IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 December 2024, accepted 27 January 2025, date of publication 4 February 2025, date of current version 7 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3538642

RESEARCH ARTICLE

Multimodal Emotion Recognition: Emotion Classification Through the Integration of EEG and Facial Expressions

SONGÜL ERDEM GÜLER AND FATMA PATLAR AKBULUT, (Member, IEEE)
Department of Computer Engineering, Istanbul Kültür University, 34158 İstanbul, Türkiye
Corresponding author: Songül Erdem Güler (2100008117@stu.iku.edu.tr)

ABSTRACT Despite advances in the field of emotion recognition, the research field still faces two main limitations: the use of deep models for increasingly complex calculations and the identification of emotions through various data types. This study aims to advance the knowledge on multimodal emotion recognition by combining electroencephalography (EEG) signals with facial expressions, using advanced models such as Transformer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The results validate the effectiveness of this approach, demonstrating the high accuracy of the Gated Recurrent Unit (GRU) model, which achieved an average of 91.8% classification accuracy on unimodal (EEG-only) data and an average of 97.8% classification accuracy on multimodal (EEG and facial expressions) datasets in the multi-class emotion categories. The findings emphasize that by applying a multi-class classification framework, multimodal approaches offer significant improvements over traditional unimodal techniques. This work presents a framework that captures complex neural dynamics and visible emotional cues, enhancing the robustness and accuracy of emotion recognition systems. These results have important practical implications, showing how integrating various data sources with advanced models can overcome the limitations of single-modality systems.

INDEX TERMS Human computer interaction, emotion recognition, deep learning, EEG signals, facial expressions.

I. INTRODUCTION

Since emotions have different impacts on life, they greatly affect how we see the world and interact with others. The intricate interaction of emotions not only gives us unique experiences but also influences how we see our surroundings. Emotions shape personal interactions and impact our environmental views; therefore, our memory, decision-making, and relationships depend on them. This important impact cuts across boundaries and penetrates the dynamic field of human-computer interaction, where the understanding and application of emotions become essential in ever-changing and complex situations. We see the rapid progress of interactive systems that identify and respond to our emotional states [1]. Emotional and technological fusion has enormous potential to completely transform human-computer interaction. In this context, robots that can adjust to our emotional states, customize responses according to our internal circumstances, and combine the digital and human domains have the potential to be very useful instruments for promoting deeper connections.

EEG signals utilize the electrical activity patterns in the brain with great accuracy in terms of time, facilitating the observation of the neural processes that control emotional responses [2], [3], [4]. A detailed knowledge of the intricate relationship between emotion and cognition is made possible by the identification of different electrocortical markers associated with emotional states [5], [6]. Apart from the EEG signal, facial expressions, which are outward manifestations of inner emotional states, provide essential

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

information for emotion recognition systems. Our faces are a rich source of emotional information that displays many microexpressions and muscle movements [7]. While facial expressions offer a recognizable and comprehensible set of emotional cues, EEG signals offer insight into the complex neural responses [8]. The relationship between facial expression and electroencephalography advances the field of emotion identification and highlights the importance of this research's multidisciplinary approach. By fusing knowledge from computer vision and neuroscience, this method goes beyond the constraints of single modality systems and opens the door for more reliable and context-sensitive emotion-aware systems.

This work aims to improve the efficiency of emotion classification in human-computer interaction using data fusion techniques. In this specific context, the goal is to overcome the limitations of traditional emotion recognition methods that rely solely on assessing isolated components such as facial expressions and voice tones. Instead, the emphasis is on combining EEG signals and facial expressions to create a more comprehensive approach. The study utilizes the DEAP dataset comprising synchronized EEG and face video recordings. In this paper, Russell's Valence-Arousal model, a widely used dimensional emotion model, was adopted to classify emotional states based on valence and arousal levels. This model provides a framework for representing emotional states as points on a two-dimensional plane. The valence ranges from positive, which is the axis of happiness, down to negative, which can be characterized as sadness; the axis of arousal also ranges from low, indicating calmness, to high, which may mean excitement and stress. By positioning each emotion on the valence-arousal axes [9], the model enables the quantitative definition of emotions. According to this scale, emotional states can be represented in four main categories on a two-dimensional plane: High Valence-High Arousal (HVHA), High Valence-Low Arousal (HVLA), Low Valence-High Arousal (LVHA), and Low Valence-Low Arousal (LVLA). This categorization provides a more granular analysis of emotional states compared to commonly used binary classifications in existing literature. Additionally, each category corresponds to distinct states. HVHA represents positive emotions with high energy, such as enthusiasm and happiness. HVLA defines positive emotions with low energy, such as peace and relaxation. LVHA includes negative emotions with high energy, such as anger and fear, while LVLA describes negative emotions with low energy, such as sadness and depression. This model facilitates the dimensional analysis of emotional states and allows the application of a multimodal approach by integrating various data sources, such as EEG signals and facial expressions, in emotion classification studies.

Previous studies [10], [11] investigating the relationship between EEG-based graph-theoretical measures and discrete emotional states highlight the potential of EEG signals in distinguishing complex emotional conditions. Additionally, these studies provide a suitable perspective for further improving emotion classification through multimodal approaches. In this work, discrete emotional states such as fear, anger, happiness, sadness, and disgust were analyzed using facial video recordings, and the results were integrated with EEG signals.

The experimental results show that multimodal deep learning models trained with the combination of EEG signals, which provide direct insights into brain activities, and facial expressions, which reflect the outward manifestation of emotions, outperform unimodal methods in emotion classification. The main contributions of this study are as follows:

- In this research, custom transformation functions were developed and a detailed data analysis was conducted to address the dimension mismatch that arises when combining facial expression data with EEG data.
- The study analyzed emotion detection results obtained by the MTCNN and HaarCascade face detection algorithms on participant video recordings. By comparing both algorithms, selecting the emotion detection results of the algorithm that maximized the use of facial expression data and detected the highest number of emotions enabled more accurate emotion recognition.
- When EEG signals and facial expression data are combined, multimodal emotion recognition methods outperform unimodal approaches. Although the accuracy rate of the unimodal GRU model (91.8%) is good for general emotion inference, the performance of the multimodal GRU model (97.8%) demonstrates the value of integrating multiple data sources to better represent emotional states.
- In cases where the facial expression data extracted from some participants' videos were insufficient, synthetic data were generated to fill in the gaps and match the amount of data with the corresponding EEG signal data.
- Research in the literature using the DEAP dataset often classifies emotions into two classes, mainly valence and arousal. In this work, however, multiclass classification was performed for four different classes (HVHA, HVLA, LVHA, LVLA), involving both valence and arousal axes along with their high and low states.

II. LITERATURE REVIEW

Understanding emotions is becoming more and more important for developing natural and harmonious interactive systems. Multimodal emotion recognition has been greatly advanced by research combining several modalities. Using fusion techniques at the feature and decision levels, Zheng et al. [12] showed, for example, that combining EEG and eye movement data could improve emotion recognition performance. Specifically, the average accuracy rates achieved by EEG-based classification and eye movement-based classification were 71.77% and 58.90%, respectively, while different fusion strategies yielded rates of 73.59% and 72.98%. Multimodal approaches have promise,

as demonstrated by this integration, which achieved higher accuracy rates than using EEG or eye movement data alone.

Liu et al. [13] achieved significant progress in this field by utilizing a Bimodal Deep Autoencoder (BDAE) network to combine EEG signals and eye movement data. On the SEED and DEAP datasets, the researchers achieved accuracy rates of 53.9% and 85.1%, respectively, using unimodal methods. By integrating the data, they improved these rates to 91.01% and 83.25%, achieving high accuracy levels. Comparably, Ma et al. [14] investigated the relationship between EEG and other physiological signals using a multimodal residual LSTM network. Enhancing their model with weight-sharing mechanisms and residual learning, they showed how well deep learning-based multimodal methods work to enhance emotion recognition performance, achieving accuracy rates of 92.87% and 92.30% in the arousal and valence dimensions, respectively. Significant progress in emotion recognition technology has also resulted from the combination of EEG and facial expressions. These modalities were combined in studies by Huang et al. [15] and Muhammad et al. [16]. Huang et al. achieved a prediction accuracy of 74.38% using only EEG data and 66.38% using only facial expression data, both of which are unimodal methods, for the four basic emotional states in the their custom dataset. However, by implementing two different data fusion techniques, they improved these accuracy rates to 81.25% and 82.75%. This demonstrated that the classification accuracy obtained through multimodal approaches, combining facial expression and EEG data, surpasses the accuracy achieved with unimodal classification using only EEG or only facial expression data. Muhammad et al. compared the performance of various deep learning methods, including Deep Canonical Correlation Analysis (DCCA), SVM classifiers, and multi-task CNNs, for three different emotion classes (happy, neutral, and sad) on the DEAP and MAHNOB-HCI datasets. They also demonstrated the positive effects of these methods on the classification results compared to unimodal methods. The best results were achieved with the DCCA method, yielding average accuracy rates of 91.54% on the DEAP dataset and 93.86% on the MAHNOB-HCI dataset. Using only EEG data, average accuracy rates of 74.57% and 83.27% were achieved on the DEAP and MAHNOB-HCI datasets, respectively, while using only facial expressions, accuracy rates of 90.5% and 92.4% were obtained on the same datasets, respectively. With these results, the authors demonstrated the advantages of the DCCA multimodal fusion method over unimodal methods on different datasets.

