IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 February 2025, accepted 19 March 2025, date of publication 28 March 2025, date of current version 23 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3555934

RESEARCH ARTICLE

Deep Fusion of Neurophysiological and Facial Features for Enhanced Emotion Detection

FARSHAD SAFAVI, (Member, IEEE), VIKAS REDDY VENKANNAGARI, (Member, IEEE), DEV PARIKH, AND RAMANA KUMAR VINJAMURI, (Senior Member, IEEE)

Vinjamuri Laboratory, Department of Computer Science and Electrical Engineering, University of Maryland at Baltimore County, Baltimore, MD 21250, USA

Corresponding author: Ramana Kumar Vinjamuri (rvinjam1@umbc.edu)

This work was supported by the National Science Foundation Faculty Early Career (CAREER) Development Award HCC-2053498.

ABSTRACT The fusion of facial and neurophysiological features for multimodal emotion detection is vital for applications in healthcare, wearable devices, and human-computer interaction, as it enables a more comprehensive understanding of human emotions. Traditionally, the integration of facial expressions and neurophysiological signals has required specialized knowledge and complex preprocessing. With the rise of deep learning and artificial intelligence (AI), new methodologies in affective computing allow for the seamless fusion of multimodal signals, advancing emotion recognition systems. In this paper, we present a novel multimodal deep network that leverages transformers to extract comprehensive features from neurophysiological data, which are then fused with facial expression features for emotion classification. Our transformer-based model analyzes neurophysiological time-series data, while transformer-inspired methods extract facial expression features, enabling the classification of complex emotional states. We compare single modality with multimodal systems, testing our model on Electroencephalography (EEG) signals using the DEAP and Lie Detection datasets. Our hybrid approach effectively captures intricate temporal and spatial patterns in the data, significantly enhancing the system's emotion recognition accuracy. Validated on the DEAP dataset, our method achieves near state-of-the-art performance, with accuracy rates of 97.78%, 97.64%, 97.91%, and 97.62% for arousal, valence, liking, and dominance, respectively. Furthermore, we achieved a precision of 97.9%, a ROC AUC score of 97.6%, an F1-score of 98.1%, and a recall of 98.2%, demonstrating the model's robust performance. We demonstrated the effectiveness of this method, specifically for EEG caps with a limited number of electrodes, in emotion detection for wearable devices.

INDEX TERMS Affective computing, emotion detection, deep learning, multimodal emotion recognition, transformer.

I. INTRODUCTION

Emotion recognition is becoming an essential aspect of digital health, allowing machines to understand and react to human emotional states. This technology is integral to AI-based clinical decision support systems and digital health applications. Enhancing user experiences in healthcare, supports task monitoring, patient well-being, mental health assessment, and even personalized medicine by providing healthcare solutions based on individual emotional states. To recognize emotion, there are two main emotion recognition representations: categorical and dimensional. The categorical approach classifies emotions into eight distinct categories including happiness, sadness, surprise, fear, anger, disgust, contempt, and neutral. While the dimensional theory explains emotion recognition using two axes: valence (ranging from pleasant to unpleasant) and arousal (ranging from calm to energized) [1]. Emotions are combinations of these two dimensions in 2D space. Our study utilizes dimensional theory to measure emotions.

A. MOTIVATION

Despite advancements in digital health, many digital health applications lack emotional intelligence, hindering their ability to enable authentic and effective human interaction. Personalized systems capable of understanding

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

and responding to human emotions are essential for improving user experiences in social settings. For example, during contagious situations such as pandemics, robots can be deployed to minimize direct human contact while ensuring care quality. Emotionally intelligent robots, capable of understanding and responding to emotional cues, can assist patients and healthcare workers more effectively, providing personalized and empathetic support [2]. In mental health monitoring, our model could assist therapists by detecting shifts in a patient's emotional state through EEG and facial cues. Similarly, in immersive gaming and adaptive learning, systems could dynamically adjust content based on user engagement and emotions, improving user experience.

While advances in deep learning and representation learning have greatly improved emotion recognition through individual modalities—such as computer vision for facial expression recognition—there remains a lack of attention to the integration of neurophysiological signals and emotional perception. To address this limitation, we are developing a multimodal emotion recognition system that combines neurophysiological features (e.g., EEG signals) with facial expressions. Our approach employs a deep neural network architecture to extract features from each modality (bio-sensing and vision) and fuse them for more accurate results. The system evaluates users' emotional responses and identifies their expressions through a dimensional representation. We validate our proposed multimodal system using the DEAP dataset [3], demonstrating its effectiveness in predicting emotional states accurately.

B. CONTRINUTIONS

In this work, we present a comprehensive approach to multimodal emotion recognition by leveraging deep learning techniques and integrating facial and neurophysiological signals to address key challenges in emotion recognition systems.

First, we propose a novel end-to-end deep network for multi-modal emotion recognition that integrates both facial and neurophysiological signals. This deep network is rigorously validated on the DEAP dataset [3] using standard evaluation metrics such as accuracy, F1-score, sensitivity, and ROC-AUC.

Second, we introduce a Transformer-based architecture to capture intricate patterns in neurophysiological data. Our approach effectively learns a comprehensive feature representation by handling the simultaneous nature of multi-channel EEG signals. This model captures nuanced patterns across both temporal and spatial dimensions, significantly improving the granularity and accuracy of emotion recognition. To further enhance performance, we employ a transformer-inspired technique for facial expression recognition. The proposed architecture demonstrates superior performance on the DEAP dataset [3] and is also validated on the LieWaves dataset [4], highlighting its robustness and versatility compared to traditional methods.

Third, we demonstrate the importance of combining bio-signals with facial features for emotion recognition, particularly in the context of wearable devices. Our experiments reveal that while transformers applied to single or dual EEG channels may exhibit lower accuracy due to reduced signal information, integrating these channels with facial features substantially improves the overall performance. This fusion is especially critical for wearable devices that utilize only one or two EEG electrodes, where accuracy typically declines. The proposed approach enables real-time emotion recognition, making it highly suitable for applications in human-computer interaction and medical scenarios involving social robots. In summary, our key contributions of this paper include:

- Development of a multi-modal deep network integrating facial and neurophysiological signals, tested on the DEAP dataset [3].
- Introduction of a Transformer-based model for EEG data to capture nuanced temporal patterns, validated on the DEAP [3] and LieWaves datasets [4].
- Proposal of an effective fusion strategy for combining EEG and facial features to enhance model accuracy and efficiency, demonstrating its applicability in wearable EEG systems for emotion recognition.

C. ORGANIZATION

The remainder of this paper is organized as follows: Section II reviews related works on emotion recognition, covering advancements in facial expression analysis, EEG-based approaches, and multimodal techniques that integrate neurophysiological and visual data. Section III provides the mathematical formulation of the proposed model for fusion of facial and bio-sensing signals. Section IV introduces the proposed multimodal architecture, detailing the integration of facial and neurophysiological signals for emotion recognition. This is followed by subsections describing the neurophysiological feature extractor and facial feature extractor, outlining the methodologies for extracting features from EEG signals and facial images and fusion block, which combines features from both modalities to enhance prediction accuracy. Section V elaborates on the experimental setup, describing datasets, preprocessing steps, and model training procedures. Section VI highlights the results, providing performance comparisons with existing methods and evaluating the effectiveness of the proposed system. In Section VII, the discussion interprets the results, addressing the significance and applications of multimodal emotion recognition. Finally, Section VIII concludes the paper by summarizing key contributions and outlining future directions for research in affective computing and emotion recognition.

