MSF-Net: Multi-stage fusion network for emotion recognition from multimodal signals in scalable healthcare

Information Fusion 119 (2025) 103028

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

MSF-Net: Multi-stage fusion network for emotion recognition from multimodal signals in scalable healthcare

Md. Milon Islam [a,*], Fakhri Karray [a,b], Ghulam Muhammad [c]

[a] Centre for Pattern Analysis and Machine Intelligence, Department of Electrical and Computer Engineering, University of Waterloo, N2L 3G1, Ontario, Canada
[b] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[c] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, 11543, Saudi Arabia

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Multi-stage fusion
Vision transformer
Bi-directional Gated Recurrent Unit
Triplet attention
Scalable healthcare

ABSTRACT

Automatic emotion recognition has attracted significant interest in healthcare, thanks to remarkable developments made recently in smart and innovative technologies. A real-time emotion recognition system allows for continuous monitoring, comprehension, and enhancement of the physical entity's capacities, along with continuing advice for enhancing quality of life and well-being in the context of personalized healthcare. Multimodal emotion recognition presents a significant challenge in terms of efficiently using the diverse modalities present in the data. In this article, we introduce a Multi-Stage Fusion Network (MSF-Net) for emotion recognition capable of extracting multimodal information and achieving significant performances. We propose utilizing the transformer-based structure to extract deep features from facial expressions. We exploited two visual descriptors, local binary pattern and Oriented FAST and Rotated BRIEF, to retrieve the computer vision-based features from the facial videos. A feature-level fusion network integrates the extraction of features from these modules, directing the output into the triplet attention technique. This module employs a three-branch architecture to compute attention weights to capture cross-dimensional interactions efficiently. The temporal dependencies in physiological signals are modeled by a Bi-directional Gated Recurrent Unit (Bi-GRU) in forward and backward directions at each time step. Lastly, the output feature representations from the triplet attention module and the extracted high-level patterns from Bi-GRU are fused and fed into the classification module to recognize emotion. The extensive experimental evaluations revealed that the proposed MSF-Net outperformed the state-of-the-art approaches on two popular datasets, BioVid Emo DB and MGEED. Finally, we tested the proposed MSF-Net in the Internet of Things environment to facilitate real-world scalable smart healthcare application.

1. Introduction

Emotion recognition allows machines to perceive and interpret human emotions based on the data provided from the individuals through the use of various smart technologies. Nowadays, the research for emotion recognition draws a lot of interest in analyzing affective data [1]. Emotion recognition encompasses various applications, including human–computer interaction, surveillance, security monitoring, psychological analysis, and healthcare monitoring [2]. Automated emotion recognition is essential for the advancement of smart healthcare facilities because it allows patients to be monitored with conditions such as stress and depression that allowed timely treatment and prevention [3].

There are several ways to recognize emotions from the individuals depending on the data modalities. The available data modalities for emotion recognition are video, audio, texts, and physiological signals [4]. In the State-of-the-Art (SOTA), a few of the systems developed based on only the use of single modal data such as facial cues, texts, acoustic signals, posture, or physiological data. Moreover, some of the frameworks exploited different combinations of above-mentioned data modalities for achieving robust outcomes. The unimodal gives useful information but offers only partial information to the systems, making them inadequate, less reliable, and inefficient [5]. Besides, the integration of data from numerous sources through multimodal emotion recognition provides a more efficient depiction of emotions. The multimodal representation of one subject from multiple perspectives ensures a more accurate view of human emotions, allowing a greater understanding of human behavior [6].

* Corresponding author.
E-mail addresses: milonislam@uwaterloo.ca (M.M. Islam), karray@uwaterloo.ca (F. Karray), ghulam@ksu.edu.sa (G. Muhammad).

https://doi.org/10.1016/j.inffus.2025.103028
Received 8 October 2024; Received in revised form 31 December 2024; Accepted 10 February 2025
Available online 19 February 2025

Over the decades, several emotion recognition systems have been introduced based on Machine Learning (ML) methods using several data modalities [7]. In these systems, the features from the raw signals are extracted and input into the ML algorithms for classification. The ML-based systems performed well on comparatively small dataset and are able to demonstrate the interpretability by making relation with the retrieved features to emotional expressions. Domínguez-Jiménez et al. [8] exploited Random Forest (RF) method to detect emotion from physiological signals by analyzing the frequency and time domain features. Multiple ML algorithms are employed in [9] to recognize speech emotion by extracting Mel-Frequency Cepstral Coefficient (MFCC) and mel spectrogram from the corpus. Azam et al. [10] utilized six supervised ML techniques to detect patient's emotions in healthcare where Multi-Layer Perceptron (MLP) performed the best with 87% accuracy. However, the major issues involved in the ML-based emotion detection systems are the manual feature extraction process, and they are not generalized efficiently. The recent trend for emotion recognition is the exploitation of Deep Learning (DL) methods to detect and interpret emotions from multiple input modalities, including facial cues, speech features, and physiological data [11]. The capability of automatic complex and hierarchical feature extraction from the raw signals, making the DL-based systems more accurate for emotion recognition. Among the DL algorithms, the Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are widely adopted to detect emotion due to the capacity of learning hierarchical depiction of image-based features and temporal dependencies. Dwijayanti et al. [12] introduced face and emotion recognition simultaneously ultizing CNN architectures for humanoid robot application. A two-level ensemble classifier is developed based on the pyramidal one-dimensional CNN in [13] to classify emotion states. Kansizoglou et al. [14] trained RNN architecture through facial landmarks to identify emotions for long-term behavior modeling. A Long Short-Term Memory (LSTM) is presented in [15] to recognize emotion from multi-channel EEG data. A few of the difficulties in the DL-based emotion recognition systems are the lack of labeled datasets for efficient model training, and resource intensive for implementation in practical scenarios.

In recent years, a few research groups have focused on developing hybrid networks for emotion recognition [16]. These hybrid networks comprise two parts: CNN to capture spatial hierarchies in visual data and RNN or LSTM to model temporal dependencies in sequential data. The visual network (CNN) extracted the local feature which are sent to the LSTM network for emotion classification. Singh et al. [17] proposed a hybrid architecture consisting of 3D-CNN and ConvLSTM to recognize emotions from facial videos. Similarly, CNN-RNN based hybrid architecture is developed in [18] to detect human emotions using facial expressions derived from visual data. In these scenarios, some part of the features retrieved by CNN are ignored whenever concatenating with the RNN network due to the gradient vanishing problem. To address the challenges associated with feature concatenation, recent research has primarily focused on holistic feature fusion, bringing forward some effective fusion approaches that leverage deep learning and attention mechanisms. The attention-based fusion techniques focus on relevant feature patterns within multimodal data sources to detect emotions. These types of frameworks provide weights to various data modalities, enabling the network to prioritize the cues that are most significant for effective emotion recognition. Zhao et al. [19] proposed a sensor fusion module using attention mechanism to retrieve the significant scores of each sensor to create fused feature representation. An attention-based architecture in conjunction with a pre-trained CNN is developed in [20] for emotion recognition from the acoustic signals. However, the most of existing fusion techniques are one-stage fusion, which cannot clearly model the interaction between model pair of the multimodal signals.

In this article, we introduce a Multi-Stage Fusion Network (MSF-Net) for multimodal emotion recognition in scalable healthcare. In this system, the physiological signals are input into the Bi-directional Gated Recurrent Unit (Bi-GRU) network to extract the significant temporal features through both directions. Considering the visual data, the Vision Transformer (ViT) retrieves the robust visual features. At the same time, the Local Binary Pattern (LBP) and Oriented FAST and Rotated BRIEF (ORB) feature descriptors are exploited to extract the computer vision-based important features from the facial expression. To emphasize the important visual modality in the fusion procedure, we simultaneously fuse visual features extracted from the ViT module along with the LBP and ORB modules. To dynamically concentrate on more relevant features, we introduce to fuse multimodal visual features through attention mechanism, called triplet attention module that fuses the visual signals using three branch architecture. Lastly, the MSF-Net aggregates the extracted multimodal features (both temporal and visual) and efficiently recognizes the emotional categories. Finally, the developed system is tested on two publicly available popular datasets, including BioVid Emo DB as well as MGEED and obtains SOTA performance for emotion recognition. Moreover, we deploy the MSF-Net in the IoT network to monitor the patient's emotion status in scalable healthcare. To the best of our knowledge, this is the first research of proposing multi-stage feature fusion architecture to recognize emotion from multimodal data to fuse deep features and computer vision-based features by exploiting ViT along with the LBP and ORB feature descriptors. Moreover, this architecture exploits the triplet attention mechanism to exhibit more importance on the fused visual features for the first time. This proposed MSF-Net has the following main strengths:

(1) We introduce a multi-stage fusion network, called MSF-Net, to effectively leverage the visual features and temporal relationships of the data for recognizing discrete dimensional emotions in a multimodal context.
(2) We propose to exploit an approach that enables the efficient extraction of both visual features and temporal dependencies from the data, without any interference between them.
(3) We design an innovative architecture based on a triple attention technique, enabling the extraction of significant emotional feature representations by developing three attention branches to record cross-interactions among the visual features.
(4) We conduct comprehensive evaluations to investigate the performance of our proposed fusion network on two popular multimodal databases and obtain the new SOTA record. Moreover, the proposed MSF-Net is deployed in IoT network to demonstrate the scalability.

The rest of the article is structured as follows. We provide an overview of previous related studies, including recognition of emotions from physiological signals, facial cues, and multimodal data in Section 2. Section 3 demonstrates the proposed multi-stage fusion architecture for emotion detection in detail. In Section 4, we describe two multimodal experimental datasets along with the comprehensive evaluations and experimental results. At the end, Section 5 provides the conclusions of the article.

2. Related study

Emotion recognition has captured the interest of researchers throughout the globe because of its broad spectrum of applications that contributes a substantial volume of research efforts to this domain. In this section, we describe the frameworks that have been developed based on deep learning methods for emotion recognition recently. Considering the data modalities used to express emotions, emotions can be classified using unimodal and multimodal signals.

2.1. Unimodal emotion recognition

Unimodal emotion recognition employs a single data modality, including physiological signals, facial expressions, textual data, or voices [21]. Although some systems developed based on unimodal data are vulnerable to noise and variability in facial images and speech patterns, a single modality is not sufficient to meet the necessary requirements due to emotion recognition's complexity.

2.1.1. Emotion recognition from physiological data

The research on emotion recognition through physiological signals has drawn a lot of interest recently since it offers a fair assessment of the true emotional states of people. The most recent research has focused on Electroencephalogram (EEG) signals together with a few other physiological signals, including Electromyography (EMG), Electrocardiography (ECG), and Galvanic Skin Response (GSR), creating notably promising outcomes [22].

Algarni et al. [23] proposed an emotion recognition system aimed at generating more robust outcomes to facilitate accurate medical decision making. This framework retrieved the wavelet features, the Hurst exponent, and statistical patterns from the EEG signals. The Binary Gray Wolf Optimizer was exploited to select the features that enables to find out the most significant patterns. Lastly, the stacked bi-directional Long Short-Term Memory (Bi-LSTM) network recognized the emotions from the extracted features. The experimental study was carried out with a benchmark dataset called DEAP to evaluate the proposed architecture. The authors of [24] developed an EEG-based emotion recognition framework that exploited CNN to learn discriminative patterns from raw data. The collected signals are converted to spectrogram-like representations and then input into the 2D CNN to classify emotions. It is found from the experiments that the proposed architecture obtained an accuracy of 96.32% from DEAP database, and 92.54% from SEED dataset. In another study, Rahman et al. [25] presented a system to recognize emotions that converted EEG data to topographic samples that included frequency and spatial information. The two-dimensional topographic images are generated using the relative power spectral density by considering the spatial location of the EEG signals on the brain. A CNN architecture extracted features from the topographic samples and recognized the emotions. The findings revealed that the developed system appraised an accuracy of 94.63% from the SEED database. Ozdemir et al. [26] proposed an approach to recognize emotions that involved taking a series of multi-channel EEG data and transforming them into a set of multi-spectral topology images. The EEG signals are transformed into topologies based on the electrode locations. A 2D CNN extracted the features from the topological images and these high-level features are input into the LSTM to recognize the emotions. The system demonstrated an accuracy of 86.13%, 90.62%, 88.48%, and 86.23% for arousal, valence, dominance, and like–unlike emotions, respectively, from the DEAP database. Islam et al. [27] introduced a simplified system to recognize emotions that required less processing power, less processing time, and less memory. In this system, Pearson's Correlation Coefficient was exploited to transform one-dimensional EEG data into images illustrating the channel correlations of EEG sub-bands. The emotional states of the individuals are recognized through a CNN network by feeding the images into the architecture. With the DEAP dataset, the highest accuracy was achieved, measuring 78.22% for valence and 74.92% for arousal. However, the obtained accuracy falls short of the requirements for an effective real-time emotion recognition system.

2.1.2. Emotion recognition from facial cues

Emotion recognition through facial cues is a natural depiction of an individual's mental state that contains an extensive amount of emotional information. Facial Emotion Recognition (FER) analyzes facial expressions from several sources, such as images and videos, to reveal emotional feelings. Due to the advancements of multimedia technologies, FER has drawn a lot of attention from the researchers for its potential use cases [28].

