Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077569

ARTICLE

NeuroVision: Multimodal Emotion Recognition via Dynamic Frame Enhancement and EEG-Guided Fusion

Ramakrishna Gandi[1,*], **Geetha A.**[1] **and Ramasubbareddy B.**[2]

[1]Computer Science & Engineering Department, Annamalai University, Annamalainagar, India
[2]Department of CSE, Mohan Babu University, Tirupati, India

*Corresponding Author: Ramakrishna Gandi. Email: gandiramakrishna2@gmail.com

Received: 11 December 2025; Accepted: 03 May 2026; Published: 15 June 2026

ABSTRACT: In the fields of affective computing, human-computer interaction, and psychological evaluation, the capacity to recognize emotions is crucial. Unimodal systems in the form of visual systems or of the physiological type are usually not designed to capture the complexity that exists in emotional states. The paper proposes NeuroVision: Multimodal Emotion Recognition System, combining facial video frames information and electroencephalogram (EEG) based information to enhance the accuracy and stability of the system. The system applies ResNet50 on the spatial information of facial expressions, Vision Transformer (ViT) on the temporal movements in the video, and an EEG-MLP Encoder to read the signal, without preprocessing, that captures pure neural patterns. The fully connected layers receive the fused characteristics and label them into three emotional conditions, using a softmax operational condition, which are: Happy, Sad, and Neutral. This model was evaluated on the LUMED-2 set, and the obtained accuracy of classification (96.9%) is higher than that observed by the existing unimodal and multimodal systems. Performing a lot of evaluations, like learning curves, confusion matrices, and benchmarking, proves that NeuroVision is effective and shows generalization capability, and can be used in real-time with an adaptive system, like mental health tracking and responsive interfaces.

KEYWORDS: Multimodal emotion recognition; EEG; facial expression analysis; vision transformer; ReseNet50; MLP encoder; human-computer interaction; affective computing; deep learning; neurophysiological signals

1 Introduction

Recognizing emotions is essential in fields like human-computer interaction, affective computing, and psychological assessments that play a vital role in improving user experiences in virtual environments and creating more responsive systems for mental health care [1]. Emotion identification is generally addressed by either visual modality, based on facial expressions [2], or by Physiological signals like EEG [3], to identify emotions. Similarly, using either of these methods as a sole source of analyzing human feelings is not always informative enough. In the study, we introduced the multimodal emotion recognition model NeuroVision [4], in which we merge face video frames and EEG inputs to make the recognition of emotions more exact and trustworthy. Using a mix of these two independent data sets, NeuroVision may make use of both exterior emotional information based on facial expression, as well as internal cues obtained from brain activity, giving a more unique perspective of the nature of a subject's emotional condition.

The recommended model employs a Convolutional Neural Net (CNN) [5], ResNet50, which is a deep learning (DL) model that extracts spatial information of video frames [6,7] and provides important details

regarding face expressions [8]. Visual features can then be recorded using the ViT [9] in the video sequence, so that the model can interpret the development of the emotion over time. Meanwhile, brainwave patterns linked to various emotional states are decoded using a lightweight MLP Encoder that ranges over EEG signals [10]. Features gathered from film and the EEG data are then integrated into a composite vector and placed through a classification layer that predicts the emotion.

This study examines the advantages of combining facial video data with EEG signals for automated emotion recognition. We assess the proposed model utilizing a compilation of synchronized EEG data and movies of facial expressions from many individuals. This shows that using more than one method works better than using only one. A methodical comparison with single-modality models is performed to elucidate the impact of each data stream on overall classification performance. The experimental results show that combining spatial and temporal patterns from video with neural activity data from EEG makes both accuracy and robustness much better. NeuroVision is a new architecture that combines a ResNet50 backbone, a Vision Transformer (ViT), and an EEG-MLP Encoder into one late-fusion pipeline. This design gives you a strong and complete way to arrange feelings into three groups: joyful, sad, and neutral. In general, it is superior to unimodal systems. The findings demonstrate a significant progression in emotional computing and human-computer interaction, yielding concrete consequences for the evolution of more responsive and contextually aware intelligent systems.