II. RELATED WORKS

A. FACIAL EXPRESSION RECOGNITION

The recognition of facial expressions has gained significant attention with the rise of deep learning, as it enables the identification of human emotions through visual cues captured by cameras. Convolutional Neural Networks (CNNs) and their variants are particularly viable solutions at processing

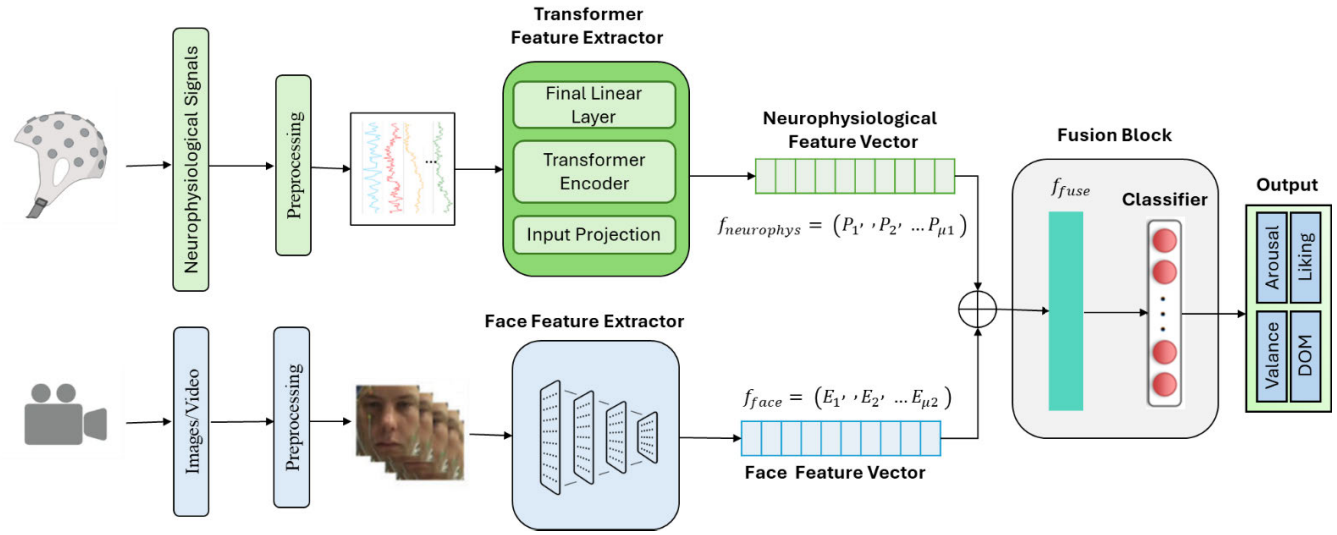

FIGURE 1. The multimodal architecture integrates neurophysiological and facial features using transformers and a facial expression network, fusing $f_{neurophys}$ and f_{face} into f_{fuse}, to predict emotional dimensions.

visual data for this purpose. For analyzing facial image sequences or videos, advanced methods such as Recurrent Convolutional Neural Networks (RCNN) [5] and the combination of CNN with Long Short-Term Memory (LSTM) networks excel at extracting both spatial and temporal features of facial expressions [6]. Additionally, the increasing popularity of transformers has led to the development of innovative facial expression recognition techniques utilizing Mix Transformers [7].

B. NEUROPHYSIOLOGICAL SIGNALS

Emotion recognition leverages neurophysiological and bio-sensing signals such as electroencephalogram (EEG), electrocardiogram (ECG), and galvanic skin response (GSR) data. Advanced models have begun exploring transformers and graph-based methods for this purpose [8], [9]. Hybrid models that combine CNNs with sparse autoencoders and deep neural networks have achieved high accuracy. Additionally, incorporating attention mechanisms and regional feature extraction through graph convolutional networks has yielded promising results [10]. These studies highlight the potential of advanced neural network architectures to significantly enhance emotion recognition from EEG data.

C. FUSION OF FACIAL AND NEUROPHYSIOLOGICAL SIGNALS

While single-modality approaches offer certain advantages, integrating multiple modalities can lead to more comprehensive and accurate emotion recognition. Relying solely on facial expressions can be problematic due to the potential for deceptive cues and using wearable EEG devices with only one or two electrodes' limits accuracy. However, combining facial recognition with neurophysiological signals, which are involuntary, significantly enhances both the reliability and accuracy of emotion detection. Multi-modal affective computing has gained significant interest due to its ability to enhance emotion recognition accuracy by leveraging diverse types of data. Most contemporary methods focus on using audio and video inputs, as noted in studies [11], [12], [13]. In these methods, audio data is often converted into Mel-spectrograms, which are treated as images, and selected video frames are processed through CNNs. The features extracted from these inputs are then fused to improve emotion prediction. These architectures consistently suggest that multimodality is the most effective approach for predicting emotions.

D. RECENT MULTIMODAL EMOTION RECOGNITION

Recent advancements in multimodal emotion recognition have significantly enhanced the accuracy and comprehensiveness of affective computing systems. A novel transformer-based architecture improves upon traditional late-fusion methods by incorporating "fusion bottlenecks" at multiple layers, enabling early and efficient information exchange between modalities like vision and audio [14]. Similarly, the TransFuser framework utilizes a transformer-based network to fuse visual and LiDAR perception for autonomous driving, demonstrating the applicability of transformers in extracting synergistic functionalities from diverse inputs [14].

Zhang et al. proposed the MART framework, which addresses the challenge of obtaining sufficient training data in video emotion analysis (VEA) by introducing a masked affective representation learning approach [15]. Siriwardhana et al. introduced a transformer-based fusion mechanism that incorporates self-attention to combine high-dimensional features from text, audio, and videos [16].

Li et al. proposed the Dual-level Disentanglement Mechanism (DDM) to disentangle modality and utterance features,

along with the Contribution-aware Fusion Mechanism (CFM) and Context Refusion Mechanism (CRM) to handle the fusion of multimodal and contextual information [17]. Another study by Li et al. introduced the HiLo model, which incorporates holistic interaction and labeling protocols with concerns to modality and utterance level using different attentions [18].

The PanoSent benchmark introduces new subtasks for multimodal conversational ABSA, including Panoptic Sentiment Sextuple Extraction and Sentiment Flipping Analysis [19]. The Latent Emotion Memory network (LEM) addresses multi-label emotion classification by learning latent emotion distribution without external knowledge [20].

A non-autoregressive encoder-decoder framework has been proposed for end-to-end Aspect-based Sentiment Triplet Extraction (ASTE), modeling it as an unordered triplet prediction problem [21]. Fei et al. proposed the Three-hop Reasoning (THOR) framework based on chain-of-thought (CoT) reasoning for implicit sentiment analysis [22].

In a separate study, Fei et al. suggested a multi-pronged strategy for improving the resilience of Aspect-Based Sentiment Analysis (ABSA) from model, data, and training perspectives [23]. Lastly, the Finsta tool enhances video-language representations through fine-grained structural spatiotemporal alignment, improving performance on various video-language tasks [24].

These studies collectively highlight the importance of multimodal approaches, transformer architectures, and advanced fusion techniques in emotion recognition and sentiment analysis. They provide valuable insights for integrating various modalities for enhanced emotion detection and sentiment understanding, particularly in the context of human-computer interaction and affective computing [25].