The benefits of combining facial expressions with EEG signals were further highlighted by research by Wu et al. [17] and and Rayatdoost et al. [18]. Wu et al. applied hierarchical LSTM techniques to data obtained by multimodally combining facial video features with corresponding EEG signals. They claimed that their proposed emotion recognition model achieved at least a 2% increase in classification accuracy and a 0.015 improvement in F1 score compared to other methods. Rayatdoost et al. proposed a model fusion approach called Gated Coordinated Information Fusion. To evaluate the performance of this approach, they utilized two different datasets: DAI-EF and MAHNOB. In their study, facial expression and EEG data were used to perform emotion classification in valence and arousal categories. The results of the proposed model were compared to those of some models trained on unimodal data. For the valence and arousal categories, the fusion-based method achieved accuracy rates of 75.4% and 63.9% on the DAI-EF dataset, and 66.9% and 55.22% on the MAHNOB dataset, respectively. Unimodal EEG-CNN model, trained only on EEG signals, obtained accuracy rates of 69.5% and 61.4% for valence and arousal on the DAI-EF dataset, and 64.8% and 50.3% on the MAHNOB dataset, respectively. The study suggested that model fusion approaches could enhance emotion classification performance based on results obtained from different datasets.

Chaparro et al. [19] combined EEG and facial expression data from the MAHNOB-HCI dataset at the feature level and trained both a Neural Network model and a Random Forest Classifier on this data. Using the Random Forest Classifier, which achieved better results, they obtained a classification accuracy of 97.7% with multimodal data, 93.9% with EEG signals alone, and 91.3% with facial expression features alone.

The potential of combining several data sources was demonstrated by Keshari and Palaniswamy [20] who created a bimodal model combining facial expressions and upper body gestures. In the study, a classification accuracy of 86% was achieved with facial expressions alone and 83.57% with EEG data alone. This accuracy rate was increased to 94% with the proposed bimodal method. Tan et al. [21] and Cimtay et al. [22] both found that multimodal methods improved accuracy and robustness, especially in situations involving deceptive facial expressions or human-robot interaction systems. For instance, Cimtay et al. introduced hybrid fusion methods capable of working with different emotional states on the DEAP dataset. Using this method, they performed both multimodal and unimodal classification based on facial data. The mean one-subject-out accuracy achieved using unimodal approaches was 37.11%, whereas this rate increased to 53.87% with multimodal approaches. Tan et al. integrated their method into an HRI system. On their custom dataset, they achieved an average emotion classification accuracy of 79.25% using only EEG data. On the other hand, their multimodal approach, combining facial expressions and EEG data, resulted in a recognition rate of 83.33%. These thorough approaches highlight the possibility of multimodal techniques to improve human-computer interaction technologies by offering a more complex comprehension of emotional states.

In conclusion, the field of emotion recognition has been much advanced by the integration of EEG and facial

expressions using advanced deep learning methods. These techniques have been improved further by studies by Wang et al. [23] and Chen et al. [24], who have included hybrid models and attention mechanisms to attain high classification accuracy. Wang et al. achieved classification accuracies of 85.72% and 87.92% for the valence and arousal dimensions, respectively, using unimodal EEG data on the DEAP dataset. However, by combining EEG and facial expression data multimodally, they improved these accuracy rates to 96.63% and 97.15%, respectively. The efficiency of merging EEG signals and video data was shown by Chakravarthy et al. [25] in their optimization-supported multimodal model. Their model achieved a high accuracy rate of 90.84% using a hybrid fusion strategy. Together, these efforts demonstrate the significant advancements in multimodal emotion recognition and the benefits of combining several modalities and advanced models to improve the robustness and accuracy of emotion recognition systems.

III. METHODOLOGY

A. DATASET OVERVIEW

In this work, the extensively used DEAP dataset for emotion recognition has been preprocessed. Using peripheral physiological data and electroencephalography (EEG) signals, DEAP seeks to understand emotional reactions to multimedia stimuli. The dataset consists of 32-channel EEG recordings together with physiological signals including skin temperature, blood volume pulse (BVP), respiration rate, galvanic skin response (GSR), and electrooculogram (EOG) as well as facial expression video clips. The facial expressions of the first 22 volunteers were also videotaped [9]. The data in the preprocessed DEAP were re-referenced to the average reference, downsampled, had EOG artifacts eliminated, had a 4.0–45.0 Hz band-pass filter applied, and were split into 60-second trials with a 3-second baseline eliminated. With measurements of valence, which indicate how pleasant or unpleasant (positive or negative) emotions are, and arousal, which indicates how excited or calm one is, the DEAP dataset provides a thorough framework. Understanding the intricacy of human emotional experience is made possible in large part by this multidimensional approach. Applications of the dataset include the creation of emotion classification algorithms, analysis of emotional reactions to multimedia material, and investigation of the interaction between several modalities in emotion recognition.

B. EEG-BASED EMOTION RECOGNITION

Through the analysis of EEG data, a novel method for precisely categorizing emotions was developed. To this end, a complete preprocessing pipeline was built for the EEG signal, and distinct machine-learning models were evaluated. The reliability of subsequent research and the integrity of the dataset depend on data preprocessing. Initially, EEG signals were systematically preprocessed using an algorithm that was painstakingly developed and included participant identifiers, EEG channel specifications, segment duration, and inter-segment distance parameters. The window and step size of 256 were carefully selected so that every step corresponds to one second. Thus, the EEG signals were split into 256 data sample, one-second segments. From 40 participants' EEG channels, the function segments the data. As seen by Equation 1, the windowing process starts from the initial variable and continues for the length of the window, increasing the starting position at each step to traverse the data.

$$start\ position = start\ position + window\ size \quad (1)$$

This method was designed to effectively record the dynamics of emotional states and make it easier to analyze the features of EEG data over time windows. Then, every data segment was labeled using the participant's emotional state at that specific moment. A one-dimensional form was created from the four distinct numerical groups of the two-dimensional labels, valence and arousal values. This modification allows a numerical and simplified representation of emotional states, which facilitates data processing and analysis.

Algorithm 1 Label Mapping Based on Valence and Arousal

Require: valence, arousal
Ensure: label
if valence $>$ 5 and arousal $>$ 5 **then**
  label $\leftarrow$ 0 {HVHA}
else if valence $>$ 5 and arousal $\leq$ 5 **then**
  label $\leftarrow$ 1 {HVLA}
else if valence $\leq$ 5 and arousal $>$ 5 **then**
  label $\leftarrow$ 2 {LVHA}
else
  label $\leftarrow$ 3 {LVLA}
end if
return label

With the `label_mapping` function, which was presented in Algorithm 1, the label transformation process was carried out. This function produces a matching label depending on the numerical values of valence and arousal in DEAP dataset as input. Emotional states were categorized into four discrete classes based on predefined thresholds for high and low valence/arousal scores. The operation takes the subsequent course:

- HVHA refers to cases where both valence and arousal values are greater than 5.
- HVLA encompasses cases where the valence value is greater than 5, while the arousal value is 5 or lower.
- LVHA includes cases where the valence value is 5 or lower, while the arousal value is greater than 5.
- LVLA refers to cases where both valence and arousal values are 5 or lower.

With this approach, the dataset is ensured to contain both EEG signals and one-dimensional labels suitable for multi-class classification.

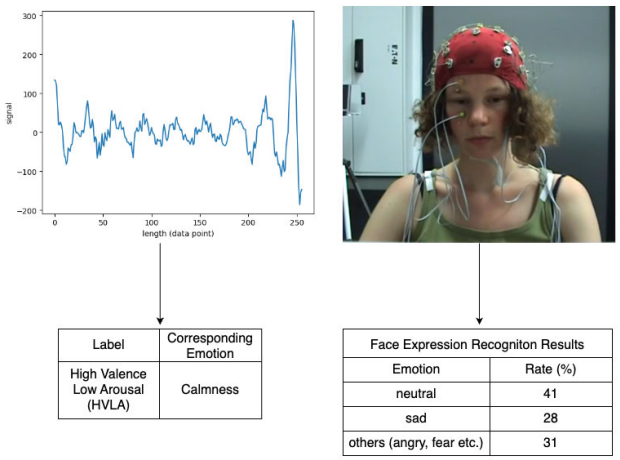


FIGURE 1. An illustrative depiction of EEG signal processing alongside the corresponding facial image and resulting recognition outcomes for subject 11.

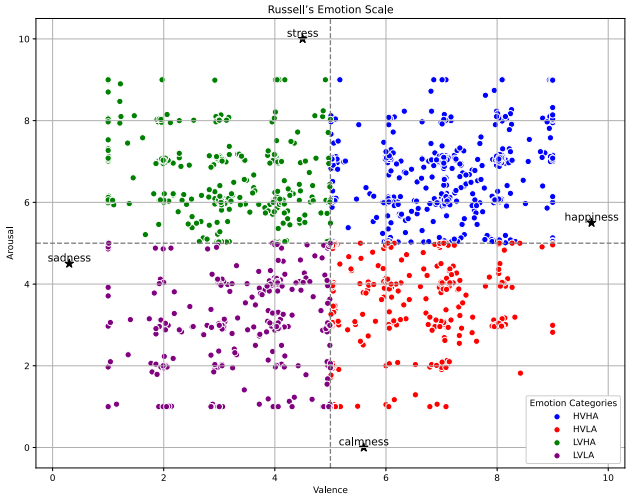


FIGURE 2. Russell's emotion scale.

A one-second segment from a video watched by Participant ID 11 has been visualized in terms of EEG data for this time frame. It was determined that the corresponding emotion label for this EEG data is HVLA. The emotions represented by the four different labels (HVHA, HVLA, LVHA, LVLA) corresponding to EEG signals were identified using Russell's emotion scale (Figure 2). In this context, the emotion associated with the HVLA label was determined to be calmness. Facial expression recognition analyzes conducted on the image taken from the video of the facial expression during the same one-second EEG signal interval are presented under ''Face Expression Recognition Results'' in Figure 1. These analyses determined that the highest intensity emotion detected in the image frame was neutral at 41%. EEG signals and facial expression recognition results were evaluated together to perform a comprehensive emotional analysis.

