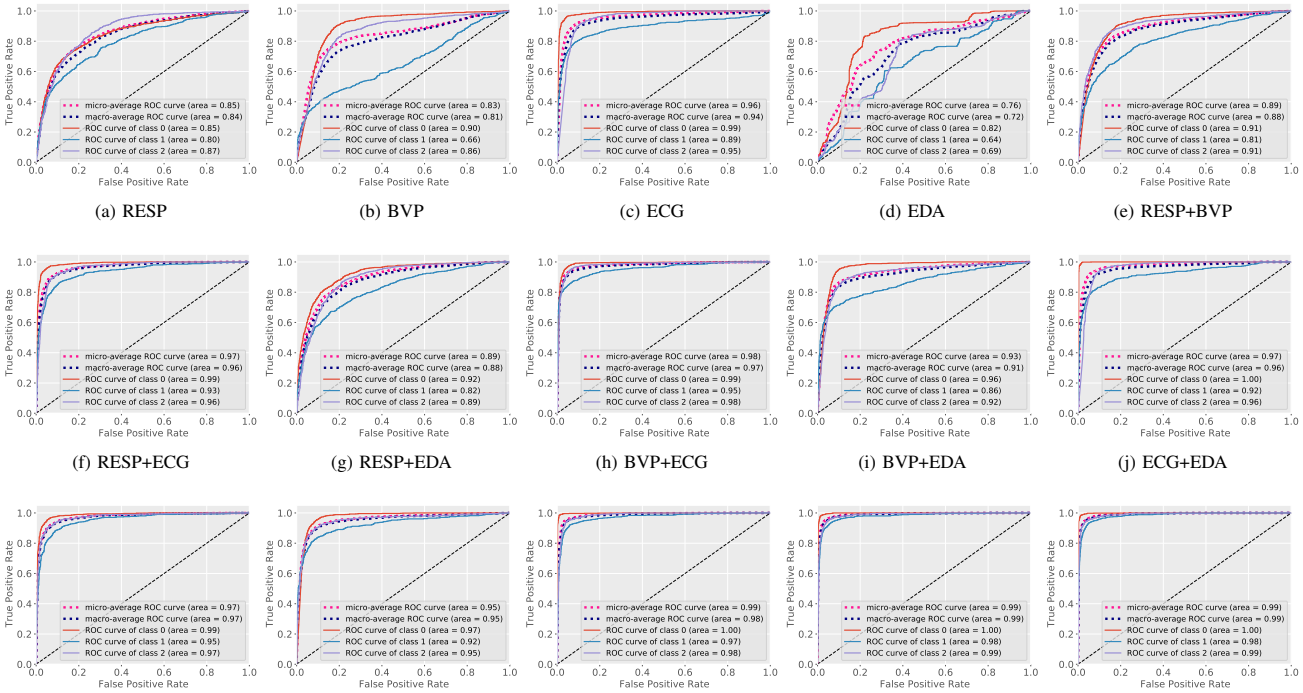
Combinations of multi-sensor signals. Leveraging physiological signals can provide comprehensive diagnostic indicators assisting healthcare applications. However, adding more modalities may not always lead to better results, and reducing redundant features and mutual interference is challenging. We explore the effects of different signals and their combinations, as displayed in Fig. 11. Employing the ECG signal captures more emotional information, achieving an F1-score and accuracy of 85.72% and 88.17%. For bimodal signals, we recommend the ensemble of BVP and ECG, as experiments have proven that they complement each other well. On this basis, increasing the EDA accessorially enhances the F1-score and accuracy by 2.96% and 2.56%, respectively. Diff-MT performs best with four-modal feature fusion, achieving 93.98% F1-score and 94.78% accuracy. In addition, we illustrate the compatibility of Diff-MT to various signals with Receiver Operating Characteristic (ROC) curves, shown in Fig. 12. On the whole, fusing multi-source signals is generally better than only covering a single signal. Promising results were achieved under all combination setups, except for the single-modal setting, where the effect on EDA and BVP was restricted. The above variants have a slightly lower recognition ability for neutral emotions (blue curve) than other labels, caused by unbalanced sample data.