E. RESEARCH GAP AND SOLUTIONS

A potential drawback is that cues from audio and facial expressions can be deliberately manipulated, making them less reliable. In contrast, the use of neurophysiological signals, such as EEG, is non-invasive and cannot be easily manipulated. This makes them a more accurate and dependable source of data for emotion recognition. By integrating neurophysiological signals with traditional audio-visual inputs, multi-modal approaches can achieve higher accuracy and robustness in emotion prediction. While the combination of bio-sensing and visual data has been less explored, our study aims to fill this gap by employing a deep multimodal fusion approach that incorporates both neurophysiological signals and visual data for emotion recognition. We utilize transformers, which are highly effective in processing time-series data, to analyze neurophysiological signals and capture detailed features. Transformers are well-suited for this task because of their ability to handle sequential data and identify complex patterns within it. Additionally, we introduce a new facial feature extraction method that uses a transformer-inspired technique for patch extraction. This method enhances our ability to detect subtle facial expressions and integrate them with physiological data for a more accurate and comprehensive emotion recognition system.

III. MODELING FACIAL AND BIO-SENSING SIGNLAS

Consider a scenario where we have a set of facial emotion video frames for the i_{th} instance, denoted $I_i = \{I_t \mid 1, \ldots, n_i\}$, where I_t representing the t_{th} image. Additionally, we have neurophysiological signals $N_i = 1, \ldots, m_i$, where e_t represents the t_{th} data point. The lengths of these video frames and neurophysiological signals are n_i and m_i, respectively. For the i_{th} instance, the ground-truth labelfor i_{th} indicates either a valence, arousal, liking or dominance.

Our objective is to train a comprehensive model G_θ using tuples $\{(I_i, N_i), y_i : i \in [0, T]\}$, where $y_i \in [0, 1]$, and T is the total number of instances in the dataset. Here, y_i indicate the levels of valence, arousal, liking or dominance, with 0 being low and 1 being high. During the prediction phase, when given a test video and corresponding neurophysiological signal pair (I_j, N_j), the trained model G_θ generates an estimated output $\hat{y}_j$, This estimate $\hat{y}_j$ aims to closely match the actual ground-truth annotation y_j. Formally, the prediction process is represented as:

$$\hat{y}_j = G_\theta \left((I_j, N_j) ; \Theta \right) \tag{1}$$

This process allows the model to effectively predict emotional states by leveraging both visual and neurophysiological data.

IV. MULTIMODAL ARCHITECTURE

Our multimodal architecture is inspired by the DeepVaNet model [6] and our previous work on a deep multimodal emotion recognition model [26], utilizing both neurophysiological and facial features to predict emotional states. As depicted in Figure 1, the architecture of our model includes a transformer-based neurophysiological feature extractor, a facial feature extractor, and a fusion block. In this study, we utilize transformers for neurophysiological signals, which have demonstrated their significance in predicting time-series data such as text translation and speech recognition. Additionally, we employ a unique neural network that combines convolutional layers for initial patch extraction and embedding, followed by an LSTM for sequence processing, to extract facial expressions. Preprocessing steps involved downsampling to 128 Hz, removing electrooculography (EOG) artifacts, segmentation, and baseline removal by subtracting the initial 3 seconds of resting-state data for each subject. The EEG data was then divided into one-second intervals. For face preprocessing, 5 frames per second were extracted, and image cropping to a size of 64 × 64 pixels was performed based on facial landmarks.

A. NEUROPHYSIOLOGICAL FEATURE EXTRACTOR

The transformer model is designed for neurophysiological feature extraction, leveraging the capabilities of transformer architectures to capture long-range dependencies

in neurophysiological signals. The data flow in the transformer-based architecture for processing neurophysiological signals begins with the input data, which has a shape of (40, 128), representing 40 neurophysiological signals including the integration of thirty-two EEG (electroencephalogram) and eight physiological signals (EOG: electro-oculogram, GSR: galvanic skin response, BVP: blood volume pulse, RSP: respiration, EMG: electromyogram, SKT: skin temperature, and pulse wave). With a sampling rate of 128 Hz, this data can be analogous to NLP, with 128 data points assumed to be words and 40 channels assumed to be word embeddings, like sentences. We apply this analogy to bio-sensing data to capture the advantages of transformers in understanding complex patterns and contextual relationships. In our implementation, the neurophysiological feature extractor processes the data through an encoding process where the tensor is permuted along the sequence length dimension to be (128, 40) as per our analogy. The transformer can effectively learn important spatial patterns within the EEG data, dynamically adjusting its focus based on the relevance of different points and channels. As depicted in Figure 1, this input is passed through the linear projection layer. In our model, positional encoding was omitted because it did not empirically enhance performance during our experiments. Prior studies on similar recognition tasks, as well as our ablation studies [7], [27], consistently demonstrated that the inclusion or exclusion of positional encoding had negligible impact on the model's accuracy. We selected 2 encoder layers for this task. After extensive research on real-time segmentation models [27], [28] and deep multimodal fusion [26], we applied the Transformer model based on [29], which utilizes the self-attention mechanism. Self-attention mechanism enables the model to simultaneously attend to different parts of an input sequence or multiple sequences, dynamically weighting their importance based on context and capturing cross-channel dependencies.

The structure of the neurophysiological extractor incorporates residual connections, feed-forward layers, and layer normalization. It employs six heads in its multiheaded attention mechanism, with a learning rate of 0.001, a dropout rate of 0.1, and ReLU activation. Finally, we flatten the output of the transformer block, and a linear layer is mapped to get the desired number of feature vectors, resulting in an output of neurophysiological feature scores. This final output represents the encoded features of the input neurophysiological data. This feature vector, represented as a $1\times\mu 1$ vector, is denoted as:

$$f_{neurophys} = \left(P_1, P_2, \ldots, P_{\mu 1}\right) \tag{2}$$

B. FACIAL FEATURE EXTRACTOR

The facial expression feature extractor network is designed to process a sequence of facial images and extract meaningful features for fusion. The use of Convolutional LSTM [30] helps us grasp the spatial and temporal features from CNNs and LSTMs, respectively. The actual facial recordings are of a length of 1 minute. In this case, we try to extract 5 frames every second. We preprocess each frame based on facial landmarks to crop the image to the size of (64, 64). The process begins with the input images of dimensions (batch_size, 5, 3, 64, 64), where 5 denotes the frames per second. These frames are preprocessed and individually passed through a pretrained CNN on the AEFW dataset [31].