Both unimodal and multimodal emotion classification datasets were stratified into 70% for training, 15% for validation, and 15% for testing. While the data were split in the same proportions for both unimodal and multimodal studies, the amount of data used in each method differed. In the multimodal approach, the differing dimensions of facial expression data and EEG data from the DEAP dataset led to a change in the overall data dimension when they were combined. This is explained in more detail in Section III-C. The number of samples used for each method is presented numerically in Table 1. For the purpose of guaranteeing model consistency, this partitioning procedure was carried out once, and all models were trained using the same datasets. The datasets were visualized and analysis procedures were carried out using these visualizations. To avoid hierarchical relationships among the class labels, the class labels in the training, validation, and test sets were transformed into binary class matrices (0 and 1). This allowed the model to evaluate each class as a distinct category and to classify them accurately.

TABLE 1. Distribution of samples across the training, validation, and test sets.

Dataset type	Unimodal	Multimodal
Train	388864	267344
Validation	83328	57288
Test	83328	57288

The data then went through procedures of normalization, scaling, and reshaping. Scaling and normalization were done to guarantee uniformity among participants and various EEG channels. Although normalization adjusts the data to have a mean of 0 and a standard deviation of 1, min-max scaling scales the data between 0 and 1. While preventing overfitting, these procedures maximize the training process of the model. Ultimately, the data were reorganized to fit the model, and where needed, data augmentation methods were used. Improved efficiency and classification performance of the model depend on these augmentation and reshaping operations. Carefully planned to maximize information extraction from the EEG data, each of these phases guarantees the correct categorization of emotional states. Emotion recognition research can produce better results when the preprocessing pipeline is implemented with care since it improves machine learning model performance.

C. MULTIMODAL EMOTION CLASSIFICATION: FACIAL EXPRESSIONS AND EEG SIGNALS

In this section, we focus on integrating preprocessed EEG data from the DEAP dataset with emotion analysis results from participants' facial videos. Within the DEAP dataset, a Facial Expression Recognition (FER) algorithm analyzed participants' facial responses to 40 different videos. FER uses advanced image processing techniques to detect micro-expressions on the human face and categorizes them into predetermined emotion groups [26]. Since only 22 participants consented to the video recording of their faces, only the facial videos of these participants were used for analysis. Additionally, although each participant was supposed to have

40 facial videos, technical issues during recording the last few videos resulted in fewer videos for some participants.

The videos, containing a total of 3000 images (50 FPS x 60 seconds), were processed using the FER tool. The analysis results, including facial coordinates (x, y, w, h) and 7 different emotional states ('anger', 'disgust', 'fear', 'happiness', 'sadness', 'surprise', 'neutral'), were saved as tables in files within the corresponding participant ID folders.

This work used the FER tool for emotion analysis of facial expressions together with the MTCNN [27] face detection algorithm. MTCNN recognizes faces and sends the coordinates of those faces to the FER program for emotional analysis. We repeated this procedure for every participant's facial video. Nevertheless, occasionally MTCNN's failure to produce the expected face detection results or FER's failure to precisely extract emotional expressions resulted in some videos having fewer than the anticipated 3000 rows in the emotion analysis datasets. Each participant watched 40 distinct videos, which are represented by the 17360 processed EEG signal samples in the DEAP dataset. The same amount of EEG signal samples was obtained from each participant. A participant's total number of EEG signal samples divided by the number of videos yields 434 EEG signal samples taken from each video. This computation suggests that each video analysis should have produced an emotion analysis dataset of at least 434 samples during the merging of EEG signals with facial expression data. The anticipated number of samples was not reached by the facial analysis results from some videos, though. The Haar Cascade [28] algorithm was applied in the FER tool for face detection to solve this problem. With this method, the videos with missing data were reprocessed and some results with more data were obtained.

At this point, if the earlier results had more data and were larger in size, they were substituted for the recently produced results. These results were preferred if those obtained with the MTCNN algorithm offered more or equal information than those obtained with Haar Cascade. Thus, in every case, the most thorough analysis results were attained.

After these processes, a function was developed to produce synthetic data in situations where the emotion analysis results taken from the facial videos could not produce a minimum of 434 samples. This function generates synthetic data until the specified target length is reached, starting from the original dataset. During synthetic data generation, extra data samples are produced, equal to the difference between the length of the current dataset and the target length. Each synthetic data sample is created by applying a specific modification to the numerical values of a randomly selected row from the original dataset. This modification involves adding a random amount of "noise" generated by a normal distribution with a mean of zero and a standard deviation of 0.01. This method aims to create more realistic and diverse synthetic data by adding small variations to the dataset. Since the values in the original dataset are between 0 and 1, negative values in the newly obtained data row are rounded to 0, and values greater than 1 are rounded to 1. This approach ensures that the data values remain within the specified range, contributing to the preservation of data integrity. Subsequently, all values are rounded to two decimal places, which helps eliminate insignificant differences and ensure consistency in the dataset. Finally, this process is repeated for all missing data, and all generated synthetic data rows are converted into a dataframe. As a result of the process, the created synthetic dataset is saved in the record directory, replacing the original data with a higher size. This method preserves the structure and characteristics of the original dataset while increasing data diversity and volume.

Moreover, the sample counts were equalized by downsampling 434 samples from each excess result in situations where the number of facial expression analysis results exceeded the number of EEG signals. The downsampling procedure used for dataset equalization includes three primary stages: calculating the downsampling factor, performing the downsampling operation, and correctly combining the downsampled FER data with the EEG data. During the data reduction phase, downsampling was utilized to align the FER data with the EEG data on the time axis. The function performing the downsampling operations reduces the size of the dataset by selecting the element at the index corresponding to the downsampling factor from the FER data. This strategy successfully reduced the frequency of the FER dataset by aligning it with the frequency of the EEG dataset.

$$\text{downsample\_factor} = \left\lfloor \frac{\text{fer\_total\_length}}{\text{eeg\_length}} \right\rfloor \tag{2}$$

The formula first presented in Equation 2 was used to compute the downsampling factor. This formula is applied by a function that performs the these steps: A ratio is obtained by this function by dividing the length of all FER data (fer_total_length) by the length of all EEG data (eeg_length, default = 17360). The downsampling factor is next calculated by rounding this ratio down to the closest integer. The FER data is shortened to the same length as the EEG data by using this factor. When combining EEG and FER data, each FER file had its data read using a function and extra columns (such as "box") eliminated. The obtained FER data was next downsampled with the determined downsampling factor and positioned on the time axis in line with the EEG data. An integrated dataset was produced by combining each FER data sample with the matching EEG sample in this way. As such, the final facial expression analysis findings of the relevant videos were combined with the EEG signals acquired from each video.

The procedure outlined in the "EEG signals with emotion classification" section was followed for the remainder of the data processing pipeline. The final seven features in the dataset were not normalized, as the facial analysis findings were already between 0 and 1.

D. MODEL TRAINING

Three different deep learning approaches were employed for emotion classification on the data: GRU (Gated Recurrent Unit), LSTM (Long Short-Term Memory), and Transformer models.

1) GRU

GRU models have two gates: update and reset. GRUs utilize gating units to manage the flow of information, allowing relevant information to be retained for longer durations [29].

The update gate, z_t, enables the model to decide which information to discard and which new information to add to the cell state. This gate supports the model's decisions regarding how much of the past information should be retained. Equation 3 represents the update gate, where W_z is the weight matrix for the update gate, σ is the sigmoid activation function, h_{t-1} is the previous hidden state, and x_t represents the input at time step t.

$$z_t = \sigma(W_z \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{3}$$

The reset gate, r_t, is used by the GRU model to decide how much past information to forget. It allows the model to reset its memory when necessary. In Equation 4, W_r represents the weight matrix for the reset gate.

$$r_t = \sigma(W_r \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t]) \tag{4}$$

The GRU model used in the study consists of three layers (128, 64, 32 units). In the first two layers, the sequential data structure is preserved during the transition between GRU layers with the ''return_sequences = True'' parameter. This enables the model to learn time-related features in deeper layers. In the third and final GRU layer, the ''return_sequences'' parameter is set to ''False''. This setting aggregates the information from all time steps into a single vector in the last layer and passes it to the output layer.

Table 2 shows the model's characteristics and parameters. L1 and L2 regularization [30] were applied to this model to prevent overfitting and enhance the model's generalization ability. While L1 regularization aims to create a simpler model by forcing some values in the model's weight matrix towards zero, L2 regularization minimizes the sum of the squares of the weights, preventing weight values from increasing significantly. Dropout layers [31], added after each GRU layer, randomly deactivate selected neurons during training, allowing the model to utilize different neuron combinations and achieve a more robust structure. The dropout approach is a widely used technique to prevent overfitting in deep learning models. Additionally, BatchNormalization [32] layers were added to enable faster and more stable training of the model. These layers normalize the distribution of activations within each mini-batch, optimizing the scaling of activations between network layers. The output layer includes a Dense layer with a softmax activation function [33] to generate the predicted class probabilities.

2) LSTM

LSTMs have a control flow that can process data to provide forward information flow. LSTM cells contain gates that determine which information to remember and which to forget, and a cell state that carries data to other cells. There are three main gates in LSTMs: Forget Gate, Input Gate, and Output Gate. These gates regulate the flow of information entering and leaving the cell state, allowing LSTMs to remember important information for extended periods [34].

The forget gate, f_t, decides which information from the previous cell state should be discarded or retained. It generates a value between 0 and 1 for each number in the previous cell state. A value closer to 0 indicates forgetting, while a value closer to 1 indicates retaining. In Equation 6, the symbol W_f represents the weight matrix for the forget gate, and b_f represents the bias for the forget gate.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$

The input gate (Equation 7) decides which values from the input will be used to update the cell state. The candidate memory (Equation 8) allows for the addition of new potential values to the cell state.

$$f_t = \sigma(W_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_f) \tag{6}$$

The input gate (Equation 7) decides which values from the input will be used to update the cell state. The candidate memory (Equation 8) allows for the addition of new potential values to the cell state.

$$i_t = \sigma(W_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_i) \tag{7}$$

$$\tilde{C}_t = \tanh(W_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_C) \tag{8}$$

Similarly, three distinct LSTM layers were employed for the training process of the LSTM model. In the first two layers, the ''return_sequences'' parameter was set to True. This ensures that the layer outputs remain as sequential data for the subsequent layer, thus enabling the preservation of time-related information. In the third LSTM layer, however, the ''return_sequences'' parameter was set to ''False.'' This compresses the time dimension in the final layer, transmitting it as a single vector to the following Dense layer. As indicated in Table 2, L1 and L2 regularizations were applied to each LSTM layer, similar to the GRU model, to prevent overfitting. Dropout layers were added after each LSTM layer to enhance the model's robustness. Batch Normalization layers, which normalize activations between layers, were also included to accelerate and stabilize the training process. The output layer of the model is a Dense layer equipped with the softmax activation function, used to represent class probabilities. Finally, the model was compiled using categorical cross-entropy loss and Adam optimization.