Furthermore, this research contributes to the growing field of multimodal emotion recognition, which provides a comprehensive framework for understanding human emotional expression. Accurately predicting people's emotional states is crucial for creating responsive and emotionally aware systems since it preconditions the usage of more complicated applications in areas like affective computing, mental state detection, and individual human-computer interactions.

The main goals of this investigation are:

NeuroVision's Cutting-Edge Technology: It is suggested that a hybrid multimodal framework be employed to provoke substantial emotional responses in individuals by merging synchronized EEG data with video frames of facial expressions. This method uses both touch and sight cues, which unimodal systems don't have.

Multimodal Fusion: This method uses both spatial facial information from video footage and temporal EEG signals to make categorization more accurate and reliable. It does this with the help of an MLP Encoder and a ResNet50 backbone. There are so many methods to show how you feel that it's hard to understand this idea.

Improved Emotion Classification: The suggested method shows better separation between three emotions (happy, sad, and neutral) than older methods that only looked at one emotion at a time. A thorough comparison examination confirmed this conclusion.

Hybrid Deep Learning: This novel approach employs a Vision Transformer (ViT) and ResNet50 to look at movie photos and see how they fit into the correct scenes. It also exhibits EEG signals with an encoder that has more than one layer. It has a unique shape and may be trained from start to finish.

Real-World Evaluation: Evaluates the model using a real-world dataset, highlighting its potential for applications in human-computer interaction and mental health monitoring.

This document is organized like this: Section 2 examines the most pertinent research on emotion recognition, emphasizing prior work that employed EEG-based models, video-based models, and multimodal fusion techniques. Section 3 has all the information you need on the proposed NeuroVision framework. This comprises the architecture of the model, the dataset, the preprocessing pipeline, and the method for combining the two. This means using ResNet50 to locate spatial characteristics and Vision Transformer (ViT)

to find contextual features. We talk about how the experiment was set up in Section 4. This explains how we set up the hardware, taught the staff, and checked the results. The paper talks about how fantastic the multimodal method is in Section 5 and presents ideas for how to undertake more research on how to find out how individuals really feel.

2 Literature Review

Affective computing, human-computer interaction, and cognitive state monitoring all now rely heavily on emotion recognition. Nevertheless, unimodal systems based on visual or physiological stimuli can prove insufficient in terms of modulating the richness and uncertainty of human feelings. In order to increase the reliability and precision of emotion recognition, there has been an increase in interest in multimodal systems that incorporate supplementary modalities, such as EEG, facial expressions, voice, eye movements, and galvanic skin response (GSR). Numerous studies explored various deep-learning networks and methods of hybridization to be able to integrate these modalities in an appropriate way. The state-of-the-art in multimodal emotion identification is covered in detail in this section, with a focus on hybrid fusion techniques, transformer-based networks, and cross-modal learning structures that try to improve a system's robustness and classification accuracy.

Choi et al. [11] used EEG and video to create a multimodal attention network with bilinear pooling to address the challenge of emotion recognition in the continuous-time domain. Their method was demonstrated to offer an enormous boost to unimodal practices on a couple of databases, the MAHNOB-HCI and in-house data. Fu et al. [12] have stressed the complementary character of the EEG and eye movement measurement in emotion recognition and mentioned the difficulties of subjective noise in the data collected with eye tracking. They proposed a cross-modal guiding neural network that uses EEG to better extract the eye movement features, thereby improving the performance of classification within SEED-IV, and a custom data set was designed. Rayatdoost et al. [13] considered a deep multimodal representation-based framework, which, cosine-similarly, fuses the EEG and facial expression characteristics as well as adds a facade to hybridize the bundle by being locked. Their strategy of excellent performance and domination of generalization on the existing techniques of fusion was observed on DAI-EF and MAHNOB-HCI datasets. Safavi et al. [14] introduced a transformer-based multimodal DL model to fuse neurophysiological (EEG) and facial expression representative parameters in order to improve emotion recognition. They demonstrate that their model was both accurate (displaying competitive accuracy on the Lie Detection dataset and performing near state-of-the-art accuracy on the DEAP dataset) and, when using only modest numbers of EEG electrodes, the model can be used in wearable systems. Pan et al. [15] focused on some of the most crucial problems associated with multimodal emotion recognition and presented Deep-Emotion, a framework that uses specialized neural networks working with each modality (facial, speech, and EEG). They have also shown much better performance in their level of decision fusion across CK+, EMO-DB, and MAHNOB-HCI datasets, which guarantees high stability in emotion detection and real-time emotion detection. Muhammad et al. [16] suggested that such a multimodal emotion recognition system, based on DCCA, the EEG, and the facial video-based control features, can be combined to make emotion recognition even more accurate. In 1D-CNN and ResNet50 feature extraction, their two-step process had high accuracy in MAHNOB-HCI and DEAP databases and compared well with some of the existing methods. Fang et al. [17] proposed the multi-level fusion, which uses visual signals and physiological signals to form a multi-level fusion to recognize emotion while still having serial fusion and EEG feature set into parallel fusion to improve the representation of the facial expression and the EEG feature set, respectively. A mix of feature-level fusion and decision-level fusion approaches is employed, demonstrating superior performance in recognizing intricate emotional states. Guo et al. [18] presented a novel multimodal methodology that concurrently utilizes stimulus source