VI. Discussions

The proposed emotion-aware system has extensive potential applications, such as mental health monitoring, stress management, resilience building, and enhancement of virtual reality therapy. For instance, it can be employed to assess emotional states and stress levels in individuals with autism spectrum disorder who have communication and mental disorders, providing a more personalized and comfortable approach to treatment and recovery. We believe that the proposed emotion-aware system will have significant implications for human-centric artificial intelligence, especially in healthcare. However, this work still has some limitations. The individual variation in emotional reaction is a critical factor that impacts the ability to recognize emotions. Although our method achieves satisfactory performance on emotion perception for most subjects, the performance of specific individuals still needs to be improved. It will be significant to provide more personalized services for patients or other users. Moreover, this work mainly focuses on the binary and ternary classification tasks of sentiment. With the rapid advancement of human-centered artificial intelligence technology and the increasing demands of diverse scenarios, providing fine-grained emotion perception will be even more essential.

VII. Conclusion

In this article, we present a novel IoMT-perspective framework for emotion recognition based on multiple sensor signals in WBAN to facilitate intelligent healthcare. We introduce



(a) RESP (b) BVP (c) ECG (d) EDA (e) RESP+BVP

(f) RESP+ECG (g) RESP+EDA (h) BVP+ECG (i) BVP+EDA (j) ECG+EDA

(k) RESP+BVP+ECG (l) RESP+BVP+EDA (m) RESP+ECG+EDA (n) BVP+ECG+EDA (o) RESP+BVP+ECG+EDA

Fig. 12. The ROC curves for models with different signals. Labels 0, 1, and 2 correspond to "stress", "amusement", and "neutral", respectively. The closer the curve is to the upper left corner, the better the classification performance. It can be seen that Diff-MT can deal with single-modal and multi-modal signals and performs well in recognizing emotions, especially stress states.

the deep learning model named Diff-MT, which consists of DHE, MGCE, and DFF modules and can extract multi-modal signal features without manual feature engineering for emotion recognition. It is a versatile signal processing model, which can potentially be employed for other physiological signal analysis. We conduct the experiments on the two public datasets: WESAD and DEAP. The results show that our method is real-time and achieves excellent recognition accuracy and high performance in binary and non-binary classification tasks, a fact that has significant advantages in emotion-aware IoMT-based applications. Our future work will focus on personalized emotion recognition for specific subject and fine-grained sentiment perception.

REFERENCES

[1] J. Atkinson and D. Campos, "Improving BCI-based Emotion Recognition by Combining EEG Feature Selection and Kernel Classifiers," in Expert Syst. Appl., vol. 47, pp. 35-41, 2016.

[2] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad, "A Transformer Architecture for Stress Detection from ECG," in Proc. Int. Symp. Wearable Comput. (ISWC), Virtual, USA, 2021. pp. 132-134.

[3] A. Bhatti, B. Behinaein, D. Rodenburg, P. Hungler and A. Etemad, "Attentive Cross-modal Connections for Deep Multimodal Wearable-based Emotion Recognition," in Proc. Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW), Nara, Japan, 2021, pp. 01-05.

[4] M. M. Bradley and P. J. Lang, "Measuring Emotion: the Self-Assessment Manikin and the Semantic Differential," in J. Behav. Ther. Exp. Psychiatry, vol. 25, no. 1, pp. 49-59, Mar 1994.

[5] J. Choi, G. Hwang, J. S. Lee, M. Ryu, and S. J. Lee, "Weighted Knowledge Distillation of Attention-LRCN for Recognizing Affective States from PPG Signals," in Expert Syst. Appl., vol. 233, no. 120883, pp. 1-10, 2023.

[6] J. Choi, J. S. Lee, M. Ryu, G. Hwang, G. Hwang, and S. J. Lee, "Attention-LRCN: Long-term Recurrent Convolutional Network for Stress Detection from Photoplethysmography," 2022, pp. 1-6.

[7] A. de Souza, M. B. Melchiades, S. J. Rigo and G. d. O. Ramos, "MoStress: a Sequence Model for Stress Classification," in Proc. Int. Jt. Conf. Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8.

[8] Y. Ding, N. Robinson, S. Zhang, Q. Zeng and C. Guan, "TSception: Capturing Temporal Dynamics and Spatial Asymmetry From EEG for Emotion Recognition," in IEEE Trans. Affect., vol. 14, no. 3, pp. 2238-2250, 1 July-Sept 2023.