The CNN model consists of 4 convolutional layers. Each of the first 3 layers has a kernel size of 3×3 and is followed by ReLU activation and max pooling with a stride of 2×2. The last convolutional layer has padding of 1, a kernel size of 3×3, and a stride of 3×3, resulting in an output shape of (batch_size, 5, 768, 2, 2). To extract the temporal features, an LSTM is deployed, where all the extracted features are flattened to a dimension of (batch_size, 5, 3072). The hidden size of the LSTM is 128. The architecture of network [30] is comprised of following key components:
Input Gate:

$$i_t = \sigma(W_{\{ii\}}x_t + W_{\{hi\}}h_{\{t-1\}} + b_{\{ii\}}) \tag{3}$$

Forget Gate:

$$f_t = \sigma\left(W_{\{if\}}x_t + W_{\{hf\}}h_{\{t-1\}} + b_{\{if\}}\right) \tag{4}$$

Cell Gate:

$$g_t = \tan \mathrm{h}\left(W_{\{ig\}}x_t + b_{\{ig\}} + W_{\{hg\}}h_{\{t-1\}} + b_{\{hg\}}\right) \tag{5}$$

Output Gate:

$$o_t = \sigma\left(W_{\{io\}}x_t + W_{\{ho\}}h_{\{t-1\}} + b_{\{io\}}\right) \tag{6}$$

Cell State:

$$c_t = f_t \odot c_{\{t-1\}} + i_t \odot g_t \tag{7}$$

Hidden State

$$h_t = o_t \odot tanh\,(c_t) \tag{8}$$

In this context, h_t is the hidden state, $c_t$is the cell state, x_t is the input, $h_{\{t-1\}}$ is the previous hidden state, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. These gates and their interactions enable the LSTM to effectively manage long-term dependencies in sequence data. After processing, the output features are passed through a fully connected layer to reduce the dimensionality to the desired feature size. The final output is a feature vector capturing the essential facial features from the input image sequence. This feature vector, represented as a $1\times\mu 2$ vector, is denoted as:

$$f_{face} = \left(E_1, E_2, \ldots, E_{\mu 2}\right) \tag{9}$$

C. FUSION BLOCK

The fusion block is designed to fuse facial and neurophysiological data for tasks such as emotion prediction. It incorporates two main feature extractors: a facial feature extractor and a neurophysiological extractor. The facial

feature extractor reduces spatial dimensions using convolutional embedding and captures essential facial features. The neurophysiological extractor is implemented using a transformer model, which processes bio-sensing data by capturing long-range dependencies within the sequence and encoding it into a feature vector. For feature-level fusion, the face appearance feature f_{face} and the neurophysiological feature $f_{neurophys}$ are concatenated to generate a multi-modal feature vector $f_{fuse} = f_{face} \oplus f_{neurophys}$. This combined feature vector is passed through a multi-layer classifier that sequentially reduces the dimensionality and applies RELU activation functions, culminating in a sigmoid function to produce the final output. This fusion approach leverages both spatial and temporal aspects of the data, making it highly effective for complex and dynamic tasks, particularly in emotion prediction from multi-modal inputs.

V. EXPERIMENT SETUP

A. DEAP DATASET

The Database for emotion analysis using physiological signals (DEAP dataset) is a frequently employed multimodal dataset intended for the analysis of human emotional states. This dataset comprises physiological data collected from 32 participants (16 males and 16 females) while they watched 40 music videos, each selected for its potential to elicit a range of emotional responses. EEG signals were recorded using a 32-channel electrode cap conforming to the ''10-20'' international standard at a sampling frequency of 512 Hz. For emotion classification, the DEAP dataset provides labels for valence, arousal, likin, and dominance enabling categorization of emotional states into distinct classes. In addition to EEG data, video recordings of facial expressions were made for 22 of the 32 participants, providing a rich dataset for multimodal emotion analysis. The DEAP dataset offers a comprehensive resource for investigating the neural and physiological correlates of emotion, facilitating the development and evaluation of emotion recognition models across various modalities.

B. LieWAVES DATASET

The performance of the Transformer classifier was tested on an EEG dataset with a limited number of channels using the LieWaves dataset [4]. This dataset includes both truth and lie trials, with detailed timestamps indicating when each subject answered each question. During data collection, each subject was asked a set of 10 questions, with each question answered twice—once truthfully and once deceitfully. The raw data consists of 160 samples, recorded at a sample rate of 1000 Hz, with 2 EEG channels.

C. TRAINING AND VALIDATION

We employ binary cross-entropy as our loss function. For training, we use a batch size of 64, considering the target emotion label and the predicted score. During inference, we input the test video and physiological signal into our proposed network, which generates a fusion score. This score determines the final prediction: a score above 0.5 results in a ''High'' prediction, while a score of 0.5 or below leads to a ''Low'' prediction. We train and test our model on each individual subject, a process referred to as a per-subject and inter-subject experiments. For the inter-subject experiment, data from all subjects is used to train and test a multimodal network, with the goal of evaluating the generalization ability of our proposed network. This approach is applied to both per-subject and inter-subject experiments. Our model undergoes 10-fold cross-validation, and the average testing accuracy is used to measure performance. For validation, we utilize the mean recognition accuracy of both valence and arousal. The facial features extracted are 16, whereas the neurophysiological features extracted are 64.

D. EVALUATION METRICS

Accuracy: To thoroughly evaluate the model's ability to predict emotional states, we utilized subject-specific evaluation metrics. This metric assesses the model's performance for each individual subject. It is determined by calculating the average accuracy across all folds within a 10-fold cross-validation for each subject.

$$Accuracy_{subject} = \left(\frac{1}{\mathrm{n}}\right) \Sigma \frac{(TP_i + TN_i)}{(TP_i + TN_i + FP_i + FN_i)} \quad (10)$$

Precision: the proportion of true positive predictions among all positive predictions made by the model.

$$Precision = \frac{(TP)}{(TP + FP)} \quad (11)$$

Recall: the proportion of true positive predictions among all positive predictions made by the model.

$$Recall_i = \frac{(TP)}{(TP + FN)} \quad (12)$$

F1-Score: The F1-score is a measure that combines Precision and Recall into a single metric by calculating their weighted average.

$$F1_i = \frac{2\,(precisionxRecall)}{precision + Recall} \quad (13)$$

ROC_AUC: Receiver Operating Characteristic — Area Under Curve, measures a classification model's performance across different thresholds.

E. HYPER-PARAMETER TUNNING

Fine-tuning hyperparameters required careful modifications. Notably, our conclusion on the ideal feature count aligned with the DeepVanet [6] paper's conclusion. We found that using 64 EEG features and 16 face features yielded the best optimal results during testing and cross-validation by methodical grid search from {16, 32, 64, 128, 256, 512} features.

To determine the optimal configurations, we started by selecting a reasonable default architecture for the model,

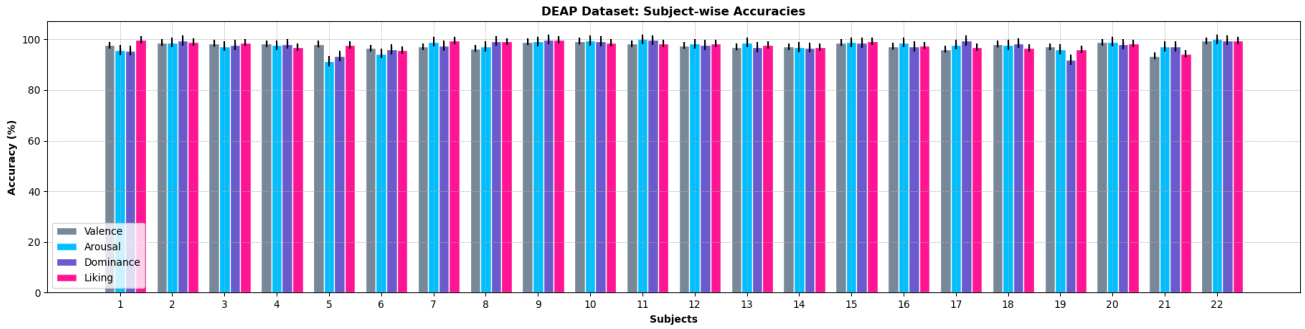

FIGURE 2. The mean accuracies for valence, arousal, dominance, and liking across all subjects, determined through 10-fold cross-validation, are presented. The error black bars illustrate the standard deviation across trials for each subject.