3) TRANSFORMER

Transformers are a type of deep learning model that efficiently perform sequence-to-sequence tasks without relying on recurrent connections, and they have been particularly

TABLE 2. Training parameters of unimodal and multimodal models.

Feature	Parameters
Base Model	GRU, LSTM, Transformer
Additional Layers	Global Average Pooling, Dense Layer, 2 Dropout Layers, 2 Batch Normalization Layers
Epochs	330
Regularization	L1, L2
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Batch Size	64
Number of Classes	4 (HVHA, HVLA, LVHA, LVLA)

successful in natural language processing (NLP) tasks. Transformer utilize self-attention mechanisms to process the entire sequence simultaneously, allowing for better parallelization and more effective capture of long-term dependencies. Recently, in addition to natural language processing, they have also been used in various application areas such as time series forecasting and classification.

The Transformer model consists of transformer encoder blocks containing MultiHeadAttention, layer normalization, dropout, and convolutional layers. Each encoder block starts with an attention mechanism followed by a feed-forward network. MultiHead Attention layers (Equation 9), incorporating attention mechanisms, enable the model to assess relationships between features in the dataset from multiple perspectives [35].

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \quad (9)$$

The feed-forward network consists of two convolutional layers that follow each MultiHead Attention block. These layers are used to increase the model's learning capacity and enable it to extract more complex features. Layer normalization and dropout are implemented to prevent overfitting and ensure a more stable learning process in the model. Similar to LSTM and GRU, the Transformer model is compiled using categorical cross-entropy as the loss function and Adam optimization as the optimization technique.

The hyperparameters for the Transformer model, shown in Table 3, were determined through experimental studies. The table specifies two different values for the input_shape parameter due to the differing input dimensions of unimodal and multimodal data. The value specified as 'None' indicates that the model can handle data batches of varying sizes. The other hyperparameters, excluding input_shape, remain the same for both unimodal and multimodal approaches. The success of the modeling depends on the proper tuning of these hyperparameters, the model parameters presented in Table 2, the quality of the dataset, and the model's training process.

During the training of the models, ModelCheckpoint and EarlyStopping callbacks were utilized. ModelCheckpoint ensures that the best model is saved during training, allowing for later reuse. EarlyStopping [36] prevents unnecessary computation by stopping training early if there is no improvement in model performance for a certain period. Each model was trained for 330 epochs, and the entire training dataset was processed by the model in each epoch. The learning rate was determined to be 0.0001 for each model as a result of experimental studies. Additionally, the models were trained on previously created training and validation datasets, and the training process assesses the models' ability to learn patterns in the dataset and their performance on the validation set based on this learned knowledge. Finally, the same model architecture was used for both unimodal and multimodal data in each model, with only changes made to the input dimensions.

TABLE 3. Model hyperparameters and descriptions.

Hyperparameters	Values	Descriptions
Input Shape	256x1 / 263x1	Shape of the input data
Classes	4	Number of classes for prediction
Head Size, Heads	256, 4	Dimensions of the attention mechanism and the number of heads
Dim	4	Size of the feed-forward network
Transformer Blocks	4	Number of transformer encoder blocks
MLP Units	[128]	Number of units in the MLP network
Dropout, MLP Dropout	0.2, 0.2	Dropout rates applied to prevent overfitting

The overall architecture outlined in Figure 3 illustrates the unimodal and multimodal experiments aimed at emotion classification using EEG signals and facial expressions derived from the DEAP dataset. In this study, for the unimodal experiment, EEG data underwent various preprocessing steps such as noise reduction, normalization, scaling, and reshaping. Subsequently, this data was used as input to different deep learning models, including Transformer, LSTM, and GRU. The performance of the models was evaluated using metrics such as accuracy, precision, recall, and F1 score.

In the multimodal experiment, data obtained from facial videos were processed using an algorithm that analyzes facial expressions. These analysis results were downsampled using the EEG downsampling method and then combined with EEG signals. The resulting multimodal dataset was then fed as input to the models used for emotion classification. In both types of experiments, the emotion recognition performance of the models was thoroughly analyzed based on the obtained results.

IV. RESULTS

In this study, the performance of GRU, LSTM, and Transformer models was investigated using both EEG data alone

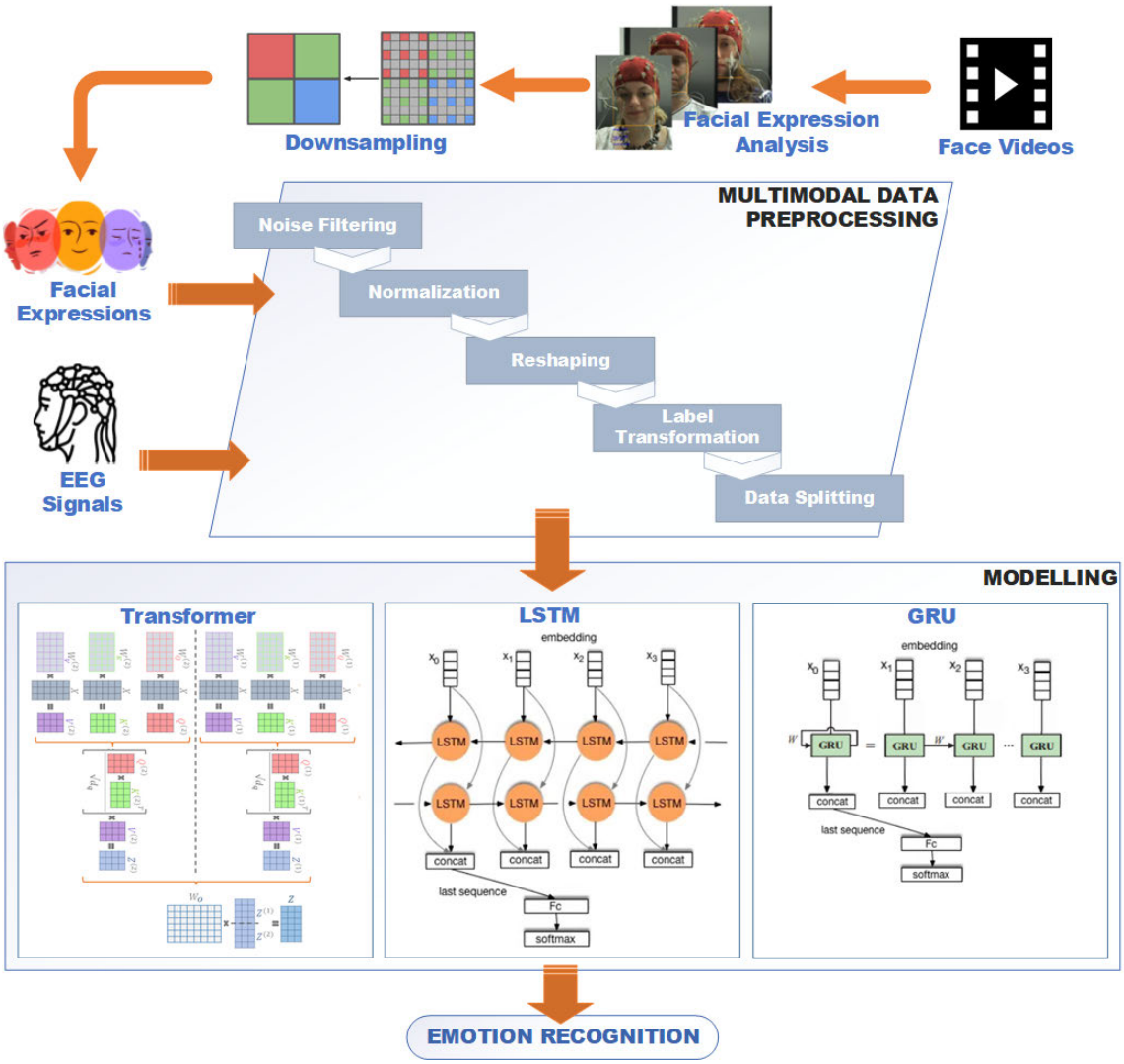


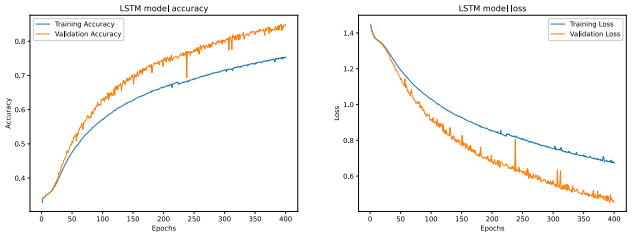
FIGURE 3. Overall architecture for both unimodal and multimodal approaches.

and a combination of EEG and facial expression data. The findings encompass the accuracy values, classification reports, confusion matrices, and loss and accuracy graphs for each model.

A. EMOTION CLASSIFICATION RESULTS WITH EEG SIGNALS

This section evaluates and discusses the performance of LSTM, Transformer, and GRU models trained exclusively on EEG signals.

The accuracy graph in Figure 4a illustrates the performance of the LSTM model on the combined EEG and facial expression training and validation datasets for each training epoch. Initially low, the accuracy increases as the number of training epochs progresses. The close alignment of training (blue line) and validation (orange line) accuracies indicates that the model is not overfitting. However, occasional spikes in the validation accuracy curve, which cause sudden drops, are observed. This suggests that the model exhibits inconsistencies on the validation set during certain training epochs. Overall, the model achieves a high accuracy rate, which increases gradually, indicating good generalization capacity.