Sarvakar et al. [29] proposed a method to recognize five different types of emotion from facial expressions. During the initial processing phase, the background of the image is eliminated, followed by extraction of the facial vector. The system proposed a double-layer CNN architecture to detect emotions from expressional vectors. From the experiments, the proposed scheme found relatively low performances with an accuracy of 54% from FER-2013 dataset. Minaee et al. [30] presented a DL system that exploited an attentional convolutional network for end-to-end emotion recognition in facial expressions. The attention mechanism used in this research focused on the significant parts of the face area for accurate emotion recognition. The proposed approach developed a DL model with fewer than ten layers that achieved an accuracy of 70.02% on the FER-2013 database. In another study, Chowdary et al. [31] demonstrated a transfer learning-based method for facial emotion recognition utilizing four CNN pre-trained architectures: MobileNet, ResNet50, Inception V3, and VGG19. The Fully Connected (FC) layers of the pre-trained CNNs were removed and new FC layers adapted to the particular requirements of the task were added. The performance was evaluated on 918 images from the Cohn Kanade (CK+) database and obtained the average accuracy of 96% to recognize emotion. Huang et al. [32] introduced a facial emotion recognition framework utilizing CNN architecture that combined squeeze-and-excitation network and the residual neural model. The feature maps are generated from the residual blocks that shows the significant facial landmarks which are nose and mouth. Two popular facial emotion recognition databases, including AffectNet and RAF-DB are used to evaluate the proposed architecture. The developed system was found to obtain an accuracy of 56.54% from the AffectNet database and 65. 67% from the RAF-DB database for emotion detection. Meena et al. [33] developed a solution for facial emotion recognition through CNN that can handle a large amount of signals. For better results, this research suggested to use a large batch size, adding more convolutional layers, including a robust dropout mechanism, and running a large number of epochs. The proposed system is evaluated with two benchmark databases: CK+ and FER-2013, and achieved an accuracy of 79% and 95% for the FER-2013 and CK+ datasets, respectively.

2.2. Multimodal emotion recognition

Nowadays, multimodal emotion recognition has received a great deal of attention, demonstrating that it is a key approach for understanding human emotions. There has been substantial advancement in this area for practical applications as the researchers have proposed a number of approaches to achieve great performances in the recognition of emotion from the multimodal signals [34].

Muhammad et al. [35] described an emotion recognition framework based on Deep Canonical Correlation Analysis (DCCA) using the combination of EEG and facial sequences. In this study, CNN-based ResNet50 extracted the features from facial data and 1D-CNN was utilized to retrieve high-level patterns from the EEG signals. The highly correlated features are fused using the DCCA technique, and sent to the SoftMax classifier for emotion recognition. In the MAHNOB-HCI dataset, the accuracy was 93. 86%, and in the DEAP dataset, it was 91.54% from the experiments. Similarly, Aguiñaga et al. [36] introduced a two-stage DL architecture to recognize emotion through the correlation of facial expressions and brain data. The method leveraged facial expressions as identifiers to derive attributes from EEG signals, then combined the two modalities to determine emotions. CNN architecture is used as the backbone for the experiments, while the 2D CNN is used to recognize facial expression, and the 1D-CNN is used for EEG analysis. The recognition rate for facial expression detection is 88.2%, and 89.6% for the brain signals analysis. Tan et al. [37] fused two data modalities, including EEG and facial cues to detect emotions for the application of Human-Robot Interaction (HRI). All features of both EEG and facial images are extracted from CNN network. Moreover, the Monte Carlo technique was utilized to combine the outcomes and tackle the issue presented by the small dataset size. In this framework, the recognition rate was appraised near 83.33% while experimenting with HRI application. In another study, Mocanu et al. [38] developed an end-to-end deep learning technique to recognize emotion, integrating speech and video fusion to exploit the mutually complementary aspects of high-level features. The proposed model combined 2D and 3D CNNs in a single space to record intra-modal features and efficiently detected

salient correlations among different data modes. Moreover, the inter-modal features are captured through audio-video cross-attention fusion approach. The experimental study revealed that the proposed fusion technique achieved accuracy scores of 89. 35% from the RAVDESS dataset and 84. 57% from the CREMA-D dataset. Mamieva et al. [39] used facial and speech features to detect emotions through attention-based fusion approach. The facial features from the videos are extracted using a CNN architecture, and the speech encoder utilized MFCC to retrieve spectral features via CNN. The output from two data modalities is fused to generate a multimodal feature vector through the attention network. The weighted accuracy and F1-Score of 74.6% and 66.1% are found from the IEMOCAP dataset, while the CMU-MOSEI database achieved a weighted accuracy and F1-Score of 80.7% and 73.7%.

In another work, Middya et al. [40] studied model-level fusion techniques to explore the best model to recognize emotion, integrating speech and video signals. The fusion of flattened audio and video features is performed at the model level. The combined feature vectors are sent to the FC layer of a CNN architecture for possible emotion recognition. The model-level and decision-level fusion method obtained an accuracy of 86% and 79% from the RAVDESS database. Hu et al. [41] developed a graph-based dynamic fusion technique to incorporate data from audio, video, and textual modalities. This approach analyzed the single-modal and cross-modal relations in a graph architecture to improve the integration of information. The module captured contextual information dynamics across different semantic spaces, therefore minimizing redundancy and enhancing complementarity among different data modes. The extensive experiments on IEMOCAP and MELD datasets found an accuracy of 68.21%, and 62.49%, respectively. Tzirakis et al. [42] fused three data modalities including raw texts, audio, and video to recognize multimodal emotions. The proposed architecture extracted the sentence semantics through a transformer network. Besides, a CNN architecture combining with the attention mechanism retrieved visual descriptors. At the same time, the LSTM network estimated the temporal dynamics of the speech and video data. The recognition rates for arousal and valence are 69% and 78.3% from the SEWA dataset, as evident from the evaluation studies. A transformer-based fusion method is presented in [43] to create a joint representation to recognize multi-label emotion from audio, video, and texts. In this system, the label-level representation technique is exploited to enhance the recognition performances and deal with the multi-label classification. The average accuracy scores are 85.9%, and 67.8% are assessed for the IEMOCAP and CMU-MOSEI databases, respectively, for emotion recognition. Fan et al. [44] developed a pairwise fusion technique for emotion recognition through dialog and event relation-aware graph convolutional neural network in conversations. The system provides an in-depth understanding of the emotional context by allowing each modality to present distinct features. A density loss is utilized to eliminate redundancy in pairwise fusion approaches by evaluating the density of fused features during training. The best accuracy and F1-Score of 69.57% and 69.34% are found from the experiments using IEMOCAP dataset. Amanda, Adaptively Modality-Balanced Domain Adaptation method is proposed in [45] to detect human emotions from multimodal signals. The proposed approach independently learnt modality representations and synchronized the source and target domains by correlating their representations and label distributions. The system adaptively adjusted the weights of the modalities by minimizing the representative alignment losses. The evaluations revealed that Amanda achieved the highest F1-Score of 64.82% using the combinations of IEMOCAP and MSP-IMPROV datasets for emotion recognition.

Although several studies have been conducted recently, multimodal emotion recognition still faces various challenges. One of the main challenges is efficiently fusing features from multiple sources to retrieve complimentary and redundant data. Another difficulty is the cross-dimensional interactions between spatial and temporal features due to the complex nature of these relationships. The proposed MSF-Net addresses these challenges by introducing multi-stage fusion that is able to extract significant features from each data modality using ViT, LBP, and ORB for facial features and Bi-GRU for physiological data. The fusion of visual feature extraction methods allows the system to retrieve diverse features, such as global contextual and local texture features. Moreover, the fused visual features are passed through the triplet attention mechanism that captures cross-dimensional interactions using three parallel branches to improve feature discriminability. In addition, the temporal dynamics are modeled by the Bi-GRU architecture in physiological signals, synchronizing the spatial features from visual data.

3. Proposed multi-stage fusion architecture

3.1. Overview of MSF-Net architecture

The proposed multi-stage fusion architecture aims to recognize emotion from the multimodal signals in the scalable healthcare environment. The developed emotion monitoring system comprises of several requisites as illustrated in Fig. 1. The major components are the following: Internet of Things (IoT) devices, deep learning architectures to develop fusion model, cloud server, communication technologies, and user interfaces both for physicians and patients. The end-user devices collect the data using several sensors, including camera, EEG, ECG, EMG, and skin sensor. The multimodal data are recorded in the shape of time series (physiological data) and facial videos. The collected signals are directed to the cloud server using short-range communication (Bluetooth) through the IoT hub. In the cloud server, the proposed multi-stage fusion architecture is deployed that is able to receive both physiological signals and visual data. The weights of the trained model are stored in the cloud to recognize patient's emotions depending on the input data. The service managers utilize the inference results to monitor the patient's emotional status to provide suitable services that are tailored for the patient's current conditions. At the same time, the decisions from the fusion architecture are forwarded to the patient monitoring interfaces through 5G communication networks.

An overview of the proposed multi-stage fusion architecture that is deployed in the cloud server is shown in Fig. 2. The fusion architecture contains several modules, including Bi-GRU network, vision transformer, LBP, ORB, triplet attention module, and classification module. The visual data are fed into the ViT, LBP, and ORB architectures simultaneously to retrieve the significant deep and computer vision-based features, F_{ViT}, F_{LBP}, and F_{ORB}, respectively. The retrieved features are fused F_{visual} and input into the triplet attention to refine the most important visual features $F_{visual\_refined}$. In Bi-GRU network, the temporal features $F_{temporal}$ are captured from the physiological data using the two-way connection of Bi-GRU's. Finally, the extracted features of two different branches, $F_{visual\_refined}$ and $F_{temporal}$ are concatenated to generate the final feature set F_{fuse} and fed into the classification module to recognize emotion. The mathematical flow of the proposed MSF-Net for emotion recognition is shown in Algorithm 1.

3.2. Visual feature extraction modules

3.2.1. Vision transformer

The Vision Transformer (ViT) [46] has demonstrated impressive image recognition capability by using a transformer architecture (as shown in Fig. 2) that receives a series of image patches as input. For 2D image input, the ViT architecture splits the original image $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{C\times H\times W}$ into a sequence of flattened 2D $\mathcal{M}$ fixed-size image patches $\boldsymbol{\mathcal{X}}' \in \mathbb{R}^{\mathcal{M}\times P^2 C}$, where $\mathcal{M} = HW/P^2$. Here, P = patch size, H = height, W = width, and C = number of channels in the original image. The obtained sequence of vectors $\mathcal{X}'$ is sent across a trainable projection layer to form a 1-dimensional facial feature embedding $\mathcal{Z}_0$ as shown in (1). A learnable positional embedding is integrated to each feature embedding

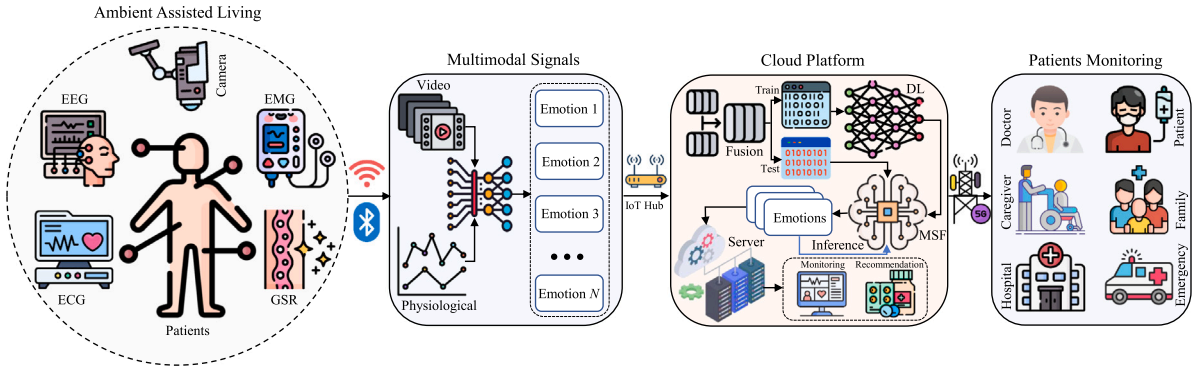

Fig. 1. A framework of the proposed architecture for emotion recognition from the multimodal data in scalable healthcare. The visual and physiological signals are collected from the patients in ambient assisted living environment through multiple sensors. In the cloud platform, the MSF-Net developed using deep learning tools is deployed for efficient emotion recognition. The patients status is informed to the immediate mediators through proper communication channel.