TABLE 1. Comparsion of EEG vs. proposed transformer for EEG classification.

Model	Valence %	Arousal %	Average %
Graph-based	60.18	59.19	59.68
LSTM	84.75	82.16	83.46
CNN+HMM	79.77	83.09	81.43
HOLO-FM	76.61	77.72	77.17
ERTNet	73.31	80.99	77.15
STS-Transformer	89.86	86.83	88.35
CNN multispectral	90.62	86.13	88.38
DCNN	87.84	87.69	87.77
EEG Classification	**91.42**	**92.98**	**92.20**

TABLE 2. Comparsion of existing fusion models vs. proposed fusion feature classification.

Deep Models	Arousal	Valance	Average
3D CNN Ensemble	96.13	96.79	96.46
CNN and Attention	96.63	98.18	97.40
Proposed Model	**97.64**	**97.78**	**97.71**

which included 4 encoder blocks, ReLU activation, and a dropout rate of 0.2. Next, we conducted a grid beam search over the learning rate and the number of features. The learning rates tested were {0.01, 0.001, 0.0001, 0.005, 0.05, 0.005}, and the feature sizes explored were {16, 32, 64, 128, 256, 512}. From this search, we identified that a learning rate of 0.001, combined with 16 facial features and 64 neurophysiological features, yielded good results over 10 epochs. Following this, we performed another grid search, this time focusing on the number of encoder blocks, activation functions, and dropout rates. We found that using 2 encoder blocks, ReLU activation, and a dropout rate of 0.1 produced the best results.

VI. RESULTS

A. MULTIMODAL FUSION RESULTS

By employing our feature-level fusion method, we attained the average recognition accuracy across two specific categories: Arousal and Valence as shown in TABLE 1.

As illustrated in TABLE 2, Multi-modal methods have demonstrated superior accuracy compared to similar fusion models from previous studies. We have run both inter-subject and per-subject models on our multimodal fusion model, achieving average accuracies of 97.71% and 97.94%, respectively, for inter-subject and per-subject evaluations. In addition, the model achieved 97.91% for Liking and 97.62% for Dominance. Despite the relatively low complexity of our fusion mechanism, it surpasses previous models. For instance, our fusion model outperforms the 3D CNN ensemble model [3], which achieved 96.13% accuracy for valence and 96.79% for arousal. The 3D CNN ensemble model employed a 3D CNN for face recognition and Mask R-CNN for face detection, experimenting with two fusion techniques and determining that stacking yielded the best performance, with accuracies of 96.13% for valence and 96.70% for arousal. Additionally, the CNN attention model [20] uses a pre-trained CNN combined with an attention mechanism to extract refined features, while a CNN is applied to extract EEG signals. After fusing the features, they are processed through a feed-forward block. This model achieved accuracies of 96.63% for valence and 97.15% for arousal classification on the DEAP dataset.

As illustrated in TABLE 3, the results demonstrate a significant progression in model performance across various architectures for the given task. Traditional models such as LGGNET-H [32], SVM [33], and BOOST [33] achieved F1 scores of 72.5, 74.8, and 77.7, respectively, indicating their moderate efficacy. Deep learning models such as LSTM [34] and DBN [35] showed notable improvements with F1 scores of 82.4 and 86.8, showcasing the potential of advanced neural architectures. The CNN-LSTM [36] model further raised the benchmark, achieving an impressive F1 score of 95.6. However, our proposed model outperformed all others, achieving a remarkable F1 score of 98.1, demonstrating its superior ability to capture and learn complex patterns in the data, setting a new standard for performance in this domain.

To further evaluate our fusion model, we achieved a precision of 0.979, a ROC AUC score of 0.976, an F1 score of 0.981, and a recall of 0.982, demonstrating the model's robust performance.

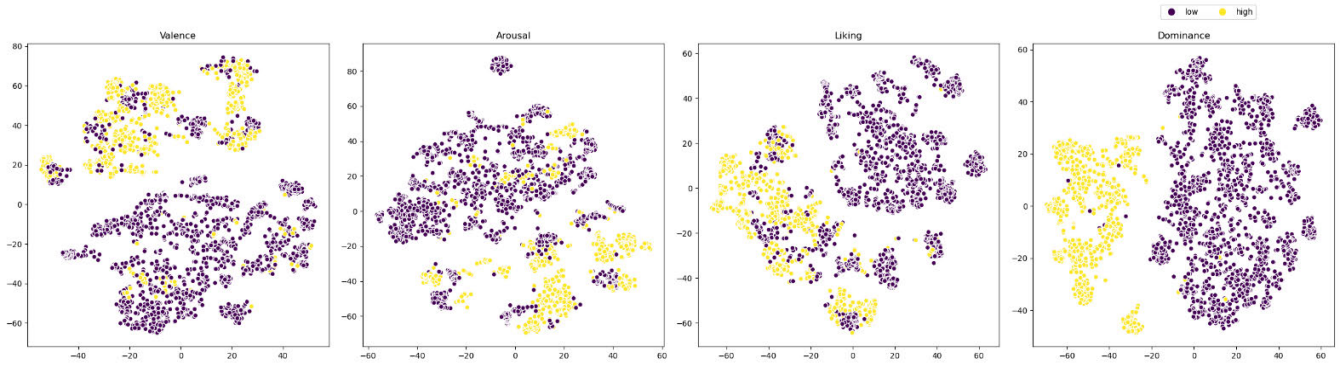

FIGURE 3. 2D t-SNE Visualization of DEAP Dataset Features: Subject 4's Emotional Landscape – Valence, Arousal, Liking, and Dominance Explored.

TABLE 3. Different evaluation metrics for fusion model.

Models	F1
LGGNET-H	72.5
SVM	74.8
BOOST	77.7
LSTM	82.4
DBN	86.8
CNN-LSTM	95.6
Ours Model	**98.1**

B. TRANSFORMER-BASED EEG CLASSIFICATION

TABLE 1 shows models that have solely used EEG data to predict emotions and compared the performance of our Transformer-based EEG classification. The CNN- multi-spectral [12] captures spatial and temporal features using multi-spectral images derived from EEG data. This model achieved accuracies of 90.62% for valence, 86.13% for arousal, 88.48% for dominance, and 86.23% for liking. A DCNN model [37], which combines a DCNN module with a ConvLSTM module, achieved accuracies of 87.84% for valence and 87.69% for arousal. our model surpasses the performance of all benchmark models, including an LSTM model, the CNN+HMM model [38] which combines convolutional neural networks (CNN) and hidden Markov models (HMM), the HOLO-FM model [39] which extracts holographic and topographic feature maps from EEG data and processes them through a CNN, the Ertnet model [40] which uses both CNNs and transformers to capture topological and spatio-temporal features, and the STS-Transformer model [41] which takes raw EEG data as input and uses a transformer architecture. Our Transformer-based EEG classifier performs comparably to single-modality approaches. This is significant because it demonstrates that we can leverage the Transformer-based approach for EEG classification.

To further test our model, we have run our model on another lie detection dataset. For the dataset in the lie detection task, our Transformer-based model achieved a precision of 0.9284, a recall of 0.9581, an F1 score of 0.9430, and an accuracy of 94.68%. The confusion matrix showed 373 true negatives, 30 false positives, 17 false negatives, and 389 true positives. Additionally, the model recorded a ROC AUC score of 0.9418, indicating strong discriminative performance.