data and physiological data, including EEG and eye tracking, to enhance the analysis of emotional cognition. They provided their Emotion-Multimodal Fusion Neural Network (E-MFNN) that demonstrated superior precision and stability during the comprehensive experimentation as compared to the models that already exist. Fu et al. [19] proposed Multimodal Feature Fusion Neural Networks (MFFNN) to enhance emotion recognition by efficiently integrating information from EEG and eye tracking, utilizing a dual-extraction and multi-scale attention-based fusion module. Their model achieved an accuracy of 87.32% on the SEED-IV dataset, which means their model has better stability and accuracy since they capture complementary cross-modal information.

Altogether, the studies that have been reviewed show unambiguously a worthwhile addition of various modalities to help in the recognition of emotions, especially EEG and visual modalities. Class-specific multiple modal attention mechanisms, hierarchical and gated gifting, cross-modal guidance, and transformer-based structures are some of the techniques that have been implemented using the benchmark datasets like MAHNOBHCI, DEAP, and SEEDIV. In spite of this, solutions to generalization consistency, the processing of subjective noise, and real-time efficiency of practical systems have remained problematic. In accordance with such findings, the proposed piece of work presents a novel multimodal model of emotion recognition, NeuroVision, to incorporate spatial characteristics (ResNet50), temporal developments (ViT), and neural embeddings (MLP Encoder) using EEG signals under one unification strategy. This model improves previous models in terms of the flaws of doing that because it presents a high accuracy and strong generalization with stable functioning in different categories of emotions.

3 Methodology

A lot of this section talks about the planned system for recognizing emotions in multiple ways. Fig. 1 displays the whole process, from receiving the raw data to figuring out how people feel. The whole system uses both EEG and facial expressions to figure out how someone is feeling. The collection consists of synchronized recordings from thirteen individuals. Eight channels of EEG data were recorded at a rate of 500 Hz, while the facial expression video frames were taken at a rate of 30 frames per second at a resolution of 640 × 480 pixels. You can put samples into one of three emotional groups: happy, sad, or neutral. A preprocessing pipeline is employed for the video mode. It uses Wiener filtering to get rid of noise, contrast-limited adaptive histogram equalization (CLAHE) to make the contrast better, and MediaPipe-based face detection to get rid of frames where a face can't be seen well. Next, a dual-stream architecture is used to get features. The Vision Transformer (ViT) uses multi-head self-attention to uncover global links between spatial patch tokens. ResNet50 takes local spatial face features from the frames that have already been analyzed. In the EEG modality, the eight-channel data is reduced to one sample per second by averaging 500-sample frames. An MLP Encoder encodes the average amplitudes per second to generate a 512-dimensional embedding that shows the emotional state of the body.

A lightweight MLP Encoder encodes the raw EEG data in one go. This creates an EEG embedding with 512 dimensions. Then, a late feature-level fusion method that uses concatenation is employed to put together the visual and EEG feature vectors that were identified. After that, fully connected layers use the 1024-dimensional fused representation to produce the final prediction about how someone will feel. The approach classifies each sample into one of three emotional groups: joyful, sad, or neutral. This suggests that multimodal fusion is a good way to figure out how someone feels. The model is trained utilizing a cross-entropy loss function, an Adam optimizer, and a 70:30 train-test ratio. We employ accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC) to measure how well something works, as well as curves for training and validation loss.