[9] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, "Stress Detection by Machine Learning and Wearable Sensors," in Proc. Int. Conf. Intell. User Interfaces (IUI), College Station, TX, USA, 2021. pp. 43-45.

[10] Z. Guo, C. Ding, X. Hu, and C. Rudin, "A Supervised Machine Learning Semantic Segmentation Approach for Detecting Artifacts in Plethysmography Signals from Wearables," Physiol. Meas., vol. 42, no. 125003, pp. 1-17, 2021.

[11] Z. He, Y. Zhong and J. Pan, "Joint Temporal Convolutional Networks and Adversarial Discriminative Domain Adaptation for EEG-Based Cross-Subject Emotion Recognition," in IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Singapore, 2022, pp. 3214-3218.

[12] L. Huynh, T. Nguyen, T. Nguyen, S. Pirttikangas, and P. Siirtola, "Stress-NAS: Affect State and Stress Detection Using Neural Architecture Search," in Proc. ACM Int. Symp. Wearable Comput. (ISWC), Virtual, USA, 2021.

[13] Q. Zhou, Z. Qu, S. Guo, B. Luo, J. Guo, Z. Xu, R. Akerkar, "On-Device Learning Systems for Edge Intelligence: A Software and Hardware Synergy Perspective," in IEEE Internet Things J., vol. 8, no. 15, pp. 11916-11934, 1 Aug.1, 2021.

[14] L. Jin and E. Y. Kim, "E-EmotiConNet: EEG-based Emotion Recognition with Context Information," in Proc. Int. Jt. Conf. Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-8.

[15] V. M. Joshi and R. B. Ghongade, "EEG Based Emotion Detection Using Fourth Order Spectral Moment and Deep Learning," in Biomed. Signal Process. Control., vol. 68, no. 102755, pp. 1-12, 2021.

[16] S. Koelstra, C. Mühl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, L. Patras, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," in IEEE Trans. Affect., vol. 3, no. 1, pp. 18-31, 2012.

[17] K. Lai, S. N. Yanushkevich and V. P. Shmerko, "Intelligent Stress Monitoring Assistant for First Responders," in IEEE Access, vol. 9, pp. 25314-25329, 2021.

[18] Q. Zhou, S. Guo, J, Pan, J. Liang, J. Guo, Z. Xu, J, Zhou, "PASS: Patch Automatic Skip Scheme for Efficient On-Device Video Perception," in IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 5, pp. 3938-3954, May 2024.

[19] J. Liao, Q. Zhong, Y. Zhu, and D. Cai, "Multimodal Physiological Signal Emotion Recognition Based on Convolutional Recurrent Neural Network," in Proc. IOP Conf. Ser. Mater. Sci. Eng., vol. 782, no. 32005, pp. 1-11, 2020.

[20] Z. Gao, Y. Li, Y. Yang, N. Dong, X. Yang and C. Grebogi, "A Coincidence-Filtering-Based Approach for CNNs in EEG-Based Recognition," in IEEE Trans Industr Inform, vol. 16, no. 11, pp. 7159-7167, Nov. 2020

[21] A. Lisowska, S. Wilk, and M. Peleg, "Catching Patient's Attention at the Right Time to Help Them Undergo Behavioural Change: Stress Classification Experiment from Blood Volume Pulse," in Lect. Notes Comput. Sci. (AIME), 2021, Virtual Event, June. pp. 72-82.

[22] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, Y. Bi, "EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder," in Front. Syst. Neurosci., vol. 14, no. 43, pp. 1-14, September 2020.

[23] J. Liu, W. Zheng, and B. Lu, "Emotion Recognition Using Multimodal Deep Learning," in Lect. Notes Comput. Sci. 2016, Cham, pp. 521-529.

[24] Y. Luo et al., "EEG-Based Emotion Classification Using Spiking Neural Networks," in IEEE Access, vol. 8, pp. 46007-46016, 2020.

[25] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion Recognition using Multimodal Residual LSTM Network," in Proc. ACM Int. Conf. Multimed., Nice, France, 2019. pp. 176-183.