(a) (b)

FIGURE 4. Training performance of the unimodal LSTM model: a) accuracy, b) loss.

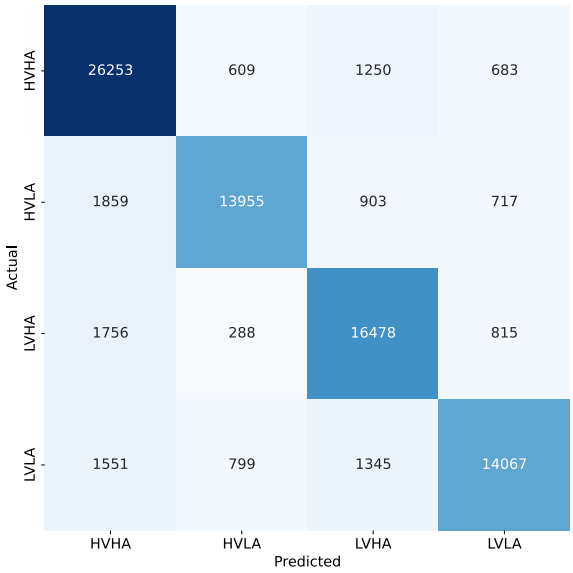


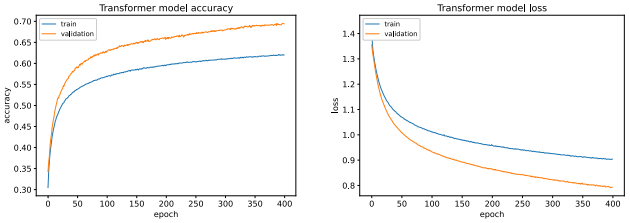
FIGURE 5. Confusion matrix of the unimodal LSTM model.

TABLE 4. Classification report for Unimodal LSTM model.

	Precision	Recall	F1-Score	Support
HVHA	0.84	0.91	0.87	28795
HVLA	0.89	0.80	0.84	17434
LVHA	0.82	0.85	0.84	19337
LVLA	0.86	0.79	0.83	17762
Accuracy			0.85	83328
Macro Avg	0.85	0.84	0.85	83328
Weighted Avg	0.85	0.85	0.85	83328

The loss graph in Figure 4b reveals how the LSTM model improves during training. The loss decreases over time for both the training and validation sets, indicating that the model is making better predictions. The validation loss (orange line) curve exhibits some sudden spikes. Such spikes usually indicate that certain data points are not well generalized by the model or that the model experiences instability during some stages of the training process. Overall, the model's loss decreases and reaches a high training accuracy, suggesting that the model demonstrates good learning performance.

Table 4 presents the classification report of the LSTM model trained solely on EEG data, evaluated on the test set. The precision metric [37] indicates the proportion of correct predictions made by the model for each class. These proportions are 84% for HVHA, 89% for HVLA, 82% for LVHA, and 86% for LVLA. The recall (sensitivity) metric [37] shows the proportion of actual positive cases that were correctly predicted. The recall values were calculated as 91% for HVHA, 80% for HVLA, 85% for LVHA, and 79% for LVLA. The F1-score [38] is the harmonic mean of precision and recall metrics, reflecting the model's balance between the two metrics. The F1-scores are 87% for HVHA, 84% for HVLA, 84% for LVHA, and 83% for LVLA, respectively. Support represents the number of samples for each class in the dataset and forms the basis for evaluation for each class.



(a) (b)

FIGURE 6. Training performance of the unimodal Transformer model: a) accuracy, b) loss.

Overall, the model demonstrated effective performance with 85% accuracy [39]. The macro_accuracy and weighted_accuracy averages are around 85%, indicating that the model performs consistently across all classes. While these results suggest the model's success, different algorithmic approaches (Transformer and GRU) were explored to address the confusion between some classes and the low recall rates.

Figure 6a displays the accuracy graph of the Transformer model trained solely on EEG signals. The model's training accuracy initially shows a sharp increase and then transitions to a slower increasing trend from the 50th epoch onwards. This indicates that the model quickly learns certain patterns in the training set and reaches a certain level of saturation. After the saturation point, the training accuracy continues to increase slightly, suggesting that the model continues to learn from the data. The validation accuracy demonstrates a similar increasing pattern to the training accuracy and flattens towards the end of training. There is a small difference between the training and validation accuracy rates. This difference is not significant, suggesting that the model does not exhibit substantial overfitting.

In the loss graph presented in Figure 6b, the training loss curve (blue line) demonstrates a rapid decrease in the initial training epochs. This observation aligns with the accuracy graph, indicating that the model learns quickly from the training data at the beginning. Subsequently, the loss curve continues to decline, but the rate of loss reduction slows down, settling into a stable trend. This suggests that the model's learning process is ongoing, albeit at a slower pace. The validation loss also exhibits a similar decreasing trend to the training loss, demonstrating a slightly better reduction than the training loss. Overall, both loss curves progress closely together, and no significant divergence is observed.

Figure 7 presents the confusion matrix for the Transformer model's performance on the EEG signals test set from the DEAP dataset. The confusion matrix reveals that the highest accuracy rate was achieved for the HVHA class, with 23880 correct predictions. However, the number of false negatives (instances that are actually HVHA but predicted as different) and false positives (other classes misclassified as HVHA) for this class are also considerable. A significant

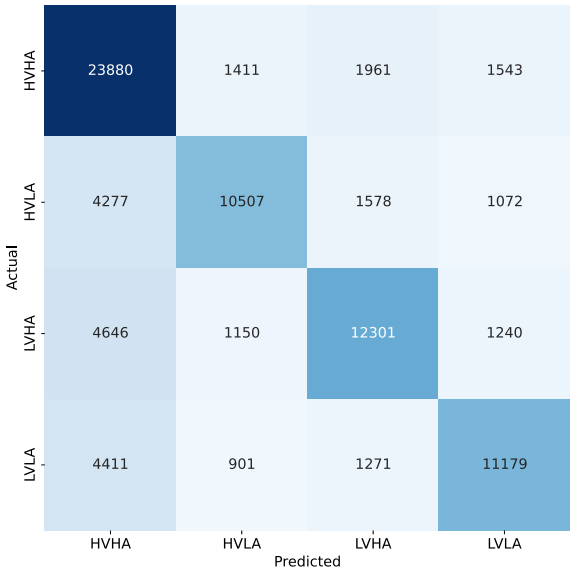


FIGURE 7. Confusion matrix of the unimodal Transformer model.

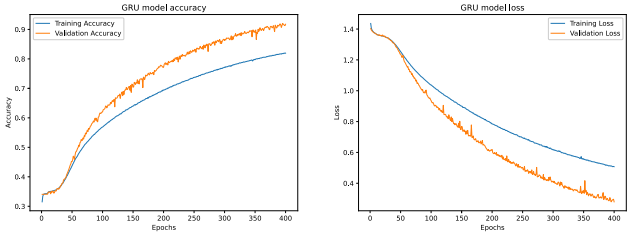
accuracy rate was obtained for the HVLA class with 10507 correct predictions. Nevertheless, the false positive and false negative counts for this class are higher compared to the results of the LSTM model in Figure 5. For the LVHA class, there were 12301 correct predictions, and for the LVLA class, there were 11179 correct predictions. However, false positives and false negatives are present for both classes.

TABLE 5. Classification report for Uni-modal Transformer model.

	Precision	Recall	F1-Score	Support
HVHA	0.64	0.83	0.72	28795
HVLA	0.75	0.60	0.67	17434
LVHA	0.72	0.64	0.67	19337
LVLA	0.74	0.63	0.68	17762
Accuracy			0.69	83328
Macro Avg	0.71	0.67	0.69	83328
Weighted Avg	0.70	0.69	0.69	83328

Table 5 presents the classification report of the Transformer model, revealing precision values of 64% for HVHA, 75% for HVLA, 72% for LVHA, and 74% for LVLA. Recall rates were calculated as 83% for HVLA, 60% for HVHA, 64% for LVHA, and 63% for LVLA. The F1-scores are 72% for HVHA, 67% for HVLA, 67% for LVHA, and 68% for LVLA. Overall, the model exhibited moderate performance with an accuracy of 69%. Macro and weighted accuracy values range between 69% and 71%, indicating that the model is consistent but not exceptionally high-performing.

In summary, the model's overall accuracy and F1-scores demonstrate that the Transformer architecture can handle the complexity of EEG data. However, the model's high false positive and false negative rates suggest that the modeling process needs further optimization.



(a) (b)

FIGURE 8. Training performance of the unimodal GRU model: a) accuracy, b) loss.

Upon examining the accuracy graph of the GRU model in Figure 8a, the training accuracy initially increases rapidly, particularly during the first 100 epochs. From the 100th epoch onwards, the rate of accuracy increase slows down, but the training accuracy continues to rise steadily, reaching around 90% at the 300th epoch. This indicates that the model continues to learn from the training data and achieves a good fit. Additionally, the validation accuracy follows a similar trend, initially showing a rapid increase and then a slower rate of increase from the 200th epoch onwards, while remaining consistently slightly above the training accuracy.

In the loss graph presented in Figure 8b for the GRU model, a rapid decrease in both training and validation loss is observed within the first 10 epochs. This rapid decrease suggests that the model quickly identifies and adapts to the patterns and relationships in the training set. Between the 10th and 40th epochs, a plateau in the losses is observed, indicating that the model reaches a certain learning saturation during these stages. Between the 35th and 80th epochs, there is another rapid decrease in losses, suggesting that the model discovers new patterns or fine-tunes existing learning. After the 80th epoch, the rate of decrease in training and validation losses slows down, but the loss continues to decline. This indicates that the model has now entered a more refined stage of the learning process, beginning to learn more complex or less obvious patterns. The close proximity and stable convergence between the training and validation loss curves reveal that the model successfully captures the general characteristics of the dataset without overfitting.