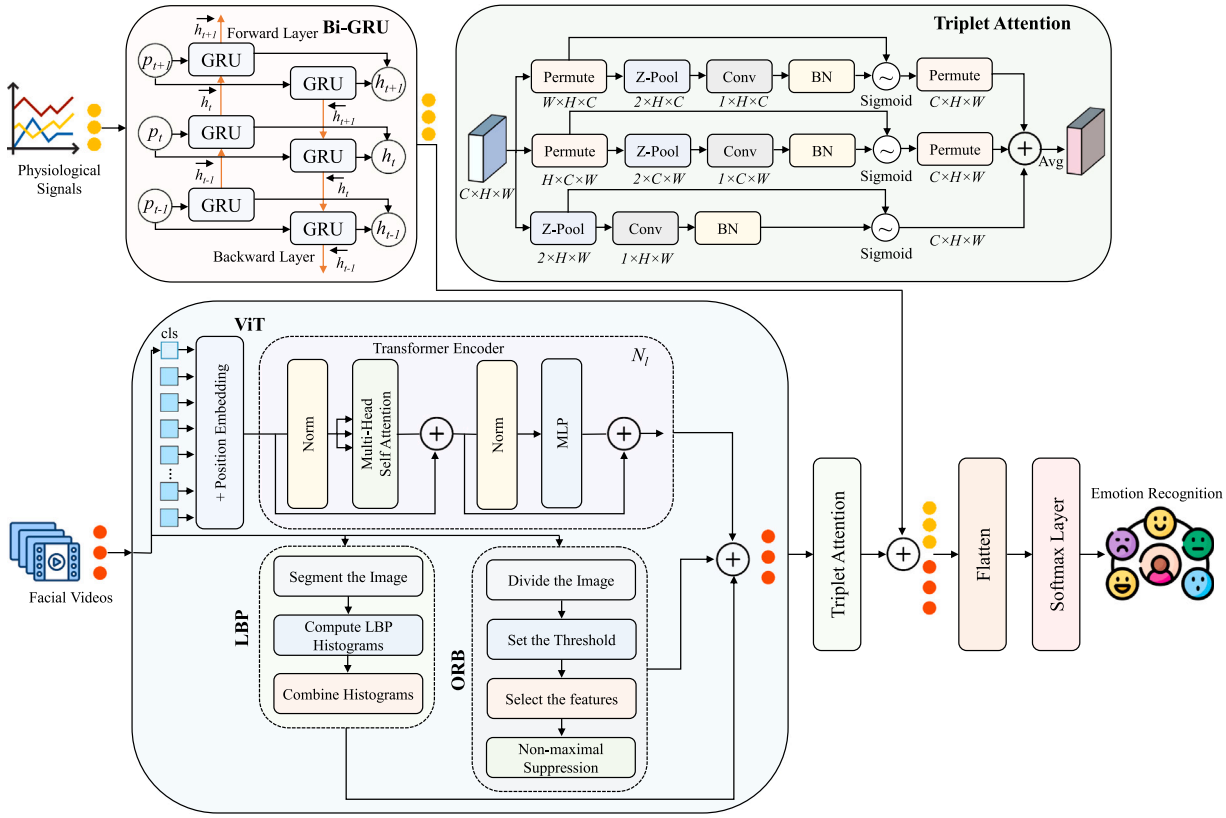

Fig. 2. An overview of the proposed MSF-Net for multimodal emotion recognition. MSF-Net includes several stages: visual feature extraction, triplet attention, temporal feature extraction, and emotion recognition. The visual feature extraction module comprises ViT, LBP and ORB methods that simultaneously extract the deep and computer vision-based features from the facial expression. The retrieved features are fused in the visual feature extraction module and sent to the triplet attention module. The triplet attention exploits three attention branches to capture the cross-interaction among the extracted signals. On the other hand, MSF-Net encodes the physiological signals with Bi-GRU to obtain the temporal features. Afterwards, the extracted features from two branches are fused to achieve the feature representation. Finally, we classify the final feature representation in the emotion classifier.

to represent the positional information of the image samples.

$$\mathcal{Z}_0 = \left[\mathcal{X}_{\text{cls}}; \mathcal{X}_p^1\mathcal{E}; \mathcal{X}_p^2\mathcal{E}; \ldots; \mathcal{X}_p^n\mathcal{E}\right] + \mathcal{E}_{\text{pos}}, \quad \mathcal{E} \in \mathbb{R}^{D\times(\mathcal{P}^2\cdot C)}, \quad \mathcal{E}_{\text{pos}} \in \mathbb{R}^{D\times(\mathcal{M}+1)} \tag{1}$$

Here, $\mathcal{E}_{\text{pos}}$ represents the position embedding, $\mathcal{X}_{\text{cls}}$ is the learnable embedding, $\mathcal{E}$ represents the linear projection, and $\mathcal{Z}_0$ is the generated position-aware feature sequence.

The next component of the ViT architecture is the transformer encoder, which includes the normalization layer, the Multi-Head Self-Attention (MHSA) mechanism, and the MLP with the residual connection. We input the facial feature embeddings $\mathcal{Z}_0$ into the encoder to design the complex interactions among the features. The MHSA module includes multiple Self-Attentions (SAs) which helps the encoder to calculate the weights of embeddings $\mathcal{Z}_0$. The SA transforms queries and key–value pairs into arrays of query ($\mathcal{Q}$), key ($\mathcal{K}$), and value ($\mathcal{V}$) matrices of dimension d_k, and computes attention. The computing

Algorithm 1: Multi-Stage Fusion Network for Multimodal Emotion Recognition.

Input: Physiological Data D_{ps} and Facial Videos D_{fv}
Output: Classifying the Human Emotions $\hat{y}$
Pre-Initialization:
$D_{ps} \leftarrow$ Load Physiological Signals;
$D_{fv} \leftarrow$ Load Facial Videos;
$ViT \leftarrow$ Load Vision Transformer;
$LBP \leftarrow$ Load Local Binary Pattern;
$ORB \leftarrow$ Load Oriented FAST and Rotated BRIEF;
$TA \leftarrow$ Load Triplet Attention;
$Bi\text{-}GRU \leftarrow$ Load Bi-GRU;
while $D_{ps} \neq 0$ ***and*** $D_{fv} \neq 0$ **do**
  for *each frame* F *in* D_{fv} **do**
    if $F^{(t)} == D_{ps}^{(t)}$ **then**
      Generate video frames: $F_t \leftarrow F^{(t)}$;
      Generate physiological sequences: $P_t \leftarrow D_{ps}^{(t)}$;
    end
  end
  Extract ViT features: $F_{ViT} \leftarrow ViT(F_t)$;
  Extract LBP features: $F_{LBP} \leftarrow LBP(F_t)$;
  Extract ORB features: $F_{ORB} \leftarrow ORB(F_t)$;
  Concatenate visual features: $F_{visual} \leftarrow F_{ViT} \oplus F_{LBP} \oplus F_{ORB}$;
  Refine visual features: $F_{visual\_refined} \leftarrow TA(F_{visual})$;
  Extract temporal features: $F_{temporal} \leftarrow Bi\text{-}GRU(P_t)$;
  Fuse features: $F_{fuse} \leftarrow F_{visual\_refined} \oplus F_{temporal}$;
  Classify emotion: $\hat{y} \leftarrow Softmax(FC(F_{fuse}))$;
end

process is stated in (2).

$$\mathrm{SA}(\mathcal{Q},\mathcal{K},\mathcal{V}) = \mathrm{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathcal{T}}}{\sqrt{d_k}}\right)\mathcal{V} \tag{2}$$

Here, $\mathcal{Q} = \mathcal{X}_i'\mathcal{W}^{\mathcal{Q}}$, is the query, $\mathcal{K} = \mathcal{X}_i'\mathcal{W}^{\mathcal{K}}$ represents the key, and $\mathcal{V} = \mathcal{X}_i'\mathcal{W}^{\mathcal{V}}$ denotes the value. All the vectors are generated through linear transformations of $\mathcal{X}_i'$. Moreover, the MHSA enables transformer architecture to learn several dependencies from different subspaces. The MHSA network divides the input into many embedding subspaces, evaluate them separately in parallel utilizing various SAs, and then combines the outcomes for the final result. The computation procedure for the MHSA is described in (3).

$$\begin{aligned}\mathrm{MHSA}(\mathcal{Q},\mathcal{K},\mathcal{V}) &= \mathrm{Concat}(h_1,h_2,\ldots,h_{\mathcal{H}})\mathcal{W}^{\mathcal{O}}\\ h_i &= \mathrm{SA}(\mathcal{Q}_i,\mathcal{K}_i,\mathcal{V}_i),\ i=1\ldots\mathcal{H}\\ &= \mathrm{SA}(\mathcal{X}_i^t\mathcal{W}^{\mathcal{Q}_i},\mathcal{X}_i^t\mathcal{W}^{\mathcal{K}_i},\mathcal{X}_i^t\mathcal{W}^{\mathcal{V}_i}),\ i=1\ldots\mathcal{H}\end{aligned} \tag{3}$$

Here, the projections $\mathcal{W}^{\mathcal{O}}$, $\mathcal{W}^{\mathcal{Q}_i}$, $\mathcal{W}^{\mathcal{K}_i}$, and $\mathcal{W}^{\mathcal{V}_i}$ are parameters' matrices.

In the transformer structure, Layer Normalization (LN) is employed before MHSA and MLP, with residual connections added to each SA and MLP layer to enhance trainability. Each transformer encoder consists of $\mathcal{N}_l$ layers of MHSA blocks, and the output of each encoder is calculated using (4) and (5). Here, $\mathcal{Z}_l'$ is the intermediate output and $\mathcal{Z}_l$ represents the final output at layer l. The MLP architecture comprises two FC layers and GELU non-linearity activation function.

$$\mathcal{Z}_l' = \mathrm{MHSA}(\mathrm{LN}(\mathcal{Z}_{l-1})) + \mathcal{Z}_{l-1}, \quad l = 1,2,\ldots,\mathcal{L} \tag{4}$$

$$\mathcal{Z}_l = \mathrm{MLP}(\mathrm{LN}(\mathcal{Z}_l')) + \mathcal{Z}_l', \quad l = 1,2,\ldots,\mathcal{L} \tag{5}$$

In our proposed MSF-Net, the original image size is set to 128×128. To deploy the ViT architecture, we set the patch size of 16×16 to divide the original image into patches, which are then embedded into a 768-dimensional space. The proposed architecture contains 12 transformer encoders, each with 12 attention heads to learn the important features from the given signals. Lastly, the hidden dimension of the MLP is set to 4608 in this study.

3.2.2. Local binary pattern

Local Binary Pattern (LBP) [47] is a textual feature descriptor (as illustrated in Fig. 2) that represents the spatial structure of images at the local level. The texture features of the images are obtained by varying the circle's radius and the pixel number. The binary operator assigns labels to the image pixels by evaluating the center pixel value with the 3×3 adjacent pixels, generating an 8-bit binary number that is then transformed into a decimal value. The formulas for calculating the LBP code for an input pixel are shown in (6) and (7).

$$\mathrm{LBP}_{\mathcal{P},\mathcal{R}}(x,y) = \sum_{n=0}^{\mathcal{P}-1} 2^n S(i_n - i_{x,y}) \tag{6}$$

$$S(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{7}$$

Here, $\mathrm{LBP}_{\mathcal{P},\mathcal{R}}(x,y)$ denotes the LBP texture feature centered at pixel (x, y), $\mathcal{R}$ represents the radius, $\mathcal{P}$ indicates the number of neighboring pixels, i_n signifies the gray value of nth neighboring pixel, and $i_{x,y}$ denotes the gray value of the center pixel.

The output is a binary number that represents the local texture adjacent to the center pixel. The LBP values are utilized to generate histograms that depict the textural patterns in the images.

3.2.3. Oriented FAST and rotated BRIEF

In this paper, we exploit the region-based Oriented FAST and Rotated BRIEF (ORB) [48] feature descriptor (as shown in Fig. 2) that divides the region of the images in a unique way compared to conventional ORB. In the region-based ORB algorithm, the number of feature points per region is calculated by dividing the total feature values and the number of regions. In the calculation procedure of this algorithm, the input images are equally divided into $\mathcal{A} \times \mathcal{B}$ regions, where $\mathcal{A}$ indicates the number of rows and $\mathcal{B}$ denotes the number of columns. The feature points are randomly distributed in the created regions, and the regions are labeled as $\{h_1, h_2, \ldots, h_{\mathcal{A}\times\mathcal{B}}\}$. In the next operation, the threshold is computed using (8), where the number of feature points is represented by n.

$$\mathcal{T} = \frac{n}{\mathcal{A}\mathcal{B}} \tag{8}$$

We detect the feature points in each region depending on the threshold value $\mathcal{T}$. Finally, the best feature values are calculated using non-maximal suppression technique while the generated feature points are greater than n.

3.2.4. Triplet attention module

In the proposed MSF-Net, we exploit the triplet attention [49] (as depicted in Fig. 2) to extract the most significant and relevant features from visual expressions. The triplet attention module contains three parallel networks. The first network represents cross-dimensional relationships between the dimensions of the channel ($\mathcal{C}$) and height ($\mathcal{W}$). The second network addresses the interactions involving the dimensions of the channel ($\mathcal{C}$) and width ($\mathcal{H}$). The third network emphasizes on the height ($\mathcal{H}$) and width ($\mathcal{W}$) interactions, serving as spatial attention.

The first network exploits spatial attention using channel pooling (Z-Pool) on the input features $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{\mathcal{C}\times\mathcal{H}\times\mathcal{W}}$, therefore decreasing the zeroth dimension of $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{\mathcal{C}\times\mathcal{H}\times\mathcal{W}}$ to 2. The calculation method is formulated as in (9).

$$\text{Z-Pool}(\mathcal{X}) = \left[\mathrm{MaxPool}_{0d}(\mathcal{X}),\ \mathrm{AvgPool}_{0d}(\mathcal{X})\right] \tag{9}$$

Here, $0d$ represents the 0th dimension where both the maximum and average pooling are carried out.