[26] G. Manogaran, M. Alazab, H. Song, and N. Kumar, "CDP-UA: Cognitive Data Processing Method Wearable Sensor Data Uncertainty Analysis in the Internet of Things Assisted Smart Medical Healthcare Systems," in IEEE J. Biomed. Health Inform., vol. 25, no. 10, pp. 3691-3699, 2021.

[27] N. Pusarla, A. Singh, and S. Tripathi, "Learning DenseNet Features From EEG Based Spectrograms for Subject Independent Emotion Recognition," in Biomed. Signal Process. Control, vol. 74, no. 103485, pp. 1-12, 2022.

[28] N. Rashid, L. Chen, M. Dautta, A. Jimenez, P. Tseng, and M. A. Al Faruque, "Feature Augmented Hybrid CNN for Stress Recognition Using Wrist-based Photoplethysmography Sensor,", in Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., vol. 2021, pp. 2374-2377, Nov 2021.

[29] N. Rashid, T. Mortlock, and M. A. A. Faruque, "Stress Detection Using Context-Aware Sensor Fusion From Wearable Devices," in IEEE Internet Things J., vol. 10, no. 16, pp. 14114-14127, 2023.

[30] S. Samyoun, A. Sayeed Mondol and J. A. Stankovic, "Stress Detection via Sensor Translation," in Proc. Annu. Int. Conf. Distrib. Comput. Sens. Syst. (DCOSS), Marina del Rey, CA, USA, 2020, pp. 19-26.

[31] P. Schmidt, A. Reiss, and R. Duerichen, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in Proc. ACM Int. Conf. Multimodal Interact., Boulder, CO, USA, 2018. pp. 400-408.

[32] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset," in Proc. AAAI Conf. Artif. Intell. (AAAI), San Francisco, California, USA, 2017. PP. 4746-4752.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, L. Polosukhin, "Attention Is All You Need," in Proc. Int. Conf. Neural Inf. Process. Syst., Long Beach, California, USA, 2017. pp. 6000-6010.

[34] Y. Yang, Q. Wu, M. Qiu, Y. Wang and X. Chen, "Emotion Recognition from Multi-Channel EEG through Parallel Convolutional Recurrent Neural Network," in Proc. Int. Jt. Conf. Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018, pp. 1-7.

[35] Q. Zhang, X. Chen, Q. Zhan, T. Yang, and S. Xia, "Respiration-based Emotion Recognition with Deep Learning," in Comput. Ind., vol. 92-93, pp. 84-90, 2017.

[36] Y. Zhang, Y. Zhang, and S. Wang, "An Attention-based Hybrid deep Learning Model for EEG Emotion Recognition," in Signal Image Video Process., vol. 17, pp. 2305-2313, 2023.

[37] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying Stable Patterns over Time for Emotion Recognition from EEG," in IEEE Trans. Affect., vol. 10, no. 3, pp. 417-429, 2019.

[38] Z. Guo, C. Ding, X. Hu, and C. Rudin, "A Supervised Machine Learning Semantic Segmentation Approach for Detecting Artifacts in Plethysmography Signals from Wearables," in Physiol. Meas., vol. 42, no. 125003, pp. 1-18, 2021.

[39] T. Fan, S, Qiu, Z. Wang, H. Zhao, J. Jiang, Y. Wang, J. Xu, T. Sun, N. Jiang, "A New Deep Convolutional Neural Network Incorporating Attentional Mechanisms for ECG Emotion Recognition," in Comput. Biol. Med., vol. 159, no. 106938, pp. 1-11, 2023.

[40] A. Liapis, E. Faliagka, C. Katsanos, C. Antonopoulos, N. Voros, "Detection of Subtle Stress Episodes During UX Evaluation: Assessing the Performance of the WESAD Bio-Signals Dataset," in Proc. Hum. Comput. Interact., Cham, 2021, pp. 238-247.

[41] M. Nouman, S. Y. Khoo, M. A. P. Mahmud and A. Z. Kouzani, "Recent Advances in Contactless Sensing Technologies for Mental Health Monitoring," in IEEE Internet Things J., vol. 9, no. 1, pp. 274-297, Jan, 2022.

[42] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola and M. Montes-y-Gómez, "Detecting Mental Disorders in Social Media Through Emotional Patterns - The Case of Anorexia and Depression," in IEEE Trans. Affect., vol. 14, no. 1, pp. 211-222, 1 Jan-March, 2023.