Figure 9 presents the confusion matrix, illustrating the classification results of the GRU model on the test dataset. According to this matrix, the highest accuracy was achieved for the HVHA class with 27536 correct predictions. For the HVLA class, 15240 correct predictions were made. While these results indicate a need for improvement, they still offer a high accuracy rate. The model effectively recognized both the LVHA and LVLA classes with 17340 and 16385 correct predictions, respectively.

The lower number of false positives and false negatives compared to the LSTM and Transformer models suggests that the GRU model effectively distinguishes between classes in the test dataset.

Table 6 presents the classification report of the GRU model, revealing precision values of 93% for HVHA,

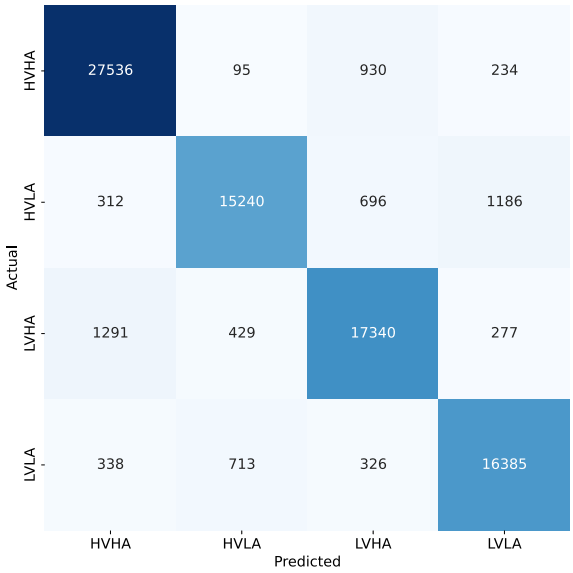


FIGURE 9. Confusion matrix of the uni-modal GRU model.

TABLE 6. Classification report for Unimodal GRU model.

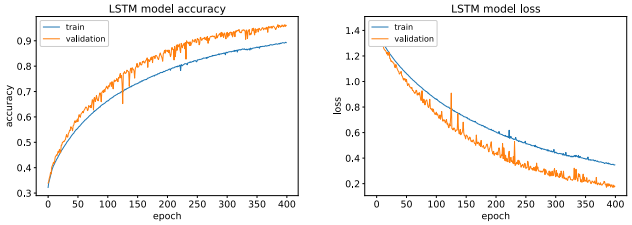
	Precision	Recall	F1-Score	Support
HVHA	0.93	0.96	0.95	28795
HVLA	0.92	0.87	0.90	17434
LVHA	0.90	0.90	0.90	19337
LVLA	0.91	0.92	0.91	17762
Accuracy			0.92	83328
Macro Avg	0.92	0.91	0.91	83328
Weighted Avg	0.92	0.92	0.92	83328

92% for HVLA, 90% for LVHA, and 91% for LVLA. These high precision values indicate that the majority of the model's positive predictions are accurate. Recall rates were determined as 96% for HVHA, 87% for HVLA, 90% for LVHA, and 92% for LVLA. The model's F1-scores are 95% for HVHA, 90% for HVLA, 90% for LVHA, and 91% for LVLA. These F1-scores suggest that the model performs well in both recall and precision metrics. The overall accuracy, macro accuracy and weighted accuracy rates of the model ranged between %91 and %92. These results demonstrate that the model performs with high accuracy across all classes.

The results indicate that the GRU model, in general, can successfully process EEG signals with complex data structures. The GRU model, trained only on EEG data, outperformed the LSTM and Transformer models, showcasing its potential for working with EEG data. However, to achieve better results and enhance success, the models were retrained and tested with various data types. The results of these experiments are explained in more detail in Section IV-B.

B. EMOTION CLASSIFICATION RESULTS WITH BOTH EEG SIGNALS AND FACIAL EXPRESSIONS

In this section, the performance of LSTM, Transformer, and GRU models trained on the combined EEG and facial expression data is evaluated and the results are discussed.



(a) (b)

FIGURE 10. Training performance of the multimodal LSTM model: a) accuracy, b) loss.

The graph in Figure 10a illustrates the performance of the LSTM model on the combined EEG and facial expression training and validation datasets for each training epoch. Initially low, the accuracy increases as the number of training epochs progresses. The close alignment of training and validation accuracies indicates that the model is not overfitting. However, there are occasional spikes where the validation accuracy drops below the training accuracy. This suggests that the model exhibits inconsistencies on the validation set during certain training epochs. Overall, the model achieves a high accuracy rate, which increases gradually until the end of training.

According to the loss graph in Figure 10b, the loss for both training and validation sets decreases over time, indicating that the model is making better predictions. The validation loss curve exhibits some sudden spikes, suggesting that the model experiences instability during some stages of the training process. Overall, the model's loss decreases and reaches a high training accuracy, demonstrating that the LSTM model achieves good learning performance.

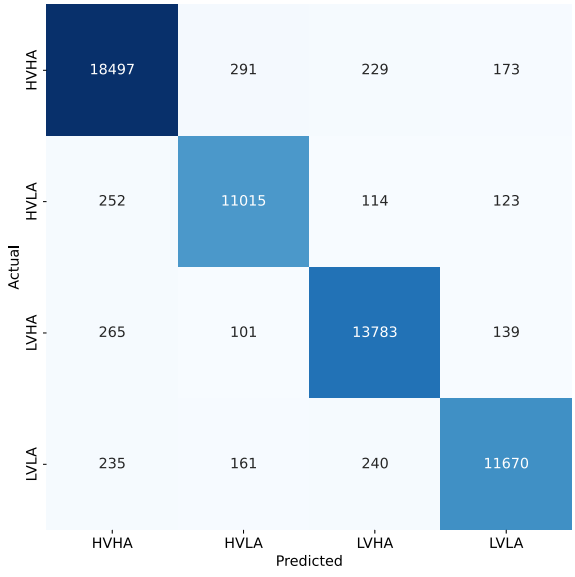


FIGURE 11. Confusion matrix of the multimodal LSTM model.

In Figure 11, the confusion matrix of the LSTM model's predictions on the combined EEG and facial expression

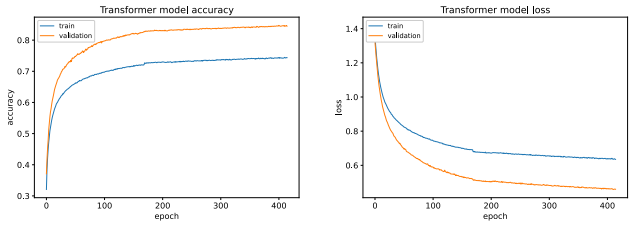
test data reveals high accuracy for the HVHA class with 18497 predictions. The model also achieved a high accuracy rate for the HVLA class with 11015 correct predictions. In the LVHA and LVLA classes, the model demonstrated similar performance, making 13783 and 11670 correct predictions, respectively. The confusion matrix shows a relatively low number of false positives and false negatives, indicating that the model predicts each class in a balanced manner.

TABLE 7. Classification report for Multimodal LSTM model.

	Precision	Recall	F1-Score	Support
HVHA	0.96	0.96	0.96	19190
HVLA	0.95	0.96	0.95	11504
LVHA	0.96	0.96	0.96	14288
LVLA	0.96	0.95	0.96	12306
Accuracy			0.96	57288
Macro Avg	0.96	0.96	0.96	57288
Weighted Avg	0.96	0.96	0.96	57288

Table 7 presents the classification report for the LSTM model trained with combined EEG and facial expression data. The precision metric ranges from 95% to 96% for all classes, indicating a low false positive rate for the model. Recall rates vary between 95% and 96%, demonstrating that the model successfully predicts a large number of true positive values. The F1-score for all classes ranges from 95% to 96%, highlighting the model's high performance.

The model achieved successful performance with an overall accuracy of 96%. The macro and weighted accuracies are also consistent and high, with values of 96%. These results indicate that there is no significant performance difference between individual classes in the model. When compared with the classification report of the unimodal LSTM approach presented in Table 4 in Section IV-A, it is evident that the LSTM model working with multimodal data shows increased success rates in accuracy, precision, recall, and F1-score metrics. The high precision, recall, and F1-scores reveal that the model produces consistent results in classifying emotional states, demonstrating the positive impact of multimodal data on model performance in emotion classification.



(a) (b)

FIGURE 12. Training performance of the multimodal Transformer model: a) accuracy, b) loss.

Figure 12a presents the accuracy graph of the Transformer model, demonstrating a rapid increase in accuracy as the model begins training. After the initial training epochs, the model achieves an accuracy exceeding 50% on both the training and validation sets. The training accuracy continues to increase throughout the training process, reaching approximately 74% accuracy by the end. This indicates that the model continues to learn from the training data. The validation accuracy also shows a similar increasing trend to the training accuracy, reaching approximately 84%. The consistent gap between training and validation accuracies throughout the training process suggests that the model might be slightly overfitting. However, both accuracy curves progress closely together without significant divergence, indicating that the overfitting is either absent or negligible.

According to the Transformer model's loss graph in Figure 12b, the training loss initially shows a rapid decrease, and in subsequent epochs, the loss continues to decrease even though the rate of decrease slows down. The continuous decrease in loss indicates that the model recognizes patterns and starts to predict them better. The diminishing rate of loss reduction over time suggests that the model finds less new information to learn from the training data after the initial rapid improvement. The validation loss also continues to decrease in parallel with the training loss, ending the training process slightly below the training loss.