Initially, the given feature $\mathcal{X} \in \mathbb{R}^{C\times\mathcal{H}\times\mathcal{W}}$ is rotated 90° anti-clockwise along the $\mathcal{H}$ axis that generates the feature maps in the form $(\mathcal{W}\times\mathcal{H}\times C)$. The generated tensor $(2\times\mathcal{H}\times C)$ after Z-Pool passes through a convolution layer followed by a batch normalization layer. In this article, we set the kernel size of 7×7 in the convolution layer. Now, the compressed tensor size is $(1\times\mathcal{H}\times C)$ which is fed into the Sigmoid activation to generate the optimized attention values. Finally, the original input shape $\mathcal{X} \in \mathbb{R}^{C\times\mathcal{H}\times\mathcal{W}}$ is retained by the rotation of 90° clockwise along the $\mathcal{H}$ axis. The operation of the first network is computed as in (10).

$$\mathcal{X}_{1st} = \mathcal{X}_{C\mathcal{H}} \odot \text{Sig}(\text{BN}(\text{Conv}(z^{th}_{pool}(\mathcal{X}_{C\mathcal{H}})))) \tag{10}$$

Here, $\odot$ is a dot product, BN is the batch normalization, and Sig is the Sigmoid activation function.

Similar to first network, the second network converts the given features $\mathcal{X} \in \mathbb{R}^{C\times\mathcal{H}\times\mathcal{W}}$ to $(\mathcal{H}\times C\times\mathcal{W})$ dimensions by rotating the input 90° anti-clockwise along the $\mathcal{W}$ axis. This network then performs pooling operation, followed by a 7×7 convolution along with batch normalization. The generated features then passes through the Sigmoid activation and finally converts them into a shape $\mathcal{X} \in \mathbb{R}^{C\times\mathcal{H}\times\mathcal{W}}$ by rotating 90° clockwise along the $\mathcal{W}$ axis. The overall procedure is formulated as in (11).

$$\mathcal{X}_{2nd} = \mathcal{X}_{C\mathcal{W}} \odot \text{Sig}(\text{BN}(\text{Conv}(z^{th}_{pool}(\mathcal{X}_{C\mathcal{W}})))) \tag{11}$$

The third network generates the tensor shape $(2\times\mathcal{H}\times\mathcal{W})$ after the Z-Pool operation which is then passed through the convolution and batch normalization, followed by Sigmoid function to create attention weights of $(1\times\mathcal{H}\times\mathcal{W})$. The mathematical calculation is followed by (12).

$$\mathcal{X}_{3rd} = \mathcal{X}_{\mathcal{H}\mathcal{W}} \odot \text{Sig}(\text{BN}(\text{Conv}(z^{th}_{pool}(\mathcal{X}_{\mathcal{H}\mathcal{W}})))) \tag{12}$$

Finally, the refined attention-based features $F_{visual\_refined}$ are derived by combining the features generated from the three networks for an input tensor $\mathcal{X} \in \mathbb{R}^{C\times\mathcal{H}\times\mathcal{W}}$ through average pooling. The final output representation of the triplet attention module follows (13).

$$F_{visual\_refined} = \frac{1}{3}(\mathcal{X}_{1\text{st}} + \mathcal{X}_{2\text{nd}} + \mathcal{X}_{3\text{rd}}) \tag{13}$$

3.3. Temporal feature extraction with Bi-GRU

In this article, the Bi-directional Gated Recurrent Unit (Bi-GRU) [50] is exploited to model physiological signals to extract temporal features. The Bi-GRU network (as demonstrated in Fig. 2) processes input signals in both directions, including forward and backward simultaneously, to improve sequence modeling. The Bi-GRU architecture has two GRUs: one unit analyzes the input signals from the beginning to the end, called forward direction, and the other unit processes it from the end to the beginning, called backward direction.

For time stamp t, physiological signals $p = \{p_1, p_2, p_3, \ldots, p_{T_p};\ p_t \in \mathbb{R}\}$ are input into the Bi-GRU architecture to extract temporal features. In the Bi-GRU network, the formulas for the forward and backward calculation functions are denoted as $\overrightarrow{GRU}(\cdot)$ and $\overleftarrow{GRU}(\cdot)$. The calculation procedure for the Bi-GRU network is shown in (14) and (15).

$$\vec{h}_t = \overrightarrow{GRU}(p_t, \vec{h}_{t-1}) \tag{14}$$

$$\overleftarrow{h}_t = \overleftarrow{GRU}(p_t, \overleftarrow{h}_{t+1}) \tag{15}$$

where, $\vec{h}_t$ indicates the output of the forward direction, and $\overleftarrow{h}_t$ signifies the output of the backward direction on the time stamp t. Finally, the output ($F_{temporal}$) of the state information of the Bi-GRU architecture is computed through (16) considering each time stamp t.

$$F_{temporal} = [\vec{h}_1, \overleftarrow{h}_1][\vec{h}_2, \overleftarrow{h}_2], \ldots, [\vec{h}_{T_p}, \overleftarrow{h}_{T_p}] \tag{16}$$

In the proposed fusion network, we exploited a Bi-GRU network with 2 layers, considering the input size of 3 and 92 for BioVid Emo DB and MGEED datasets. The hidden state dimension is defined to 16. After the operations, the total number of generated hidden state size is 32 due to the bi-directional property.

3.4. Feature fusion and classification

The multimodal features extracted from the Bi-GRU and the triplet attention module are fused and passed to the classification module for emotion recognition. The refined visual features $F_{visual\_refined}$ extracted from the triplet attention module and the temporal features $F_{temporal}$ obtained by the Bi-GRU network are concatenated to generate the final fused feature vector $\mathcal{F}_{fuse}$. The fusion process is shown in (17).

$$\mathcal{F}_{fuse} = F_{visual\_refined} \oplus F_{temporal} \tag{17}$$

The fused features are fed into a linear layer with residual connections and then pass towards a Softmax layer to achieve the probability distribution $\mathcal{P}$ of emotion categories (C).

$$\mathcal{S} = \mathcal{F}_{fuse} + \text{ReLU}(\mathcal{F}_{fuse}\mathcal{W}_{\mathcal{F}_{fuse}} + b_{\mathcal{F}_{fuse}}) \tag{18}$$

$$\mathcal{P} = \text{Softmax}(\mathcal{S}\mathcal{W}_s + b_s) \tag{19}$$

where $\mathcal{W}_{\mathcal{F}_{fuse}} \in \mathbb{R}^{d_{\mathcal{F}_{fuse}}\times d_{\mathcal{F}_{fuse}}}$, $b_{\mathcal{F}_{fuse}} \in \mathbb{R}^{d_{\mathcal{F}_{fuse}}}$, $W_{\mathcal{S}} \in \mathbb{R}^{d_{\mathcal{F}_{fuse}}\times d_C}$, and $b_{\mathcal{S}} \in \mathbb{R}^C$ are training parameters. Here, $\mathcal{S} \in \mathbb{R}^{\mathcal{N}\times d_{\mathcal{F}_{fuse}}}$ and $\mathcal{P} \in \mathbb{R}^{\mathcal{N}\times C}$.

We calculate the emotion label with the maximum probability using the argmax function as shown in (20).

$$\hat{y}^{(n)} = \arg\max(\mathcal{P}^{(n)}) \tag{20}$$

where $\mathcal{P}^{(n)} \in \mathbb{R}^C$ is the probability of emotion categories, and $\hat{y}^{(n)}$ denotes the emotion label predicted by the proposed architecture.

We exploit the cross-entropy loss for training. The formula for loss calculation is given below (21).

$$\mathcal{L} = -\frac{1}{\sum_{(m,n)=1}^{(\mathcal{V},\mathcal{U})}} \sum_{m=1}^{\mathcal{V}}\sum_{n=1}^{\mathcal{U}}\sum_{c=1}^{C} y^{(n)}_{m,c}\log_2(\hat{y}^{(n)}_{m,c}) \tag{21}$$

In this context, $\mathcal{V}$ and $\mathcal{U}$ are the total number of frames generated from the videos and physiological signals in the dataset. Moreover, $y^{(n)}_{m,c}$ and $\hat{y}^{(n)}_{m,c}$ represent the actual label and the predicted label of class c.

4. Experimental results and analysis

In this section, we assess the performance of MSF-Net on two multimodal emotion recognition databases, BioVid Emo DB and MGEED by comparing several SOTA emotion recognition frameworks. We performed comprehensive evaluations to validate the efficiency of our proposed approach and provide a thorough evaluation based on the experimental findings.

4.1. Experimental environment

All experiments for MSF-Net for emotion recognition were performed on a server computer with 8-core 4.00 GHz CPU of a Core i7 processor with 32 GB RAM (DDR4). The experimental setup includes two GPU cards with the specification of GTX 1080 with 8 GB of GDDR5 memory. Pytorch, a deep learning library, is exploited along with Python for all evaluation processes. The server was running an Ubuntu operating system, version 22.04 LTS.

4.2. Multimodal datasets

The experiments have been performed on two publicly available multimodal emotion recognition datasets: BioVid Emo DB and MGEED. The detailed description of these databases are mentioned below.

4.2.1. BioVid Emo DB

The BioVid Emo DB [51] dataset includes physiological data including Skin Conductance Level (SCL), Electrocardiogram (ECG), and Trapezius Electromyogram (tEMG), as well as synchronized videos. The major emotion categories in this dataset are *amusement, sad, anger, disgust,* and *fear*. The raw signals were collected from 94 participants (50 women and 44 men) by showing them 15 video clips. The data from 86 volunteers are available in the dataset due to incomplete and corrupted information. The video frame rate is 25 Hz and the resolution is 1388 × 1038 pixels. The sampling frequency of the physiological signals is 512 Hz.

4.2.2. MGEED

Multimodal Genuine Emotion and Expression Detection (MGEED) [52] dataset is made up of several data modalities including Optomyography (OMG), EEG, ECG, GSR, facial expressions, and depth maps. This is the first open-access dataset that contains OMG signals. The data collection framework is developed to collect multimodal data simultaneously from the participants. In the data collection system, 17 volunteers saw 9 emotional movies and provided their frame-by-frame emotional reactions mentioning seven basic emotions. The emotion classes are *neutral, happy, sadness, angry, fear, disgust,* and *surprise*. The extensive MGEED dataset contains 17 participants and 150,497 facial images, 147,539 depth map frames, and four physiological signal categories: 70-channel EEG, 20-channel OMG, as well as one-channel ECG and GSR data.

In the proposed MSF-Net, we conducted the experiments using both the BioVid Emo DB and MGEED dataset to evaluate the robustness of the fusion network across multiple data sources. We utilized all physiological data and facial expressions from the BioVid Emo DB database. For the MGEED dataset, we excluded the depth information to maintain the same information in both datasets. We partitioned the data in training and test sets following the ratio of 70:30. Initially, we generated frames from the facial videos at the data preprocessing stage. Afterwards, the face region is detected from the frames using opencv's Caffe model. Later, facial frames and physiological data are synchronously input into the deep learning network. The frame size is set to 128 × 128. The attributes for physiological data for BioVid Emo DB and MGEED datasets are considered to value 3 and 92 considering different signals combination, which are fed into the Bi-GRU architecture.

4.3. Performance metrics and baseline methods

4.3.1. Performance metrics

This section presents a summary of the evaluation methods utilized to evaluate the performance of the architecture described in this paper. This paper exploits four performance measures, including Accuracy (A), Precision (P), Recall (R), and F1-Score (F1) that are frequently used to evaluate the effectiveness of classification methods.

4.3.2. Baseline methods

To comprehensively assess our proposed MSF-Net, we exploit the following baselines and SOTA techniques for comparison. We also chose datasets that are extensively used in recent studies to allow for a direct comparison of our findings with them. The works presented in [53–56], and [57] used BioVid Emo DB dataset. The MGEED dataset is introduced in [52].

FedCMD [53] proposes the knowledge transfer scheme from teacher model to student model using physiological data and videos i.e., sensor to vision cross-modal distillation. The baseline model in this work is two-dimensional CNN.

MEmoR [54] exploits a fine-tuned ResNet50 for face recognition and CNN for physiological signals to analyze emotions. The data from these modalities are fused through decision-level weighted fusion to recognize emotions.

MDBN [55] optimizes and fuses the features from sensor and video data through Deep Belief Networks (DBNs). The features are retrieved via bimodal DBNs, and a linear SVM is exploited for emotion recognition.

WT-SVM [56] retrieves the high-level features from the physiological data through wavelet transform and recognizes the emotions using SVM.

TF-SVM [57] retrieves the time-frequency domain features from the sensor data and classifies the signals through SVM.

Fus-CNN [52] employs pre-trained ResNet101 model to retrieve and fuse the video frame and depth features. The physiological signals (EEG, ECG, and OMG) are synchronized, and normalized to their corresponding compact attributes. Lastly, all the extracted attributes are fused and input into classification block for recognition.