VII. DISCUSSION

A. FUSION MODEL PERFORMANCE ON EEG DATA

Our extensive experiments underscore the remarkable capabilities of our proposed fusion model in accurately recognizing emotions from EEG data. Figure 2 illustrates the average accuracies obtained for each subject in the DEAP dataset. It shows that the emotional dimensions—valence arousal, liking, and dominance—perform consistently well, with standard deviations of the accuracies ranging from 1.31 to 1.96. This indicates the robustness of the model's architecture in capturing various details in the data. Additionally, the accuracy range is similar across all subjects, further demonstrating the model's reliability and consistency in emotion recognition.

The advantages of multimodality over single modality are evident in facial expression recognition. Our experiments with EmoNet, a deep neural network designed for facial feature analysis, show low average accuracies of 53.52% (V1) and 53.72% (V2) on the DEAP dataset. V1 uses a fixed threshold for labels, while V2 employs a dynamic threshold based on the distribution. These results highlight the difficulty of recognizing emotions in the DEAP dataset, where subtle facial expressions make it challenging for models to perform well. Similarly, single-modality EEG classification on the DEAP dataset achieves an average accuracy of 92.20%. By integrating additional modalities, such as neurophysiological signals and temporal information, we significantly improve emotion recognition accuracy, achieving 97.71% with our fusion model. Our fusion model, which deeply integrates neurophysiological and facial features, is crucial for enhanced emotion detection, especially for wearable devices in human-machine interaction.

B. COMPARISON WITH SIMILAR MODELS

When we compare our model with similar fusion models, it performs exceptionally well. The superiority of our

proposed model can be attributed to the following reasons. First, the use of a transformer-based architecture for EEG data in our model allows for the effective capture of complex spatial and temporal patterns. In addition, the multi-head attention block contributes to capture different dependencies in the data, allowing for more comprehensive feature extraction. Moreover, the facial expression extractor, utilizing a unique technique inspired by transformers, obtains meaningful spatial features from images, and LSTM further refines these features to capture temporal dependencies. Finally, the integration of the transformer-based model for EEG data and the transformer-inspired facial emotion recognition, followed by the fusion of their extracted features, provides a robust framework for emotion prediction. This architecture effectively captures and integrates diverse aspects of neurophysiological and visual data, resulting in improved accuracy and reliability of emotion prediction. Overall, as presented in Result section our proposed model exhibits superior accuracy compared to the evaluated methods, resulting in enhanced performance in emotion recognition tasks.

C. PERFORMANCE WITH LIMITED EEG CHANNELS

The high performance of the Transformer-based architecture is due to the fact that the features extracted by the Transformer model are more discriminative than handcrafted features, which is crucial for wearable devices with fewer channels. Upon investigating our Transformer-based EEG classifier, we found that it performs exceptionally well using only the FP1 and FP2 frontal channels of the EEG cap in the DEAP dataset, achieving over 85% accuracy. We also noticed that this accuracy can be improved by fusing these features with facial expression recognition, leveraging the strengths of both modalities to achieve significantly higher accuracy, exceeding 90%, compared to using a single modality. This is especially important for wearable devices with fewer EEG channels, where integrating computer vision and facial expression analysis can enhance the emotion recognition capabilities of our models.

D. VISUALIZATION OF MULTIMODAL FEATURES

As illustrated in Figure 3, we use 2D t-SNE to reduce the feature dimensions into a two-dimensional space for visualization. t-SNE is a non-linear clustering algorithm that aims to preserve the relative structure and pair-wise similarities of the data when projecting it onto a lower-dimensional space. This technique is employed to visualize higher-dimensional datasets, allowing us to understand their clustering and the patterns associated with the data. The HIGH and LOW class features distinctly form two clusters within this space. In the fusion modal, these clusters are almost clearly separated. This observation underscores the effectiveness of our multimodal features, demonstrating their salience and discriminative power.

E. APPLICATIONS IN DIGITAL HEALTH AND HEALTHCARE

The emotion recognition capabilities achieved by our fusion model have significant applications in digital health, healthcare, and human-computer interactions. The ability to measure emotional states through neurophysiological and facial recognition is poised to facilitate mental health diagnosis and treatment. Our experiments demonstrate that deep multimodal fusion of EEG data, even with a limited number of channels (such as the frontal channels in our case), can extract features that, when fused with facial data, improve emotion detection. Our model has the potential to be seamlessly integrated into wearable devices, offering continuous, non-invasive monitoring for individuals with mood disorders or anxiety. By analyzing subtle shifts in EEG patterns, these devices could provide real-time feedback, granting individuals profound insights into their emotional triggers and empowering them to proactively manage their mental health. For clinicians, access to this precise data would improve personalized treatment plans, enhance the tracking of therapeutic efficacy, and enable the early detection of mood fluctuations.

F. IMPACT ON HUMAN-COMPUTER INTERACTION (HCI)

Our model significantly impacts the field of affective computing by advancing systems that can accurately recognize, interpret, and respond to human emotions. This progress transforms human-computer interaction (HCI), enabling adaptive interfaces, emotionally intelligent virtual agents, and personalized gaming experiences. For instance, it allows online learning platforms to adjust lesson difficulty based on a student's detected frustration or boredom, optimizing engagement and retention. Video games can adapt in real-time to the player's excitement, ensuring a consistently thrilling experience. Additionally, virtual reality environments can use emotion recognition to modify stimulus intensity, enhancing immersion and tailoring experiences to individual preferences.

G. APPLICATIONS IN SOCIAL ROBOTICS

An emotion recognition system paves the way for social robots with emotional intelligence, enabling them to perceive and respond to human emotions in real-time. These emotionally intelligent robots enhance assistance for the elderly, offering emotional support and social interaction. In therapeutic settings, robots equipped with advanced emotion recognition capabilities can deliver personalized interventions for children with autism spectrum disorder, significantly enhancing their ability to understand and express emotions. For example, a robot can guide a child through calming breathing exercises or relaxation techniques upon detecting signs of anxiety or frustration, fostering emotional regulation skills.

H. AFFECTIVE COMPUTING VIA NEUROPHYSIOLOGY

We also introduce a Transformer-based architecture to capture features in neurophysiological data, validating its performance across both the DEAP and Lie Detection datasets. Our experiment demonstrates that the Transformer is able to generalize well for the classification of EEG signals. To further test its generalizability for affective design tasks, we trained

our model on both the DEAP and Lie Detection datasets, achieving high results across all investigated metrics.

Finally, this research enriches our comprehension of the neural mechanisms underlying emotion. By identifying the intricate patterns in facial and neurophysiological data that distinguish emotional states, we enhance the foundational knowledge in bio-sensing, and these insights have the potential to deepen our understanding of multimodal emotion recognition and apply this system to various applications such as healthcare, wearable devices, and human-computer interactions.

I. LIMITATIONS AND FUTURE WORK

Recognizing emotions through EEG signals is inherently complex due to the high variability in individual emotional responses. A ''one-size-fits-all'' approach may not yield reliable results, even with a large, diverse dataset. To address this, we applied our model across different datasets and conducted experiments in both per-subject and inter-subject settings. The personalized nature of EEG signals underscores the need for high-quality hardware to ensure accurate neurophysiological signal detection, which is crucial for advancing affective computing. As future work, it is essential to explore personalized models that account for individual variability in emotional responses. Additionally, hybrid approaches combining population-based models with real-time calibration should be investigated to achieve improved adaptability and performance.