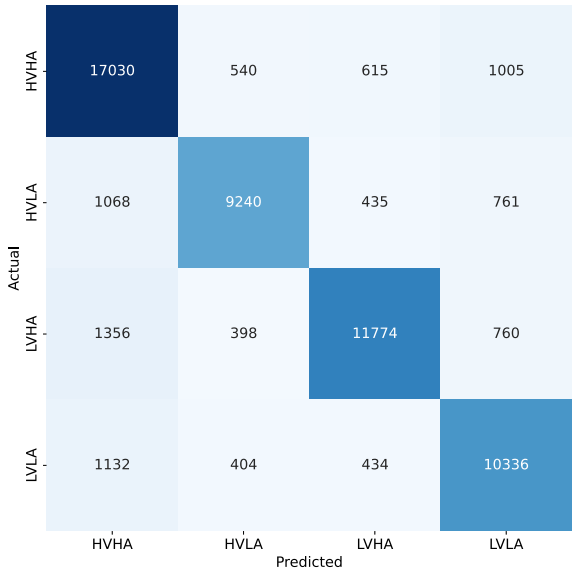


FIGURE 13. Confusion matrix of the multimodal Transformer model.

Figure 13 presents the confusion matrix for the Transformer model, revealing that the highest number of correct classifications was achieved for the HVHA class with 17030 predictions. The highest false negative count for this class was recorded for the LVLA class label. For the HVLA class, a lower accuracy was achieved compared to the HVHA class, with 9,240 correct classifications. This lower accuracy can be attributed, in part, to the fewer samples belonging to this class in the test data. A notable accuracy was achieved for the LVHA label with 11774 correct predictions. However,

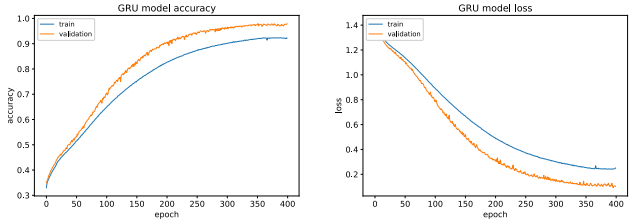
the false negative count for this class is higher than that of the HVLA class. Finally, the model achieved a high classification accuracy for the LVLA class with 10336 correct classifications. Additionally, it was observed that the model struggles to distinguish between some classes, particularly with a noticeable confusion between the HVHA and LVLA classes.

TABLE 8. Classification report for multimodal Transformer model.

	Precision	Recall	F1-Score	Support
HVHA	0.83	0.89	0.86	19190
HVLA	0.87	0.80	0.84	11504
LVHA	0.89	0.82	0.85	14288
LVLA	0.80	0.84	0.82	12306
Accuracy			0.84	57288
Macro Avg	0.85	0.84	0.84	57288
Weighted Avg	0.85	0.84	0.84	57288

Table 8 presents the classification report of the Transformer model. Examining the precision column, it is observed that the model predicts the HVHA class with 83% precision, the HVLA class with 87% precision, the LVHA class with 89% precision, and the LVLA class with 80% precision. Recall results are 89% for HVHA, 80% for HVLA, 82% for LVHA, and 84% for LVLA. The model's F1-scores are 86% for HVHA, 84% for HVLA, 85% for LVHA, and 82% for LVLA.

The model achieved a good performance with an overall accuracy of 84%. The macro and weighted averages are measured as 84% and 85%, respectively. While the results indicate minor performance differences between specific classes, the inclusion of multimodal data has led to a higher success rate in emotion classification compared to the unimodal Transformer model (Section IV-A, Table 5).



(a) (b)

FIGURE 14. Training performance of the multimodal GRU model: a) accuracy, b) loss.

The accuracy graph of the GRU model trained on the combined EEG and facial expression dataset, shown in Figure 14a, indicates a steady increase in training accuracy from the beginning. The training accuracy curve reaches almost 92% at the 400th epoch, demonstrating the model's high learning capacity on the training data. The validation accuracy initially shows a similar increase to the training accuracy and, from a certain point onwards, increases consistently with the training accuracy. The increase in validation accuracy continues until the end of the training epochs, reaching high levels. This suggests that the model has good generalization performance not only on the training data but also on unseen data. The balanced gap between training and validation accuracy and the stabilization of validation accuracy at a high level indicate that the model does not overfit and captures the general characteristics of the dataset.

The loss graph of the GRU model in Figure 14b shows how the model's loss values change for both training and validation sets. The training loss decreases from the beginning and continues to decrease throughout the training epochs. Towards the end of the graph, the rate of loss reduction slows down, and a plateau formation is observed. The validation loss also decreases from the beginning throughout the training epochs, following a similar trend to the training loss and forming a plateau towards the end of training. The decrease in loss during training indicates that the model's learning capacity on the training data increases.

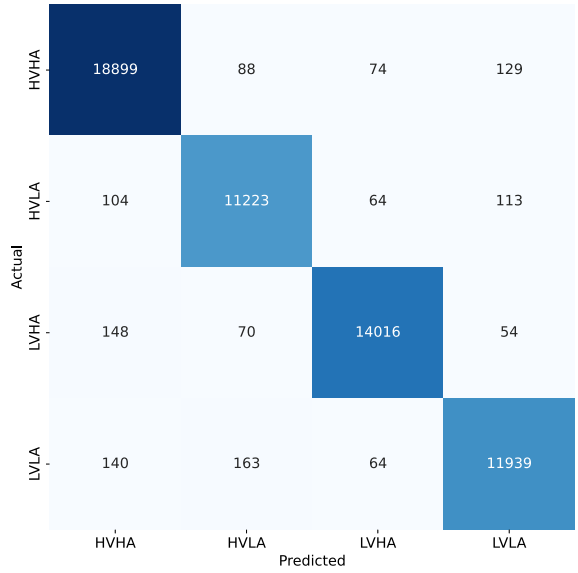


FIGURE 15. Confusion matrix of the multimodal GRU model.

The confusion matrix in Figure 15 presents the classification results of the GRU model on the test dataset for the HVHA, HVLA, LVHA, and LVLA classes. The model achieved high accuracy for the HVHA class with 18899 correct predictions, producing very few false negatives for this class. For the HVLA class, a high accuracy rate was observed with 11223 correct predictions. Similarly, the model demonstrated strong accuracy for the LVHA class with 14016 correct predictions. In the LVLA class, a noteworthy accuracy rate was obtained with 11939 correct predictions. There are very few misclassifications between class labels. This low misclassification rate demonstrates the GRU model's high effectiveness in predicting different emotional states.

Table 9 presents the classification report for the GRU model. Upon examination of this report, the precision rates,

TABLE 9. Classification report for multimodal GRU model.

	Precision	Recall	F1-Score	Support
HVHA	0.98	0.98	0.98	19190
HVLA	0.97	0.98	0.97	11504
LVHA	0.99	0.98	0.98	14288
LVLA	0.98	0.97	0.97	12306
Accuracy			0.98	57288
Macro Avg	0.98	0.98	0.98	57288
Weighted Avg	0.98	0.98	0.98	57288

ranging from 97% to 99% for each class, indicate that most of the model's positive predictions are indeed positive. Recall values also range between 97% and 98%, signifying that a large proportion of true positive cases are correctly predicted. The F1-scores, varying between 97% and 98%, demonstrate that the model exhibits a balanced performance in both precision and recall metrics. The model's overall accuracy was measured as 98%, which is higher than the overall accuracy rates of the multimodal LSTM and Transformer models. The macro and weighted accuracy values are also 98%, indicating that the model performs consistently across classes. These results suggest that the model can reliably and effectively classify emotions in practical applications.

Overall, in classifications using EEG data combined with facial expressions, each model achieved better results compared to using only EEG data. The GRU model classified emotions with the highest accuracy in both unimodal and multimodal approaches, surpassing both Transformer and LSTM models in this context. Its superior performance compared to other models makes the GRU model more suitable for predicting emotional states in this dataset.

V. DISCUSSION

Three different models, namely LSTM, Transformer, and GRU, were tested for both unimodal and multimodal approaches, with the GRU model demonstrating the highest performance in emotion classification. The analysis of confusion matrices, classification reports, accuracy, and loss graphs has facilitated the identification of areas for improvement in the model. Notably, the utilization of multimodal datasets has been identified as a crucial factor in enhancing the accuracy of the models.

TABLE 10. Overall accuracy rates of models.

Model - Overall Accuracy Rates (%)	Unimodal (EEG only)	Multimodal (EEG and face)
GRU	91.8	97.8
LSTM	84.9	95.9
Transformer	70.1	84.4

Table 10 presents the accuracy rates achieved by the models on both unimodal and multimodal data. Unimodal approaches involve classifying emotions using either EEG data alone. For unimodal approaches utilizing EEG data, the GRU, LSTM, and Transformer models demonstrated good performance in emotion classification with accuracy rates of 91.8%, 84.9%, and 70.1%, respectively. The GRU model achieved the highest accuracy rates in both unimodal (EEG only and facial expressions only) and multimodal datasets with 91.8%, and 97.8%, respectively. In contrast, the Transformer model exhibited the lowest accuracy rates in both approaches compared to other models, with 70.1%, and 84.9%. The GRU model's simple structure and fewer parameters allowed it to adapt well to the dimensionality and temporal nature of EEG data. On the other hand, the Transformer model's lower performance can be attributed to its complex structure, which may be more suitable for larger and more comprehensive datasets.

TABLE 11. Overall loss performance of models.

Model - Overall Loss Statistics	Unimodal (EEG only)	Multimodal (EEG and face)
GRU	0.278	0.104
LSTM	0.45	0.179
Transformer	0.791	0.465

Table 11 presents the loss scores achieved by the GRU, LSTM, and Transformer models on their test sets using both unimodal and multimodal data. The GRU model attained the lowest loss scores in both approaches, with values of 0.278, and 0.104. The LSTM model demonstrated a good performance, falling between the other two models, with loss scores of 0.45 for unimodal EEG signals, and 0.179 for the multimodal approach. Conversely, the Transformer model exhibited a poorer loss performance compared to the other two models, with loss scores of 0.791 for unimodal data and 0.465 for multimodal data. In conclusion, the GRU model's low loss scores provide a strong argument for its utilization in both unimodal and multimodal datasets.