[43] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance. "EEGNet: A Compact Convolutional Neural Network for EEG-based Brain-computer Interfaces," in J Neural Eng. vol. 15, no. 5, pp. 1-17, Jun. 2018.

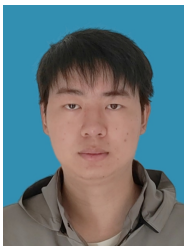
[44] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997.

[45] T. Wilaiprasitporn, A. Ditthapron, K. Matchaparn, T. Tongbuasirilai, N. Banluesombatkul and E. Chuangsuwanich, "Affective EEG-Based Person Identification Using the Deep Learning Approach," in IEEE Trans. Cogn. Dev. Syst., vol. 12, no. 3, pp. 486-496, Sept. 2020.

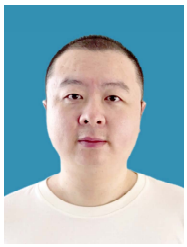
Chang Li received the B.Sc. degree from Jiangsu Ocean University, Lianyungang, China, in 2019. She received the M.Sc. degrees in computer science and technology from Hohai University, Nanjing, China, in 2021. She is currently pursuing Ph.D. degree in Hohai University, Nanjing, China. Her research interests include media computing, deep learning, distributed computing, smart sensor device, and computer vision especially action recognition and pose estimation.

Yingchi Mao was born in China. She received the B.Sc. and M.Sc. degrees in computer science and technology from Hohai University, Nanjing, China, in 1999 and 2003, respectively, and the Ph.D. degree in computer science and technology from Nanjing University, China, in 2007. She is currently a Professor with the College of Computer and Information, Hohai University. Her research interests include distributed computing, wireless sensor networks, and distributed data management.

Qian Huang received the B. Sc. degree in computer science from Nanjing University, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2010. From 2010 to 2012, he was a deputy technical manager of Mediatek (Beijing) Incorporation, Beijing, China. Since Dec. 2012, he serves as the dean of Computer Science and Technology Department, Hohai University, Nanjing, China. His research interests include media computing, data mining, and intelligent education.



Weiliang Xie received his bachelor's degree in software engineering from Nanchang Hangkong University in 2022. Currently, he is studying at the School of Computer and Software of Hohai University for his master's degree. His research interests include computer vision, deep learning, pose estimation, and action recognition.



Xiaoming He (Member, IEEE) received the Ph.D. degree in Computer Science and Software Engineering from Hohai University, Nanjing, China, in 2023. He is currently a Lecturer with the College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China. Prior to work, he was a Visiting Research Fellow in Singapore University of Technology and Design. His current research interests include edge intelligence and FPGA-based AI accelerator.

Jie Wu is the Director of the Center for Networked Computing and Laura H. Carnell professor at Temple University. He also serves as the Director of International Affairs at College of Science and Technology. He served as Chair of Department of Computer and Information Sciences from the summer of 2009 to the summer of 2016 and Associate Vice Provost for International Affairs from the fall of 2015 to the summer of 2017. Prior to joining Temple University, he was a program director at the National Science Foundation and was a distinguished professor at Florida Atlantic University. His current research interests include mobile computing and wireless networks, routing protocols, network trust and security, distributed algorithms, and cloud computing. Dr. Wu regularly publishes in scholarly journals, conference proceedings, and books. He serves on several editorial boards, including IEEE Transactions on Mobile Computing, IEEE Transactions on Service Computing, Journal of Parallel and Distributed Computing, and Journal of Computer Science and Technology. Dr. Wu is/was general chair/co-chair for IEEE IPDPS'08, IEEE DCOSS'09, IEEE ICDCS'13, ACM MobiHoc'14, ICPP'16, IEEE CNS'16, WiOpt'21, and ICDCN'22 as well as program chair/cochair for IEEE MASS'04, IEEE INFOCOM'11, CCF CNCC'13, and ICCCN'20. He was an IEEE Computer Society Distinguished Visitor, ACM Distinguished Speaker, and chair for the IEEE Technical Committee on Distributed Processing (TCDP). Dr. Wu is a Fellow of the AAAS and IEEE. He is the recipient of the 2011 China Computer Federation (CCF) Overseas Outstanding Achievement Award.