4.4. Implementation details

We exploited the validation data for hyper-parameter tuning. We utilized Adam optimizer with a learning rate of 0.001 to fine-tune the parameters. In the evaluations, we trained the models for 25 iterations. The batch size was set to 64. We implemented L_2 regularization technique considering the following parameters: momentum = 0.9, decay = 0.0005, step size = 3, and $\gamma = 0.1$ to alleviate over-fitting problems. Our proposed model was trained through cross-entropy loss as our objective. For the validation set, we chose 20% of the training data to validate the MSF-Net. For evaluation, we used all physiological signals along with facial cues of the BioVid Emo DB database. From the MGEED dataset, we utilized all physiological data and facial expressions except the depth maps to keep the same information in both datasets. In the experiments, we considered five variants of the developed MSF-Net to recognize emotion. The models vary depending on the modules used in the architecture. In all variants, physiological data are input into the Bi-GRU network to retrieve temporal dependencies. The image frames are input into the ViT, LBP, and ORB networks (depending on the modules used) and then passed directly to the next fusion stage or through the triplet attention module to retrieve the relevant patterns. Finally, the extracted attributes from both branches are fused and passed to the classification module for emotion recognition. The proposed variants of the MSF-Net are given below.

$ViT + Attention + Bi\text{-}GRU$: In this architecture, the image data are fed into the ViT module and passed through the triplet attention block.

$ViT + LBP + Attention + Bi\text{-}GRU$: In this network, the image data are sent to the ViT and LBP modules to extract the features. Afterwards, the extracted attributes from two modules are fused and passed to the triplet attention module.

$ViT + ORB + Attention + Bi\text{-}GRU$: In this variation, the ViT and ORB modules retrieve the image features and the extracted features from both branches are combined and fed into the triplet attention network.

$ViT + LBP + ORB + Bi\text{-}GRU$: In this variation, all the three modules including, ViT, LBP, and ORB are exploited to capture the features from the image frame. The three branches extracted patterns are fused and sent to the next fusion stage excluding the triplet attention block.

$ViT + LBP + ORB + Attention + Bi\text{-}GRU$: In this architecture, the ViT, LBP, and ORB blocks extract the image features and fuses them. Then the fused features are passed to the triplet attention module. All the modules proposed in our system are utilized in this architecture.

4.5. Results and comparison

To validate the robustness of the developed MSF-Net, we have conducted extensive evaluations to compare it with the SOTA algorithms on two popular databases, BioVid Emo DB and MGEED. In this part, we evaluate our proposed architecture to SOTA works on overall performance using above-mentioned evaluation metrics. Moreover, we conduct comparisons on recognition accuracy, precision, recall, and F1-Score for each emotion category.

Table 1
Performance (%) of MSF-Net against SOTA methods on the BioVid Emo DB database. The best performances are represented by bold front, while the second-best outcomes are marked with underlined text.

Methods	Modalities	Fusion	A	P	R	F1
FedCMD [53]	Image+tEMG+ECG+SCL	–	90.83	–	–	93.82
MEmoR [54]	SCL	–	63.78	–	–	68
	ECG	–	69.56	–	–	64
	tEMG	–	70.93	–	–	61
	tEMG+ECG+tEMG	Feature Fusion	73.29	–	–	76
	Image	–	68.51	–	–	65
	Image+tEMG+ECG+SCL	Early Fusion	76.32	–	–	74
		Decision Fusion	81.54	–	–	79
MDBN [55]	Image+tEMG+ECG+SCL	Feature Fusion	80.89	–	–	–
WT-SVM [56]	SCL	–	70.10	–	–	–
	ECG	–	88.76	–	–	–
	tEMG	–	80.07	–	–	–
	tEMG+SCL	Feature Fusion	63.14	–	–	–
	ECG+SCL		73.90	–	–	–
	tEMG+ECG		90.58	–	–	–
	tEMG+ECG+SCL		63.14	–	–	–
	tEMG+SCL	Decision Fusion	84.94	–	–	–
	ECG+SCL		87.76	–	–	–
	tEMG+ECG		94.81	–	–	–
	tEMG+ECG+SCL		85.19	–	–	–
TF-SVM [57]	ECG	Feature Fusion	79.51	–	–	–
ViT+Attention+Bi-GRU	Image+tEMG+ECG+SCL	Feature Fusion	<u>96.46</u>	<u>96.42</u>	96.39	<u>96.40</u>
ViT+LBP+Attention+Bi-GRU			96.37	96.31	96.28	96.30
ViT+ORB+Attention+Bi-GRU			96.39	96.33	96.33	96.33
ViT+LBP+ORB+Bi-GRU			<u>96.46</u>	96.40	<u>96.40</u>	<u>96.40</u>
ViT+LBP+ORB+Attention+Bi-GRU			**96.66**	**96.63**	**96.56**	**96.60**

Table 2
Performance (%) of MSF-Net against SOTA methods on the MGEED database. The best performances are represented by bold front, while the second-best outcomes are marked with underlined text.

Methods	Modalities	Fusion	A	P	R	F1
Fus-CNN [52]	RGB (weighted loss)	Feature Fusion	20.59	–	–	–
	RGB (over-sampling)		22.00	–	–	–
	RGB+Depth		20.69	–	–	–
	RGB+ECG		19.58	–	–	–
	OMG+EEG		26.77	–	–	–
	RGB+EEG+OMG		29.24	–	–	–
ViT+Attention+Bi-GRU	RGB+OMG+EEG+ECG+GSR	Feature Fusion	93.87	91.90	91.60	91.70
ViT+LBP+Attention+Bi-GRU			<u>94.02</u>	92.12	91.45	91.72
ViT+ORB+Attention+Bi-GRU			93.99	92.29	**92.03**	<u>92.10</u>
ViT+LBP+ORB+Bi-GRU			93.99	<u>92.54</u>	91.70	92.05
ViT+LBP+ORB+Attention+Bi-GRU			**94.26**	**92.81**	<u>91.86</u>	**92.24**

4.5.1. Comparison on overall performance

Tables 1–2 present the comparative analysis of the developed MSF-Net and other SOTA networks for emotion recognition on BioVid Emo DB and MGEED datasets, respectively. For the BioVid Emo DB dataset, our proposed MSF-Net ($ViT+LBP+ORB+Attention+Bi\text{-}GRU$) obtained superior performances: accuracy of 96.66%, precision of 96.63%, recall of 96.56%, and F1-Score of 96.60%. The findings demonstrate an improvement of 1.85% on accuracy and 2.78% on F1-Score over existing multimodal emotion recognition methods, as evident in Table 1. For MGEED dataset, the proposed $ViT + LBP + ORB + Attention + Bi\text{-}GRU$ architecture achieved the highest accuracy, precision, and F1-Score of 94.26%, 92.81%, and 92.24%, respectively. Table 2 illustrates the increase in accuracy of 65.02% over the work presented in the original MGEED dataset. The best recall of 92.03% is revealed from the $ViT + ORB + Attention + Bi\text{-}GRU$ model from the MGEED dataset. The second-best accuracy: 96.46% and F1-Score: 96.40% are evaluated in the variants of $ViT + Attention + Bi\text{-}GRU$ and $ViT + LBP + ORB + Bi\text{-}GRU$ for the BioVid Emo DB database. On the same dataset, the second-best precision of 96.42% and recall of 96.40% are found from the $ViT + Attention + Bi\text{-}GRU$ and $ViT+LBP+ORB+Bi\text{-}GRU$ models. Considering the MGEED dataset, the variants $ViT + LBP + Attention + Bi\text{-}GRU$, $ViT + LBP + ORB + Bi\text{-}GRU$, $ViT + LBP + ORB + Attention + Bi\text{-}GRU$, and $ViT + ORB + Attention + Bi\text{-}GRU$ revealed the second-best accuracy of 94.02%, precision of 92.54%, recall of 91.86%, and F1-Score of 92.10%, respectively. It is found from Table 1 that the works presented in FedCMD [53], MEmoR [54], and MDBN [55] used full set of data from the BioVid Emo DB dataset. The accuracy obtained from these SOTA methods is 90.83%, 81.54%, and 80.89%, respectively. The other SOTA approaches, WT-SVM [56] and TF-SVM [57] used only the physiological signals available in the BioVid Emo DB dataset. The highest accuracy of 94.81% is obtained in WT-SVM [56] through decision-level fusion using $tEMG + ECG$ signals. In Table 2, the best accuracy of 29.24% is achieved in Fus-CNN [52] using $RGB + EEG + OMG$ signals through the feature fusion approach. Our proposed multi-stage fusion architecture obtained superior performances over the SOTA techniques. The reason behind that we fused the deep features captured from the transformer and the significant patterns from the computer vision-based features. Moreover, we extract the most significant visual features through triplet attention along with the deployment of multi-stage feature fusion to exploit considerably fine-grained multimodal interactions.

It is evident from Tables 1 and 2 that the proposed MSF-Net obtained the best performances both for BioVid Emo DB and MGEED datasets for emotion recognition. While removing the triplet attention module, the accuracy is decreased by 0.2% compared to the proposed MSF-Net for the BioVid Emo DB dataset. On the MGEED dataset, the

Table 3
Performance (%) comparison of different approaches on the BioVid Emo DB dataset with multi-modality. We provide each emotion categories classification performances. The best performances are represented by bold front, while the second-best results are marked with underlined text.

Methods	Amusement				Anger				Disgust				Fear				Sad			
	A	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*
MEmoR [54]	91.02	–	–	–	75.31	–	–	–	77.64	–	–	–	80.38	–	–	–	79.92	–	–	–
MDBN[a]	56.19	–	–	–	87.68	–	–	–	27.60	–	–	–	71.93	–	–	–	42.15	–	–	–
MDBN[b]	64.70	–	–	–	57.32	–	–	–	12.21	–	–	–	43.75	–	–	–	93.07	–	–	–
MDBN[c]	80.10	–	–	–	53.65	–	–	–	14.08	–	–	–	67.33	–	–	–	57.64	–	–	–
MDBN[d]	81.86	–	–	–	80.03	–	–	–	22.73	–	–	–	76.95	–	–	–	79.55	–	–	–
MDBN[e]	86.35	–	–	–	87.21	–	–	–	33.57	–	–	–	70.20	–	–	–	85.82	–	–	–
MDBN[f]	89.25	–	–	–	90.74	–	–	–	52.60	–	–	–	86.45	–	–	–	85.42	–	–	–
ViT+Attention+Bi-GRU	95.68	96.05	95.68	95.86	96.04	96.04	96.04	96.04	97.20	96.92	97.20	97.06	96.30	96.57	96.30	96.43	96.72	96.49	96.72	96.61
ViT+LBP+Attention+Bi-GRU	95.53	95.78	95.53	95.66	95.67	96.01	95.67	95.84	97.24	97.10	97.24	97.17	96.17	96.33	96.17	96.25	96.77	96.33	96.77	96.55
ViT+ORB+Attention+Bi-GRU	95.81	95.59	95.81	95.70	95.83	96.48	95.83	96.15	97.34	96.88	97.34	97.11	96.04	96.21	96.04	96.12	96.61	96.48	96.61	96.55
ViT+LBP+ORB+Bi-GRU	95.79	95.85	95.79	95.82	95.93	95.93	95.93	95.93	97.30	97.16	97.30	97.23	96.25	96.54	96.25	96.39	96.70	96.52	96.70	96.61
ViT+LBP+ORB+Attention+Bi-GRU	95.61	96.23	95.61	95.92	96.05	96.50	96.05	96.28	**97.65**	**97.26**	**97.65**	**97.45**	96.49	96.59	96.49	96.54	96.99	96.52	96.99	96.75

[a] =tEMG.
[b] =ECG.
[c] =SCL.
[d] =combine feature.
[e] =DBN feature.
[f] =MDBN feature.

[a–f] results come from [55].

Table 4
Performance (%) comparison of different approaches on the MGEED dataset with multi-modality. We provide each emotion categories classification performances. The best performances are represented by bold front, while the second-best results are marked with underlined text.