The DEAP dataset assumes a stable emotional state over the duration of an entire music video, approximately one minute long. However, it is likely that arousal and valence levels fluctuate within shorter time frames, as brain signal responses are often dynamic and can stabilize within seconds. This mismatch in temporal granularity poses a challenge for accurately linking EEG responses to emotional states across the video length. Therefore, in the future, a key focus should be to develop models capable of capturing dynamic changes in emotional states over shorter time frames to align with EEG signal fluctuations. This includes incorporating time-series analysis techniques and adaptive segmentation to better represent temporal dynamics.

A significant limitation in this domain is the scarcity of datasets that include both neurophysiological signals and videos simultaneously. Overcoming this limitation requires concerted efforts to curate and share high-quality, synchronized datasets, paving the way for more robust and generalizable emotion recognition systems. Therefore, a key future direction is to curate high-quality, synchronized multimodal datasets that integrate neurophysiological signals, videos, and facial expressions.

VIII. CONCLUSION

This research presents a novel multimodal emotion recognition system integrating neurophysiological signals and facial data to enhance accuracy and reliability. The system processes face image sequences alongside EEG, EOG, GSR, BVP, RSP, EMG, SKT, and pulse wave signals to predict valence-arousal, liking, and dominance labels. Leveraging transformers and deep neural networks, it captures complex temporal and spatial patterns in the data. Experiments on the DEAP dataset, conducted on both per-subject and inter-subject bases, demonstrate superior performance compared to single-modality methods and parity with state-of-the-art multimodal approaches. The model achieved accuracy rates of 97.78% (arousal), 97.64% (valence), 97.91% (liking), and 97.62% (dominance), along with a precision of 0.979, ROC AUC of 0.976, F1 score of 0.981, and recall of 0.982, showcasing robust performance. A Transformer-based EEG classifier was also introduced, tested on the DEAP and Lie Detection datasets, demonstrating strong generalizability and effective performance with fewer EEG channels—a significant advantage for wearable devices. These findings highlight the potential of multimodal fusion in advancing emotion recognition technology, enabling empathetic and emotionally intelligent human-computer interactions and improving wearable technology usability.

IX. LIST OF ABBREVIATIONS

A list of the abbreviations introduced in this article is tabulated in TABLE 4.

TABLE 4. List of abbreviations.

Acronym	Full Name
AI	Artificial Intelligence
AEFW	Acted Facial Expressions in the Wild
BVP	Blood Volume Pulse
CNN	Convolutional Neural Network*
DEAP	Database for Emotion Analysis using Physiological signals*
EOG	Electro-Oculogram*
EMG	Electromyogram*
GSR	Galvanic Skin Response*
HMM	Hidden Markov Model*
LSTM	Long Short-Term Memory*
ROC AUC	Receiver Operating Characteristic Area Under Curve*
RSP	Respiration*
SKT	Skin Temperature*
t-SNE	t-distributed Stochastic Neighbor Embedding*
ERTNet	Explainable and Reliable Transformer Network*

REFERENCES

[1] F. Safavi, P. Olikkal, D. Pei, S. Kamal, H. Meyerson, V. Penumalee, and R. Vinjamuri, ''Emerging frontiers in human–robot interaction,'' *J. Intell. Robotic Syst.*, vol. 110, no. 2, p. 45, 2024.

[2] F. Safavi, D. Pei, P. Olikkal, and R. Vinjamuri, ''New horizons in human–robot interaction: Synergy, cognition, and emotion,'' in *Discovering the Frontiers of Human–Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control*, R. Vinjamuri, Ed., Cham, Switzerland: Springer, 2024, pp. 103–133, doi: 10.1007/978-3-031-66656-8_5.

[3] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[4] M. Aslan, M. Baykara, and T. B. Alakus, "LieWaves: Dataset for lie detection based on EEG signals and wavelets," *Med. Biol. Eng. Comput.*, vol. 62, no. 5, pp. 1571–1588, May 2024, doi: 10.1007/s11517-024-03021-2.

[5] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," *Egyptian Informat. J.*, vol. 22, no. 2, pp. 167–176, Jul. 2021, doi: 10.1016/j.eij.2020.07.005.

[6] Y. Zhang, M. Z. Hossain, and S. Rahman, "DeepVANet: A deep end-to-end network for multi-modal emotion recognition," in *Proc. 18th IFIP TC Int. Conf. Hum.-Comput. Interact. (INTERACT)*, Bari, Italy, Sep. 2021, pp. 227–237, doi: 10.1007/978-3-030-85613-7_16.

[7] F. Safavi, K. Patel, and R. K. Vinjamuri, "Towards efficient deep learning models for facial expression recognition using transformers," in *Proc. IEEE 19th Int. Conf. Body Sensor Netw. (BSN)*, Oct. 2023, pp. 1–4, doi: 10.1109/bsn58485.2023.10331041.

[8] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, and Y. Bi, "EEG-based emotion classification using a deep neural network and sparse autoencoder," *Frontiers Syst. Neurosci.*, vol. 14, p. 43, Sep. 2020, doi: 10.3389/fnsys.2020.00043.

[9] K. Patel, F. Safavi, R. Chandramouli, and R. Vinjamuri, "Transformer-based emotion recognition with EEG," in *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2024, pp. 1–4, doi: 10.1109/embc53108.2024.10781700.

[10] Z. Fan, F. Chen, X. Xia, and Y. Liu, "EEG emotion classification based on graph convolutional network," *Appl. Sci.*, vol. 14, no. 2, p. 726, Jan. 2024, doi: 10.3390/app14020726.

[11] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019, doi: 10.1016/j.inffus.2018.09.008.

[12] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio–visual emotion recognition," 2021, *arXiv:2103.09154*.

[13] J. D. S. Ortega, P. Cardinal, and A. L. Koerich, "Emotion recognition using fusion of audio and video features," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3847–3852, doi: 10.1109/SMC.2019.8914655.

[14] Q. Guo, Y. Liao, Z. Li, and S. Liang, "Multi-modal representation via contrastive learning with attention bottleneck fusion and attentive statistics features," *Entropy*, vol. 25, no. 10, p. 1421, Oct. 2023, doi: 10.3390/e25101421.

[15] Z. Zhang, P. Zhao, E. Park, and J. Yang, "MART: Masked affective RepresenTation learning via masked temporal distribution distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 12830–12840, doi: 10.1109/cvpr52733.2024.01219.

[16] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.

[17] B. Li, H. Fei, L. Liao, Y. Zhao, C. Teng, T. S. Chua, D. Ji, and F. Li, "Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 5923–5934, doi: 10.1145/3581783.3612053.

[18] B. Li, H. Fei, F. Li, T. Chua, and D. Ji, "Multimodal emotion-cause pair extraction with holistic interaction and label constraint," *ACM Trans. Multimedia Comput., Commun. Appl.*, 2024, doi: 10.1145/3689646.

[19] M. Luo, H. Fei, B. Li, S. Wu, Q. Liu, S. Poria, E. Cambria, M. L. Lee, and W. Hsu, "PanoSent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 7667–7676, doi: 10.1145/3664647.3680705.

[20] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent emotion memory for multi-label emotion classification," in *Proc. 34th AAAI Conf. Artif. Intell.*, pp. 7692–7699, 2020, doi: 10.1609/aaai.v34i05.6271.