The studies summarized in Table 12 showcase various approaches to emotion recognition using different datasets and modalities. The proposed study demonstrated a marginally higher performance, achieving an accuracy of 97.80%, compared to Wang et al.'s [23] 96.89% using the DEAP dataset by combining EEG and facial expressions. Chaparro et al. [19] and Muhammad et al. [16] achieved accuracies of 97.00% and 93.86%, respectively, using the MAHNOB-HCI dataset combined with EEG and facial expressions. Studies by Ma et al. [14] and Huang et al. [15] used EEG with other physiological signals and custom datasets, achieving accuracies of 92.87% and 82.75%, respectively. Additionally, Liu et al. [13] and Zheng et al. [12] incorporated eye movement data, obtaining accuracies of 91.01% and 73.59%, respectively. Different datasets and modalities lead to varying results. In the comparisons made, some studies used the DEAP dataset while others used different datasets. According to the table, facial expressions together with EEG usually achieve a higher performance in emotional recognition. This emphasizes the possibility

TABLE 12. Comparison of emotion recognition performance.

Study	Dataset	Approach	Class	Accuracy (%)
Proposed Study	DEAP	EEG + Facial Expressions	HVHA, HVLA, LVHA, LVLA	97.80
Chaparro et al. [19]	MAHNOB-HCI	EEG + Facial Expressions	Sad, Happy, Neutral etc.	97.00
Wang et al. [23]	DEAP, MAHNOB-HCI	EEG + Facial Expressions	Valence, Arousal	96.89, 96.4
Muhammad et al. [16]	DEAP, MAHNOB-HCI	EEG + Facial Expressions	Happy, Neutral, Sad	91.54, 93.86
Ma et al. [14]	DEAP	EEG + Other Physiological Signals	Valence, Arousal	92.87
Huang et al. [15]	Custom	EEG + Facial Expressions	Happy, Neutral, Sad, Fear	82.75
Liu et al. [13]	DEAP	EEG + Eye Movement	Valence, Arousal, Dominance, Liking	83.25
Tan et al. [21]	Custom	EEG + Facial Expressions	Happy, Neutral, Sad, Fear	83.33
Zheng et al. [12]	Custom	EEG + Eye Movement	Positive, Neutral, Negative	73.59
Cimtay et al. [22]	DEAP	EEG + Facial Expressions + GSR	Happy, Neutral, Sad etc.	53.8

of enhancing the classification performance by means of multimodal strategies.

VI. CONCLUSION

This study significantly advances emotion recognition by demonstrating the superior performance of GRU deep learning models in classifying emotional states through the multimodal integration of EEG signals and facial expressions. Our methodology, detailing the processing and fusion of the DEAP dataset, contributes to the creation of a rich and diverse dataset for training deep learning models. The GRU model's ability to effectively learn from time-series data and generalize complex emotional states underpins its exceptional accuracy. This research expands the understanding of emotion recognition and human-computer interaction by emphasizing the crucial role of multimodal data integration in enhancing the precision of deep learning models in classifying emotional expressions. The findings have broad practical implications for interactive systems, educational technologies, healthcare services, and security systems. Future research should prioritize larger and more diverse datasets to encompass a wider spectrum of emotional states and explore novel deep learning architectures and modality integrations to further advance the field. Furthermore, feature extraction techniques from EEG signals can be utilized to distinguish between contrasting emotional states (e.g., fear and anger) within the same category, and a more detailed emotion representation or a new approach can be applied alongside the valence-arousal model. This study provides a solid foundation for future development, offering valuable insights for the continued improvement and application of emotion recognition technologies in human-machine interaction.

REFERENCES

[1] E. H. Houssein, A. Hammad, and A. A. Ali, "Human emotion recognition from EEG-based brain–computer interface using machine learning: A comprehensive review," *Neural Comput. Appl.*, vol. 34, no. 15, pp. 12527–12557, Aug. 2022.
[2] L. Pessoa, "Neural dynamics of emotion and cognition: From trajectories to underlying neural geometry," *Neural Netw.*, vol. 120, pp. 158–166, Dec. 2019.
[3] G. Dogan and F. P. Akbulut, "Multi-modal fusion learning through biosignal, audio, and visual content for detection of mental stress," *Neural Comput. Appl.*, vol. 35, no. 34, pp. 24435–24454, Dec. 2023.
[4] F. P. Akbulut, "Evaluating the effects of the autonomic nervous system and sympathetic activity on emotional states," *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, vol. 21, no. 41, pp. 156–169, Jun. 2022.
[5] W. Huang, W. Wu, M. V. Lucas, H. Huang, Z. Wen, and Y. Li, "Neurofeedback training with an electroencephalogram-based brain–computer interface enhances emotion regulation," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 998–1011, Apr. 2021.
[6] S. Derdiyok and F. P. Akbulut, "Biosignal based emotion-oriented video summarization," *Multimedia Syst.*, vol. 29, no. 3, pp. 1513–1526, Jun. 2023.
[7] S. C. Leong, Y. M. Tang, C. H. Lai, and C. K. M. Lee, "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing," *Comput. Sci. Rev.*, vol. 48, May 2023, Art. no. 100545.
[8] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.
[9] 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.
[10] S. Aydın and L. Onbaşı, "Graph theoretical brain connectivity measures to investigate neural correlates of music rhythms associated with fear and anger," *Cognit. Neurodyn.*, vol. 18, no. 1, pp. 49–66, Feb. 2024.
[11] B. Kılıç and S. Aydın, "Classification of contrasting discrete emotional states indicated by EEG based graph theoretical network measures," *Neuroinformatics*, vol. 20, no. 4, pp. 863–877, Oct. 2022.
[12] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 5040–5043.
[13] W. Liu, W. Zheng, and B. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Neural Inf. Process.*, Kyoto, Japan. Cham, Switzerland: Springer, Jan. 2016, pp. 521–529.
[14] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 176–183.
[15] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Comput. Intell. Neurosci.*, vol. 2017, no. 1, 2017, Art. no. 2107451.
[16] F. Muhammad, M. Hussain, and H. Aboalsamh, "A bimodal emotion recognition approach through the fusion of electroencephalography and facial sequences," *Diagnostics*, vol. 13, no. 5, p. 977, Mar. 2023.
[17] D. Wu, J. Zhang, and Q. Zhao, "Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning," *IEEE Access*, vol. 8, pp. 133180–133189, 2020.
[18] S. Rayatdoost, D. Rudrauf, and M. Soleymani, "Multimodal gated information fusion for emotion recognition from EEG signals and facial behaviors," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 655–659.

[19] V. Chaparro, A. Gomez, A. Salgado, O. L. Quintero, N. Lopez, and L. F. Villa, "Emotion recognition from EEG and facial expressions: A multimodal approach," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 530–533.
[20] T. Keshari and S. Palaniswamy, "Emotion recognition using feature-level fusion of facial expressions and body gestures," in *Proc. Int. Conf. Commun. Electron. Syst. (ICCES)*, Jul. 2019, pp. 1184–1189.
[21] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103029.
[22] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.
[23] S. Wang, J. Qu, Y. Zhang, and Y. Zhang, "Multimodal emotion recognition from EEG signals and facial expressions," *IEEE Access*, vol. 11, pp. 33061–33068, 2023.
[24] J. Chen, Y. Liu, W. Xue, K. Hu, and W. Lin, "Multimodal EEG emotion recognition based on the attention recurrent graph convolutional network," *Information*, vol. 13, no. 11, p. 550, Nov. 2022.
[25] G. K. Chakravarthy, M. Suchithra, and S. Thatavarti, "EEG-based multimodal emotion recognition with optimal trained hybrid classifier," *Multimedia Tools Appl.*, vol. 83, no. 17, pp. 50133–50156, Nov. 2023.
[26] J. D. A. Luz Junior and M. A. F. Rodrigues, "Comparative analysis of facial expression recognition systems for evaluating emotional states in virtual humans," in *Proc. Symp. Virtual Augmented Reality*, Nov. 2023, pp. 38–47.
[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[28] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2001, pp. 1–11.
[29] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2017, pp. 1597–1600.
[30] A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 78.
[31] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2015, pp. 448–456.
[33] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Sci.*, vol. 4, no. 12, pp. 310–316, May 2017.
[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 6000–6010.
[36] L. Prechelt, "Early stopping—But when?" in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2002, pp. 55–69.
[37] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Amer. Soc. for Inf. Sci.*, vol. 45, no. 1, pp. 12–19, Jan. 1994.
[38] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[39] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informat.*, vol. 17, no. 1, pp. 168–192, Jan. 2021.

SONGÜL ERDEM GÜLER received the B.S. degree in mathematics and the B.S. degree in digital forensic engineering from Fırat University, Elâzğ, Türkiye, in 2014 and 2020, respectively, and the M.Sc. degree in computer engineering from Istanbul Kültür University, İstanbul, Türkiye, in 2024. In the early years of her career, she was involved in a company specializing in anomaly detection, developing critical problem-solving skills in the complex field of data analytics. She then transitioned to her current role as a Senior Data Scientist, working in the field of artificial intelligence in healthcare. Throughout her tenure, she has developed numerous projects in healthcare AI, leveraging advanced data analysis, and machine learning techniques. She has been actively involved in the AI sector for approximately five years, contributing to the advancement of AI methodologies and their practical applications in the healthcare industry.

FATMA PATLAR AKBULUT (Member, IEEE) received the B.S. and M.S. degrees in computer engineering from Istanbul Kültür University (IKU) and the Ph.D. degree in biomedical engineering from İstanbul University, in 2017. Her doctoral work focused on developing machine-learning-enabled wearable systems for long-term cardiovascular disease monitoring. She joined the Computer Engineering Department, IKU, in 2019, where she is currently the Department Chair of the Software Engineering Department, further contributing her expertise in the field. Prior to IKU, she worked as a Postdoctoral Researcher with the Computer Science Department and Advanced Self-Powered Systems of Integrated Sensors and Technologies (ASSIST) Center, North Carolina State University (NCSU), from 2017 to 2019. Her research interests include biomedical signal processing, affective computing, data analytics, and wearable systems for healthcare.