Methods	Neutral				Happy				Sad				Angry				Fear				Disgust				Surprise			
	A	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*
Fus-CNN[a]	25.73	–	–	–	19.61	–	–	–	7.32	–	–	–	0	–	–	–	30.15	–	–	–	5.35	–	–	–	26.32	–	–	–
Fus-CNN[b]	27.13	–	–	–	24.49	–	–	–	7.39	–	–	–	1.60	–	–	–	27.98	–	–	–	15.30	–	–	–	25.21	–	–	–
Fus-CNN[c]	25.34	–	–	–	23.07	–	–	–	7.27	–	–	–	1.60	–	–	–	28.59	–	–	–	8.89	–	–	–	25.21	–	–	–
Fus-CNN[d]	22.18	–	–	–	14.48	–	–	–	20.55	–	–	–	13.44	–	–	–	32.70	–	–	–	8.84	–	–	–	25.08	–	–	–
Fus-CNN[e]	25.73	–	–	–	20.56	–	–	–	22.14	–	–	–	20.72	–	–	–	55.96	–	–	–	30.17	–	–	–	27.13	–	–	–
Fus-CNN[f]	31.04	–	–	–	37.56	–	–	–	20.36	–	–	–	25.29	–	–	–	28.52	–	–	–	22.34	–	–	–	26.82	–	–	–
ViT+Attention+Bi-GRU	91.72	91.14	91.72	91.43	95.35	94.75	95.35	95.05	95.87	96.11	95.87	95.99	91.39	85.43	91.39	88.31	95.53	95.32	95.53	95.42	77.49	86.37	77.49	81.69	93.82	94.12	93.82	93.97
ViT+LBP+Attention+Bi-GRU	92.32	91.19	92.32	91.75	95.57	95.07	95.57	95.32	95.81	96.38	95.81	96.10	90.23	85.65	90.23	87.88	95.34	95.34	95.34	95.34	76.31	87.36	76.31	81.46	94.54	93.79	94.54	94.17
ViT+ORB+Attention+Bi-GRU	92.13	91.83	92.12	91.98	95.20	94.73	95.20	94.96	96.17	96.07	96.16	96.12	93.26	86.60	93.25	89.80	95.21	95.11	95.21	95.16	78.56	88.31	78.56	83.15	93.67	93.36	93.66	93.51
ViT+LBP+ORB+Bi-GRU	91.82	91.65	91.82	91.74	95.29	94.71	95.29	95.01	96.23	96.17	**96.23**	96.20	91.39	88.11	91.39	89.72	95.28	95.13	95.28	95.20	77.06	89.09	77.06	82.64	94.79	92.89	94.78	93.83
ViT+LBP+ORB+Attention+Bi-GRU	92.68	91.61	92.68	92.14	**95.90**	95.26	95.90	95.58	95.84	**96.60**	95.84	**96.22**	91.86	87.38	91.86	89.56	95.65	95.23	95.65	95.44	76.42	90.02	76.42	82.66	94.62	93.50	94.62	94.06

[a] =RGB (weighted loss).
[b] =RGB (over-sampling).
[c] =RGB + Depth.
[d] =RGB + ECG.
[e] =OMG + EEG.
[f] =RGB + EEG + OMG.

[a–f] results come from [52].

accuracy and F1-Score have been decreased by 0.27% and 0.19%, respectively. By varying the LBP and ORB modules, the accuracy for the BioVid Emo DB dataset is reduced by 0.27% and 0.29%, while the highest recall of 92.03% and the second-best accuracy of 94.02% are found by excluding LBP and ORB modules for MGEED dataset. Ignoring the blocks associated with computer-vision-based features makes the obtained performances comparatively better considering the BioVid Emo DB dataset. However, the accuracy and F1-Score are decreased for the MGEED dataset by 0.39% and 0.54% by removing LBP and ORB modules from the proposed MSF-Net architecture.

4.5.2. Comparison on performance for each class

Tables 3–4 demonstrate class-wise emotion recognition accuracy, precision, recall, and F1-Score for the BioVid Emo DB and MGEED dataset. From Table 3, it is shown that the best accuracy, precision, recall, and F1-Score of 97.65%, 97.26%, 97.65%, and 97.45%, respectively, are found for *disgust* emotion from the $ViT + LBP + ORB + Attention + Bi\text{-}GRU$ model. The second-best accuracy of 97.34% and recall of 97.34% are also revealed in the *disgust* emotion for the $ViT + ORB + Attention + Bi\text{-}GRU$ architecture. For the same emotion category, the second-best precision and recall are found considering the values 97.10% and 97.17%. For some variations of the models, a few of the emotions got relatively lower performances. The least accuracy of 95.67% for *anger*, recall of 95.53% for *amusement*, and F1-Score of 95.66% for *amusement* are obtained from the $ViT + LBP + Attention + Bi\text{-}GRU$ network. The $ViT+ORB+Attention+Bi\text{-}GRU$ model found the lowest precision of 95.59% for *amusement* emotion. The findings from Table 4 demonstrated that the *sad* emotion scored the highest precision: 96.60% and F1-Score: 96.22% from the $ViT+LBP+ORB+Attention+Bi\text{-}GRU$ network. Moreover, the top accuracy of 95.90% and recall of 96.23% are achieved for *happy* and *sad* emotions from the $ViT+LBP+ORB+Attention+Bi\text{-}GRU$ and $ViT+LBP+ORB+Bi\text{-}GRU$ architectures. Similarly, the second-highest performances are appraised for *happy* and *sad* emotions considering the different variations of the MSF-Net. In the case of lower performances, the $ViT + LBP + Attention + Bi\text{-}GRU$ network recorded the least accuracy of 76.31%, recall of 76.31%, and F1-Score of 81.46% for *disgust* emotion. The *angry* emotion scored the lowest precision of 85.43% from the $ViT + Attention + Bi\text{-}GRU$ model.

4.6. Visual results

In this part, we provide several examples of the sub-views to demonstrate the intuitive outcomes. Figs. 3 and 4 show the confusion matrices to represent the distribution of several emotion categories on two different datasets. Overall, the proposed MSF-Net demonstrates superior predictive performance compared to the SOTA methods on different

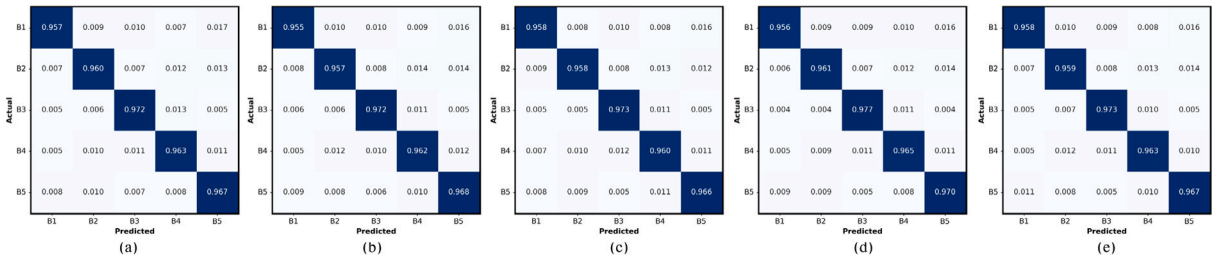

Fig. 3. Visualization of confusion matrices of MSF-Net on the testing set of the BioVid Emo DB database. The columns represent the actual labels, while the rows represent the predicted ones. (a) ViT+Attention+Bi-GRU (b) ViT+LBP+Attention+Bi-GRU (c) ViT+ORB+Attention+Bi-GRU (d) ViT+LBP+ORB+Bi-GRU (e) ViT+LBP+ORB+Attention+Bi-GRU. Here, B1: Amusement, B2: Anger, B3: Disgust, B4: Fear, and B5: Sad.

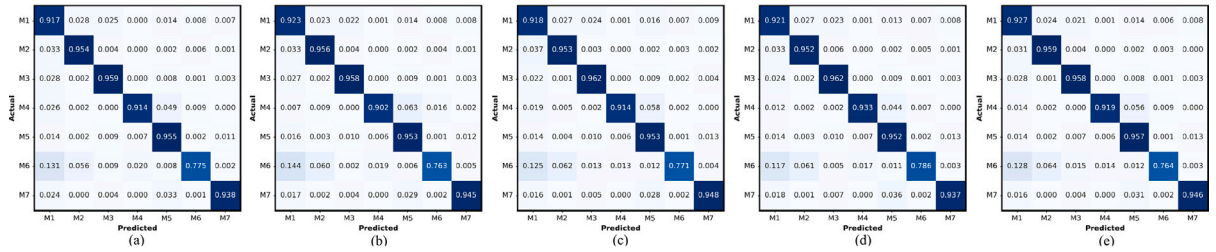

Fig. 4. Visualization of confusion matrices of MSF-Net on the testing set of the MGEED database. The columns represent the actual labels, while the rows represent the predicted ones. (a) ViT+Attention+Bi-GRU (b) ViT+LBP+Attention+Bi-GRU (c) ViT+ORB+Attention+Bi-GRU (d) ViT+LBP+ORB+Bi-GRU (e) ViT+LBP+ORB+Attention+Bi-GRU. Here, M1: Neutral, M2: Happy, M3: Sad, M4: Angry, M5: Fear, M6: Disgust, and M7: Surprise.

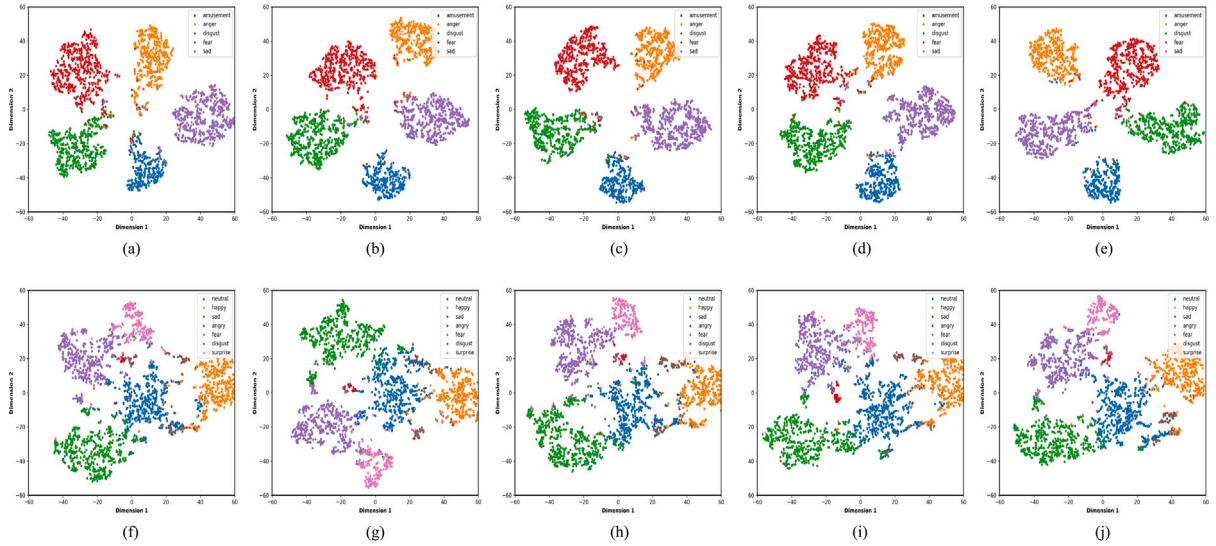

Fig. 5. t-SNE feature visualization of the MSF-Net for multimodal emotion recognition. (a) & (f) ViT+Attention+Bi-GRU (b) & (g) ViT+LBP+Attention+Bi-GRU (c) & (h) ViT+ORB+Attention+Bi-GRU (d) & (i) ViT+LBP+ORB+Bi-GRU (e) & (j) ViT+LBP+ORB+Attention+Bi-GRU. Different categories of emotions are represented with distinct colors and shapes. The first row represents the results from the BioVid Emo DB dataset, and the second row depicts the results from the MGEED database.

emotions on the BioVid Emo DB and MGEED datasets. For instance, the *disgust* emotion from the BioVid Emo DB dataset and the *happy* emotion from the MGEED dataset are correctly classified using the proposed MSF-Net.

On the BioVid Emo DB dataset (Fig. 3), it is observed that the *amusement* emotion are mostly misclassified by the *sad* emotion. In this case, the developed architecture is failed to differentiate between *amusement* and *sad* emotions. The *fear* emotion are wrongly identified as *disgust* and *sad* emotions. The reason behind it there are semantic correlations between *fear* emotion and *disgust* or *sad* emotion. At the same time, it is found that the *disgust* emotion has the tend to misclassify as *fear*. Moreover, there is shared semantic information between *anger* and *sad* emotions, which poses a challenge for the MSF-Net to differentiate between these emotions. On the MGEED dataset (Fig. 4), it is revealed that the *disgust*, *happy*, and *sad* emotions are confused with the *neutral* emotion in some cases. The reason for this is that the MGEED dataset contains an excess number of neutral emotions. Thus, MSF-Net has a tendency to acquire a greater understanding of the neutral emotion. The *sad* emotion is recognized as *fear* emotion and vice versa. This result is found because the *sad* and *fear* are highly close in the activation and arousal domains, meaning that the speakers have similar perception and interpretation of these emotions. We did not observe misclassification rates for several emotion categories, which are: *happy* and *surprise*, *happy* and *angry*, *sad* and *angry*, as well as *angry* and *surprise*.

Fig. 5 depicts the feature distribution of the generated fused representations on the BioVid Emo DB and MGEED datasets. We randomly choose 2500 samples from each emotion category and show their

Table 5
Ablation analysis (%) of MSF-Net with different combinations of multimodal signals on the BioVid Emo DB and MGEED datasets. We provide the overall performance on the testing set. The best performances are represented by bold front, while the second-best results are marked with underlined text.