[21] H. Fei, Y. Ren, Y. Zhang, and D. Ji, "Nonautoregressive encoder–decoder neural framework for end-to-end aspect-based sentiment triplet extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5544–5556, Sep. 2023, doi: 10.1109/TNNLS.2021.3129483.

[22] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T. Chua, "Reasoning implicit sentiment with chain-of-thought prompting," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 1171–1182.

[23] H. Fei, T.-S. Chua, C. Li, D. Ji, M. Zhang, and Y. Ren, "On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training," *ACM Trans. Inf. Syst.*, vol. 41, no. 2, pp. 1–32, 2022.

[24] M. Staffa and S. Rossi, "Enhancing affective robotics via human internal state monitoring," in *Proc. 31st IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Sep. 2022, pp. 884–890, doi: 10.1109/RO-MAN53752.2022.9900762.

[25] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "TransFuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878–12895, Nov. 2023, doi: 10.1109/TPAMI.2022.3200245.

[26] F. Safavi, V. R. Venkannagari, and R. K. Vinjamuri, "Deep multimodal emotion recognition: Fusion of facial features and neurophysiological signals," in *Proc. IEEE 20th Int. Conf. Body Sensor Netw. (BSN)*, Oct. 2024, pp. 1–4, doi: 10.1109/bsn63547.2024.10780554.

[27] F. Safavi, T. Chowdhury, and M. Rahnemoonfar, "Comparative study between real-time and Non-Real-Time segmentation models on flooding events," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 4199–4207, doi: 10.1109/BigData52589.2021.9671314.

[28] F. Safavi and M. Rahnemoonfar, "Comparative study of real-time semantic segmentation networks in aerial images during flooding events," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 15–31, 2023, doi: 10.1109/JSTARS.2022.3219724.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[30] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf

[31] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, Sep. 2017, doi: 10.1016/j.imavis.2017.02.001.

[32] Y. Ding, N. Robinson, C. Tong, Q. Zeng, and C. Guan, "LGGNet: Learning from local-global-graph representations for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9773–9786, Jul. 2024, doi: 10.1109/TNNLS.2023.3236635.

[33] M. Khateeb, S. M. Anwar, and M. Alnowami, "Multi-domain feature fusion for emotion classification using DEAP dataset," *IEEE Access*, vol. 9, pp. 12134–12142, 2021, doi: 10.1109/ACCESS.2021.3051281.

[34] R. D. Gaddanakeri, M. M. Naik, S. Kulkarni, and P. Patil, "Analysis of EEG signals in the DEAP dataset for emotion recognition using deep learning algortihms," in *Proc. IEEE 9th Int. Conf. Converg. Technol. (I2CT)*, Apr. 2024, pp. 1–7, doi: 10.1109/i2ct61223.2024.10543369.

[35] H. Xu and K. N. Plataniotis, "Affective states classification using EEG and semi-supervised deep learning approaches," in *Proc. IEEE 18th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2016, pp. 1–6, doi: 10.1109/MMSP.2016.7813351.

[36] M. Asif, S. Mishra, M. T. Vinodbhai, and U. S. Tiwary, "Emotion recognition using temporally localized emotional events in EEG with naturalistic context: DENS# dataset," *IEEE Access*, vol. 11, pp. 39913–39925, 2023, doi: 10.1109/ACCESS.2023.3266804.

[37] Y. An, N. Xu, and Z. Qu, "Leveraging spatial–temporal convolutional features for EEG-based emotion recognition," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102743, doi: 10.1016/j.bspc.2021.102743.

[38] Q. Zhong, Y. Zhu, D. Cai, L. Xiao, and H. Zhang, "Electroencephalogram access for emotion recognition based on a deep hybrid network," *Frontiers Hum. Neurosci.*, vol. 14, Dec. 2020, Art. no. 589001, doi: 10.3389/fnhum.2020.589001.

[39] A. Topic and M. Russo, "Emotion recognition based on EEG feature maps through deep learning network," *Eng. Sci. Technol., Int. J.*, vol. 24, no. 6, pp. 1442–1454, Dec. 2021, doi: 10.1016/j.jestch.2021.03.012.

[40] R. Liu, Y. Chao, X. Ma, X. Sha, L. Sun, S. Li, and S. Chang, "ERTNet: An interpretable transformer-based framework for EEG emotion recognition," *Frontiers Neurosci.*, vol. 18, Jan. 2024, Art. no. 1320645, doi: 10.3389/fnins.2024.1320645.

[41] W. Zheng and B. Pan, "A spatiotemporal symmetrical transformer structure for EEG emotion recognition," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105487, doi: 10.1016/j.bspc.2023.105487.

FARSHAD SAFAVI (Member, IEEE) received the B.Eng. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 2012, and the master's degree in electrical and computer engineering from the University of Toronto, Toronto, ON, in 2018. He is currently pursuing the Ph.D. degree in computer science with the University of Maryland at Baltimore County, Baltimore, MD, USA. He is also a Research Assistant with the University of Maryland. His research focuses on the development of advanced machine learning and deep learning models for human–robot interaction, human–computer interaction, and wearable technology. He has a particular interest in multimodal deep learning, affective computing, and computer vision. His work also explores the optimization of deep learning algorithms for embedded systems and the design of lightweight models for wearable and edge devices.

VIKAS REDDY VENKANNAGARI (Member, IEEE) received the Bachelor of Technology degree in computer science and engineering from the Institute of Aeronautical Engineering, India, in 2023. He is currently pursuing the M.S. degree in data science with the University of Maryland at Baltimore County. He is also a Research Assistant with the University of Maryland at Baltimore County. His research focuses on using deep learning to advance affective computing, with an emphasis on enhancing emotionally responsive robotic systems.

DEV PARIKH received the Bachelor of Engineering degree in computer engineering from Gujarat Technological University, India, in 2023. He is currently pursuing the M.S. degree in data science with the University of Maryland at Baltimore County. His research focuses on leveraging machine learning and brain–computer interfaces (BCI) for advanced robotic and automated system control to assist individuals with mobility impairments.

RAMANA KUMAR VINJAMURI (Senior Member, IEEE) received the B.S. degree from Kakatiya University, India, in 2002, the M.S. degree from Villanova University, Villanova, PA, USA, in 2004, and the Ph.D. degree from the University of Pittsburgh, Pittsburgh, PA, in 2008, all in electrical engineering. He was a Research Associate with the Department of Physical Medicine and Rehabilitation, University of Pittsburgh, from 2008 to 2012. He was a Research Assistant Professor with the Department of Biomedical Engineering, Johns Hopkins University, from 2012 to 2013. He was an Assistant Professor with the Department of Biomedical Engineering, Stevens Institute of Technology, from 2013 to 2020. He was an Assistant Professor with the Department of Computer Science and Electrical Engineering (CSEE), University of Maryland at Baltimore County (UMBC), from 2020 to 2023. Currently, he is a tenured Associate Professor with UMBC CSEE. He also holds a secondary appointment as a Visiting Professor with Indian Institute of Technology, Hyderabad, India, and an Adjunct Faculty with the Department of Physiotherapy, Manipal Academy of Higher Education, India. He received the NSF CAREER Award, in 2019; the NSF IUCRC "BRAIN" Planning Grant, in 2020; the NSF I-Corps Award, in 2024; and the NSF IUCRC "BRAIN" Center Grant, in 2024.