Datasets	Signals	A	P	R	F1
BioVid Emo DB	Image+tEMG+ECG	96.42	96.41	96.34	96.38
	Image+tEMG+SCL	96.47	96.41	96.39	96.40
	Image+ECG+SCL	96.41	96.37	96.33	96.35
	Image+tEMG	96.44	96.39	96.36	96.38
	Image+ECG	**96.52**	**96.49**	**96.43**	**96.46**
	Image+SCL	96.39	96.36	96.34	96.35
MGEED	RGB+OMG+EEG+ECG	93.59	91.82	90.61	91.11
	RGB+OMG+EEG+GSR	93.85	91.74	91.18	91.40
	RGB+EEG+ECG+GSR	93.48	91.50	90.44	90.86
	RGB+OMG+EEG	93.88	91.87	91.33	91.55
	RGB+OMG+GSR	93.63	91.73	90.83	91.20
	RGB+EEG+ECG	93.53	91.30	90.99	91.08
	RGB+EEG+GSR	93.58	91.29	90.92	91.05
	RGB+ECG+GSR	93.87	92.14	91.50	**91.73**
	RGB+OMG	93.60	91.72	**91.67**	91.63
	RGB+EEG	93.61	91.86	90.60	91.12
	RGB+ECG	**93.90**	**92.27**	91.14	91.62
	RGB+GSR	93.70	92.23	90.93	91.49

combined representations in the shared feature space. It is shown from Fig. 5 that the data distribution for MGEED dataset is slightly scattered than the BioVid Emo DB dataset meaning the proposed MSF-Net learnt the features more efficiently from the BioVid Emo DB dataset. Considering the BioVid Emo DB dataset, the *amusement* and *anger* emotion categories are more critical to differentiate from other emotions. They are found in nearly all of the other emotion clusters. For the MGEED dataset, it is quite hard to differentiate between *neutral* and *disgust* emotions. The cluster of *disgust* emotion are quite small and is mostly represented in other emotion groups. Our observation revealed that *disgust* emotional data is frequently misclassified into different emotion categories, while the misclassification of other emotions as *disgust* is very rare.

4.7. Ablation studies

This section performs ablation studies on different physiological signals to see the effects of different signals on the model efficiency on the BioVid Emo DB and MGEED datasets. The experimental outcomes are demonstrated in Table 5, Table 6, and Table 7 considering the architecture of $ViT + LBP + ORB + Attention + Bi\text{-}GRU$. Table 5 illustrates the overall performance on both datasets. Table 6, and Table 7 depict the class-wise emotion recognition performance on BioVid Emo DB and MGEED datasets by varying the physiological signals. On the BioVid Emo DB database, the highest performance metric is found for the signal combination of $Image + ECG$. The best scores are: accuracy (96.52%), precision (96.49%), recall (96.43%), and F1-Score (96.46%). The second-best performance is obtained with the combination of $Image + tEMG + SCL$. Similar to BioVid Emo DB, $RGB + ECG$ obtained the best accuracy (93.90%) and precision (92.27%) from MGEED dataset. However, it should be noted that the $RGB + ECG + GSR$ combination has the highest recall (91.50%) and F1-Score (91.73%), showing an excellent balance of precision and recall for the MGEED dataset. The other combination, including $RGB+OMG$ performed the best for recall (91.67%). Multimodal signals are found to improve performance for both datasets. The choice of modalities, including *RGB*, *ECG*, and *GSR*, has an influence on various aspects of performance. Moreover, the efficiency of various signal combinations may be dependent upon the nature of the dataset and its particular features, as evidenced by the variation in performance across datasets.

From Table 6, it is noticed that the highest metrics are achieved for *disgust* emotion considering the combination of $Image + ECG$ signals. The recorded scores are accuracy (97.57%), recall (97.57%), precision (96.90%), and F1-Score (97.24%). For the *fear* emotion, the signal combination $Image+tEMG+SCL$ received the best precision (96.53%) and F1-score (96.39%), highlighting that the SCL signal can improve the recognition of this emotion. However, the F1-Scores for *amusement* and *anger* have comparatively lower values for all signal combinations, showing difficulties in effectively recognizing these emotions. The robust performance of combinations that incorporate *ECG* indicates its crucial role in enhancing the efficiency of emotion recognition. In Table 7, we revealed that the $RGB + ECG$ found the highest accuracy (96.13%), precision (96.18%), and F1-Score (96.19%) for *sad* emotion. For the same combination, the *neutral* emotion received the best accuracy of 92.49%. The best precision of 96.52% for *sad* emotion is achieved for the $RGB + OMG + EEG$ combination. While considering the *disgust* and *surprise* emotions, the performance for almost all signal combinations is relatively low, noting the significant variation in precision and F1-Scores. This outcome demonstrated that accurate recognition of these emotions is quite a challenge. In some cases, the *GSR* signal enhanced the performance for emotion recognition. For instance, $RGB+OMG+GSR$ recorded a comparatively higher F1-Score of 80.14% for *surprise* emotion. In summary, the findings emphasized the various effects of different signal combinations. In particular, the combination of *ECG* and *EEG* consistently improved the recognition of certain emotions, whereas *GSR* has limited but focused efficiency.

4.8. IoT performance

We deployed our proposed multi-stage fusion architecture in scalable healthcare environment. We considered two performance measures, including latency and response time, to evaluate the scalability of the IoT network. Latency refers to the time to transmit the packets from the patients and the cloud server receives them. The response time is the total time to send the request from the patients and receives the classification results from the cloud server. This is the time measure for sending data from patients and receiving results, along with inference time. Inference time represents the time to make classifications through deep learning architecture on the data received from the patients.

Fig. 6 shows the performance of the IoT deployment in scalable healthcare environment. We deployed $ViT + LBP + ORB + Attention + Bi\text{-}GRU$ model in cloud server that continuously takes input from multiple patients and makes classification for multimodal emotion recognition. It is found from Fig. 6(a) that the latency for the MGEED data is comparatively higher than the BioVid Emo DB data. The reason behind this is that the MGEED dataset considered more physiological signals (70-channel EEG, 20-channel OMG, as well as one-channel ECG and GSR) compared to BioVid Emo DB dataset. The BioVid Emo DB dataset only considered the single channel tEMG, ECG, and SCL. While considering the response time as shown in Fig. 6(b), the response time

Table 6
Ablation analysis (%) of MSF-Net with different combinations of multimodal signals on the BioVid Emo DB dataset. We provide each emotion categories classification performances. The best performances are represented by bold front, while the second-best results are marked with underlined text.

Signals	Amusement				Anger				Disgust				Fear				Sad			
	A	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*
Image+tEMG+ECG	95.59	96.36	95.59	95.97	95.78	96.19	95.78	95.99	97.44	96.86	97.44	97.15	96.11	96.27	96.11	96.19	96.80	96.39	96.80	96.59
Image+tEMG+SCL	95.63	95.90	95.63	95.77	95.83	96.00	95.83	95.91	97.53	97.11	97.53	**97.32**	96.24	96.53	96.24	96.39	96.70	96.51	96.70	96.61
Image+ECG+SCL	95.75	96.06	95.75	95.90	95.60	95.99	95.60	95.80	97.38	97.12	97.38	97.25	96.12	96.24	96.12	96.18	96.81	96.44	96.81	96.62
Image+tEMG	95.65	96.07	95.65	95.86	95.79	95.90	95.79	95.85	97.23	**97.18**	97.23	97.21	96.48	96.28	96.48	96.38	96.63	96.54	96.63	96.58
Image+ECG	95.65	96.15	95.65	95.90	95.67	96.32	95.67	95.99	**97.57**	96.90	**97.57**	97.24	96.42	96.52	96.42	96.47	96.84	96.54	96.84	96.69
Image+SCL	95.99	95.90	95.99	95.94	95.68	96.27	95.68	95.98	97.08	97.04	97.08	97.06	96.17	95.96	96.17	96.07	96.77	96.61	96.77	96.69

Table 7
Ablation analysis (%) of MSF-Net with different combinations of multimodal signals on the MGEED dataset. We provide each emotion categories classification performances. The best performances are represented by bold front, while the second-best results are marked with underlined text.

Signals	Neutral				Happy				Sad				Angry				Fear				Disgust				Surprise			
	A	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*
RGB+OMG+EEG+ECG	91.37	90.95	91.37	91.16	95.50	94.08	95.50	94.78	95.69	96.10	95.69	95.89	90.70	87.44	90.70	89.04	95.48	94.96	95.48	95.22	71.28	86.03	71.28	77.96	94.27	93.22	94.27	93.74
RGB+OMG+EEG+GSR	91.50	91.91	91.50	91.71	95.85	94.52	95.85	95.18	95.75	96.00	95.75	95.87	90.93	86.50	90.93	88.66	95.77	94.96	95.77	95.36	74.28	84.82	74.28	79.20	94.19	93.48	94.19	93.83
RGB+EEG+ECG+GSR	91.55	90.75	91.55	91.15	95.43	94.57	95.43	94.95	95.75	95.95	95.75	95.85	90.70	86.09	90.70	88.33	95.11	94.81	95.11	94.96	71.17	85.79	71.17	77.80	93.35	92.57	93.35	92.96
RGB+OMG+EEG	92.07	91.18	92.07	91.62	95.33	94.85	95.33	95.09	95.85	**96.52**	95.85	96.18	90.47	86.06	90.47	88.21	95.26	95.03	95.26	95.14	75.99	85.94	75.99	80.66	94.35	93.49	94.35	93.92
RGB+OMG+GSR	91.98	90.70	91.98	91.34	95.39	94.94	95.39	95.16	95.69	96.08	95.69	95.89	89.77	85.78	89.77	87.73	94.91	95.04	94.91	94.98	74.38	86.86	74.38	80.14	93.67	92.74	93.67	93.20
RGB+EEG+ECG	91.44	91.33	91.44	91.39	95.52	94.22	95.52	94.86	95.72	96.06	95.72	95.91	90.64	84.52	90.64	87.82	95.72	95.72	95.72	95.72	74.38	84.52	74.38	79.45	93.79	87.82	93.79	91.49
RGB+EEG+GSR	91.62	91.17	91.62	91.39	95.67	94.83	95.67	95.25	95.44	96.02	95.44	95.73	90.00	84.31	90.00	87.06	94.78	94.40	94.78	94.59	74.49	84.65	74.49	79.25	94.43	93.68	94.43	94.05
RGB+ECG+GSR	91.84	91.34	91.84	91.59	95.14	94.77	95.14	94.95	**96.13**	96.18	**96.13**	96.15	92.56	86.33	92.56	89.34	95.33	95.03	95.33	95.18	75.24	88.19	75.24	81.20	94.27	93.11	94.27	93.69
RGB+OMG	91.77	91.02	91.77	91.39	94.78	94.51	94.78	94.65	95.48	96.05	95.48	95.76	93.72	85.74	93.72	89.56	95.01	94.74	95.01	94.88	77.17	86.64	77.17	81.63	93.79	93.30	93.79	93.54
RGB+EEG	91.64	91.04	91.64	91.34	95.83	94.27	95.83	95.04	95.73	95.79	95.73	95.76	90.70	87.25	90.70	88.94	95.19	95.41	95.19	95.30	71.60	86.87	71.60	78.50	93.51	92.43	93.51	92.97
RGB+ECG	92.49	91.16	92.49	91.82	95.01	94.88	95.01	94.95	96.11	96.26	96.11	**96.19**	91.86	88.17	91.86	89.98	95.31	95.20	95.31	95.26	73.20	87.01	73.20	79.51	94.03	93.21	94.03	93.62
RGB+GSR	91.77	90.60	91.77	91.18	95.01	94.40	95.01	94.70	95.78	95.95	95.78	95.86	91.40	88.51	91.40	89.93	95.50	95.29	95.50	95.40	72.45	86.78	72.45	78.97	94.59	94.10	94.59	94.34

in the case of MGEED dataset is slightly higher than the BioVid Emo DB database. For the fewer number of patients, the response time for both cases are mostly close. Moreover, the inference time for the MGEED data is relatively high. This is because of the large number of training parameters for the fusion architecture considering MGEED data. At the same time, the fusion network developed for MGEED data takes relatively more time to process a large number of input signals compared to BioVid Emo DB dataset. Overall, the deployed network is capable of handling an increasing number of patients. The performance metrics are increasing for the larger number of connected patients; however, all measures are relatively low to be applicable for real-world scalable healthcare scenario. There are few factors that affect the performance of IoT-enabled healthcare network, such as the size of the developed model, the cloud architecture, the hardware capabilities and the nature of the data.

5. Conclusion

The emotion recognition has grown in significance since it ensures specialized supports in a variety of contexts, including personalized healthcare. In this article, we introduced a multi-stage fusion network for emotion recognition from multimodal data in scalable healthcare. In this fusion network, the visual features are extracted using transformer, LBP, and ORB, which are fused together and sent to the triplet attention module to capture cross-dimensional interactions from the deep features and computer vision-based features. The Bi-GRU network modeled the physiological signals to obtain the significant temporal dependencies. Finally, the retrieved facial patterns and temporal features are fused through feature-level fusion and fed into the FC layer for multimodal emotion recognition. The effectiveness of the proposed fusion network has been extensively tested using two benchmark databases: BioVid Emo DB and MGEED. The outcomes clearly demonstrate that the efficiency of multimodal discrete emotion recognition is significantly improved by the multi-stage fusion of the extracted features through feature-level fusion. Finally, we deployed the proposed multi-stage fusion architecture in IoT network to monitor multiple patients in scalable healthcare.

The proposed improvement in future work is to enhance the robustness of multimodal fusion architectures by dynamically sharing weights to each modalities according to their relevance and usefulness,

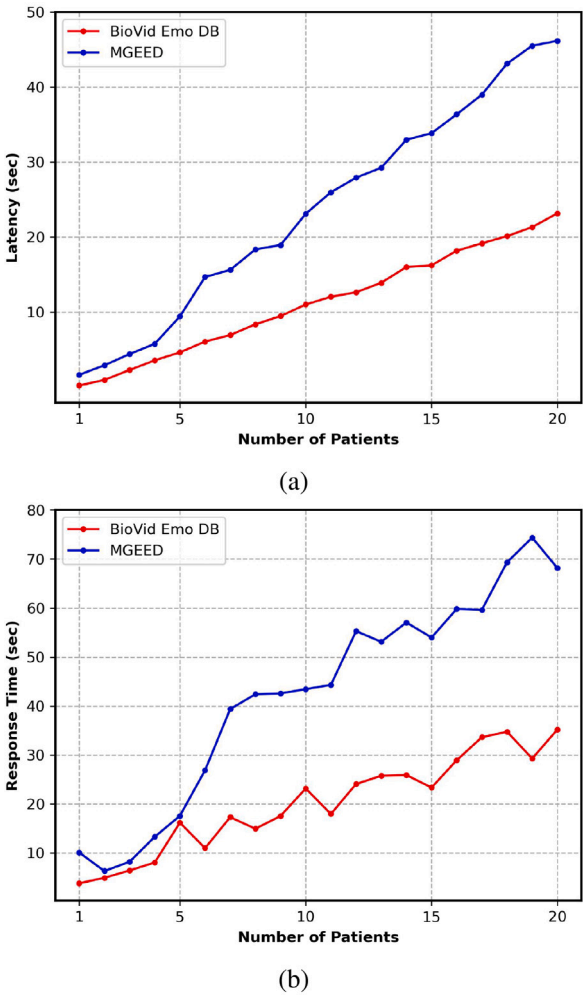

Fig. 6. Performance measures for IoT deployment in scalable healthcare scenario. (a) Latency (b) Response time.

hence boosting performance by emphasizing relevant information and reducing noise. The next future work will be to develop fully functional scalable healthcare to monitor patients by consulting with medical experts.

CRediT authorship contribution statement

Md. Milon Islam: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Fakhri Karray:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Ghulam Muhammad:** Writing – review & editing, Funding acquisition, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, Canada. Moreover, G. Muhammad acknowledges the support of Researchers Supporting Project number (RSP2025R34), King Saud University, Riyadh, Saudi Arabia. All the authors have read and agreed to this version of the manuscript.

Data availability

Data will be made available on request.

References

[1] A. Geetha, T. Mala, D. Priyanka, E. Uma, Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions, Inf. Fusion 105 (2024) 102218.
[2] S. Saganowski, B. Perz, A.G. Polak, P. Kazienko, Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review, IEEE Trans. Affect. Comput. 14 (3) (2023) 1876–1897.
[3] G. Muhammad, F. Alshehri, F. Karray, A. El Saddik, M. Alsulaiman, T.H. Falk, A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Inf. Fusion 76 (2021) 355–375.
[4] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods, Neurocomputing 561 (2023) 126866.
[5] P.S. Tomar, K. Mathur, U. Suman, Unimodal approaches for emotion recognition: A systematic review, Cogn. Syst. Res. 77 (2023) 94–109.
[6] A.A. Gladys, V. Vetriselvi, Survey on multimodal approaches to emotion recognition, Neurocomputing 556 (2023) 126693.
[7] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, Inf. Fusion 59 (2020) 103–126.
[8] J.A. Domínguez-Jiménez, K.C. Campo-Landines, J.C. Martínez-Santos, E.J. Delahoz, S.H. Contreras-Ortiz, A machine learning model for emotion recognition from physiological signals, Biomed. Signal Process. Control. 55 (2020) 101646.
[9] R.H. Aljuhani, A. Alshutayri, S. Alahdal, Arabic speech emotion recognition from Saudi dialect corpus, IEEE Access 9 (2021) 127081–127085.
[10] N. Azam, T. Ahmad, N.U. Haq, Automatic emotion recognition in healthcare data using supervised machine learning, PeerJ Comput. Sci. 7 (2021) e751.
[11] Y. Cai, X. Li, J. Li, Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review, Sensors 23 (5) (2023) 2455.
[12] S. Dwijayanti, M. Iqbal, B.Y. Suprapto, Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network, IEEE Access 10 (2022) 89876–89886.
[13] M. Hussain, E.-U.-H. Qazi, H.A. AboAlSamh, I. Ullah, Emotion recognition system based on two-level ensemble of deep-convolutional neural network models, IEEE Access 11 (2023) 16875–16895.
[14] I. Kansizoglou, E. Misirlis, K. Tsintotas, A. Gasteratos, Continuous emotion recognition for long-term behavior modeling through recurrent neural networks, Technologies 10 (3) (2022) 59.
[15] C. Fan, H. Xie, J. Tao, Y. Li, G. Pei, T. Li, Z. Lv, ICaps-ResLSTM: Improved capsule network and residual LSTM for EEG emotion recognition, Biomed. Signal Process. Control. 87 (2024) 105422.
[16] B. Chakravarthi, S.-C. Ng, M. Ezilarasan, M.-F. Leung, EEG-based emotion recognition using hybrid CNN and LSTM classification, Front. Comput. Neurosci. 16 (2022) 1019776.
[17] R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, S. Singh, Facial expression recognition in videos using hybrid CNN & ConvLSTM, Int. J. Inf. Technol. 15 (4) (2023) 1819–1830.
[18] S.C. Prince, N.V. Babu, Advancing multiclass emotion recognition with CNN-RNN architecture and illuminating module for real-time precision using facial expressions, in: 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science, AMATHE, IEEE, 2024, pp. 1–11.
[19] Y. Zhao, M. Guo, X. Sun, X. Chen, F. Zhao, Attention-based sensor fusion for emotion recognition from human motion by combining convolutional neural network and weighted kernel support vector machine and using inertial measurement unit signals, IET Signal Process. 17 (4) (2023) e12201.
[20] S. Akinpelu, S. Viriri, Speech emotion classification using attention based network and regularized feature selection, Sci. Rep. 13 (1) (2023) 11990.
[21] K. Ezzameli, H. Mahersia, Emotion recognition from unimodal to multimodal analysis: A review, Inf. Fusion 99 (2023) 101847.
[22] Y.S. Can, B. Mahesh, E. André, Approaches, applications, and challenges in physiological emotion recognition—a tutorial overview, Proc. IEEE 111 (10) (2023) 1287–1313.
[23] M. Algarni, F. Saeed, T. Al-Hadhrami, F. Ghabban, M. Al-Sarem, Deep learning-based approach for emotion recognition using electroencephalography (EEG) signals using bi-directional long short-term memory (Bi-LSTM), Sensors 22 (8) (2022) 2976.
[24] A. Mahmoud, K. Amin, M.M. Al Rahhal, W.S. Elkilani, M.L. Mekhalfi, M. Ibrahim, A CNN approach for emotion recognition via EEG, Symmetry 15 (10) (2023) 1822.
[25] M.A. Rahman, A. Anjum, M.M.H. Milu, F. Khanam, M.S. Uddin, M.N. Mollah, Emotion recognition from EEG-based relative power spectral topography using convolutional neural network, Array 11 (2021) 100072.
[26] M.A. Ozdemir, M. Degirmenci, E. Izci, A. Akan, EEG-based emotion recognition with deep convolutional neural networks, Biomed. Eng./Biomed. Tech. 66 (1) (2021) 43–57.
[27] M.R. Islam, M.M. Islam, M.M. Rahman, C. Mondal, S.K. Singha, M. Ahmad, A. Awal, M.S. Islam, M.A. Moni, EEG channel correlation based model for emotion recognition, Comput. Biol. Med. 136 (2021) 104757.
[28] F.Z. Canal, T.R. Müller, J.C. Matias, G.G. Scotton, A.R. de Sa Jr., E. Pozzebon, A.C. Sobieranski, A survey on facial emotion recognition techniques: A state-of-the-art literature review, Inform. Sci. 582 (2022) 593–617.
[29] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J.S. Kumar, R. Manjunath, S. Jaiswal, Facial emotion recognition using convolutional neural networks, Mater. Today: Proc. 80 (2023) 3560–3564.
[30] S. Minaee, M. Minaei, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, Sensors 21 (9) (2021) 3046.
[31] M.K. Chowdary, T.N. Nguyen, D.J. Hemanth, Deep learning-based facial emotion recognition for human–computer interaction applications, Neural Comput. Appl. 35 (32) (2023) 23311–23328.
[32] Z.-Y. Huang, C.-C. Chiang, J.-H. Chen, Y.-C. Chen, H.-L. Chung, Y.-P. Cai, H.-C. Hsu, A study on computer vision for facial emotion recognition, Sci. Rep. 13 (1) (2023) 8425.
[33] G. Meena, K.K. Mohbey, A. Indian, M.Z. Khan, S. Kumar, Identifying emotions from facial expressions using a deep convolutional neural network-based approach, Multimedia Tools Appl. 83 (2024) 15711–15732.
[34] M.M. Islam, S. Nooruddin, F. Karray, G. Muhammad, Enhanced multimodal emotion recognition in healthcare analytics: A deep learning based model-level fusion approach, Biomed. Signal Process. Control. 94 (2024) 106241.
[35] F. Muhammad, M. Hussain, H. Aboalsamh, A bimodal emotion recognition approach through the fusion of electroencephalography and facial sequences, Diagnostics 13 (5) (2023) 977.
[36] A.R. Aguiñaga, D.E. Hernandez, A. Quezada, A. Calvillo Téllez, Emotion recognition by correlating facial expressions and EEG analysis, Appl. Sci. 11 (15) (2021) 6987.
[37] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, C.F. Caiafa, A multimodal emotion recognition method based on facial expressions and electroencephalography, Biomed. Signal Process. Control. 70 (2021) 103029.
[38] B. Mocanu, R. Tapu, T. Zaharia, Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning, Image Vis. Comput. 133 (2023) 104676.
[39] D. Mamieva, A.B. Abdusalomov, A. Kutlimuratov, B. Muminov, T.K. Whangbo, Multimodal emotion detection via attention-based fusion of extracted facial and speech features, Sensors 23 (12) (2023) 5475.
[40] A.I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities, Knowl.-Based Syst. 244 (2022) 108580.

[41] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 7037–7041.
[42] P. Tzirakis, J. Chen, S. Zafeiriou, B. Schuller, End-to-end multimodal affect recognition in real-world environments, Inf. Fusion 68 (2021) 46–53.
[43] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, H.-J. Yang, Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning, IEEE Access 11 (2023) 14742–14751.
[44] C. Fan, J. Lin, R. Mao, E. Cambria, Fusing pairwise modalities for emotion recognition in conversations, Inf. Fusion 106 (2024) 102306.
[45] X. Zhang, J. Sun, S. Hong, T. Li, Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition, in: Findings of the Association for Computational Linguistics, ACL, 2024, pp. 14448–14458.
[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv:2010.11929.
[47] M. Pietikäinen, G. Zhao, Two decades of local binary patterns: A survey, in: Advances in Independent Component Analysis and Learning Machines, Elsevier, 2015, pp. 175–210.
[48] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2564–2571.
[49] D. Misra, T. Nalamada, A.U. Arasanipalai, Q. Hou, Rotate to attend: Convolutional triplet attention module, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3139–3148.
[50] B.-H. Su, C.-M. Chang, Y.-S. Lin, C.-C. Lee, Improving speech emotion recognition using graph attentive bi-directional gated recurrent unit network, in: INTERSPEECH, 2020, pp. 506–510.
[51] L. Zhang, S. Walter, X. Ma, P. Werner, A. Al-Hamadi, H.C. Traue, S. Gruss, "BioVid Emo DB": A multimodal database for emotion analyses validated by subjective ratings, in: 2016 IEEE Symposium Series on Computational Intelligence, SSCI, IEEE, 2016, pp. 1–6.
[52] Y. Wang, H. Yu, W. Gao, Y. Xia, C. Nduka, MGEED: A multimodal genuine emotion and expression detection database, IEEE Trans. Affect. Comput. 15 (2) (2024) 606–619.
[53] S. Bano, N. Tonellotto, P. Cassarà, A. Gotta, FedCMD: A federated cross-modal knowledge distillation for drivers' emotion recognition, ACM Trans. Intell. Syst. Technol. 15 (3) (2024) 1–27.
[54] A. Kumar, K. Sharma, A. Sharma, MEmoR: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries, Image Vis. Comput. 123 (2022) 104483.
[55] Z. Wang, X. Zhou, W. Wang, C. Liang, Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video, Int. J. Mach. Learn. Cybern. 11 (4) (2020) 923–934.
[56] J. Xie, X. Xu, L. Shu, WT feature based emotion recognition from multi-channel physiological signals with decision fusion, in: 2018 First Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia, IEEE, 2018, pp. 1–6.
[57] Z. Cheng, L. Shu, J. Xie, C.P. Chen, A novel ECG-based real-time detection method of negative emotions in wearable applications, in: 2017 International Conference on Security, Pattern Analysis, and Cybernetics, SPAC, IEEE, 2017, pp. 296–301.