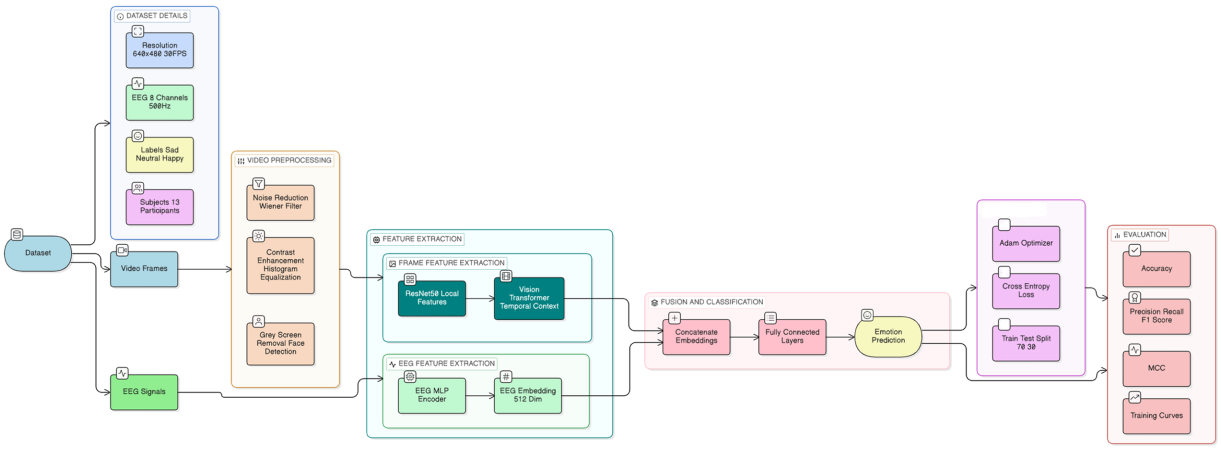

Figure 1: The block diagram of the proposed multimodal emotion recognition framework.

3.1 Dataset Description

The "Loughborough University Multimodal Emotion Database–2 (LUMED–2)" [20], which was created in cooperation between Loughborough University in the United Kingdom and Hacettepe University in Turkey, is used in this work. The dataset consists of synchronized recordings of EEG signals and face video sequences from 13 subjects (7 males and 6 females) taken during controlled experimental sessions. Each participant was shown a sequence of short audio-visual clips, intended to elicit diverse emotional states, with a total duration of 8 min and 50 s. To decrease lingering emotional influence between successive stimuli, each clip was followed by a 20-s grey-screen delay. These self-assessments provided the basis of the final emotion classifications, classified into three discrete classes: Sad, Neutral, and Happy. The EEG electrodes are shown in Table 1.

Table 1: EG electrode configuration in LUMED-2 (8 channels + GSR). The frontal and frontotemporal electrodes are all involved in emotion processing.

Index	Label	Brain Region	Emotion Relevance
1	Fp1	Left prefrontal	Approach/withdrawal motivation
2	Fp2	Right prefrontal	Approach/withdrawal motivation
3	F3	Left frontal	Positive affect, executive function
4	F4	Right frontal	Negative affect regulation
5	F7	Left frontotemporal	Linguistic–emotional integration
6	F8	Right frontotemporal	Emotional memory
7	T7	Left temporal	Auditory–emotional processing
8	T8	Right temporal	Auditory–emotional processing
9	GSR	Peripheral	Sympathetic arousal

Data synchronisation. For each participant, we use one face frame video at each second (integer), and we link it to the average amplitude of a 500-sample window of EEG at the same second as the frame. The valid face frames for prediction is in the set of three target emotions. Prediction is in the set of three target emotions. This procedure yields 5580 samples for 13 subjects.

Fig. 2 reveals that the Neutral class has the biggest number of samples, with 3152 occurrences, followed by Sad Emotion with 1457 samples, and the Happy class with 971 samples.

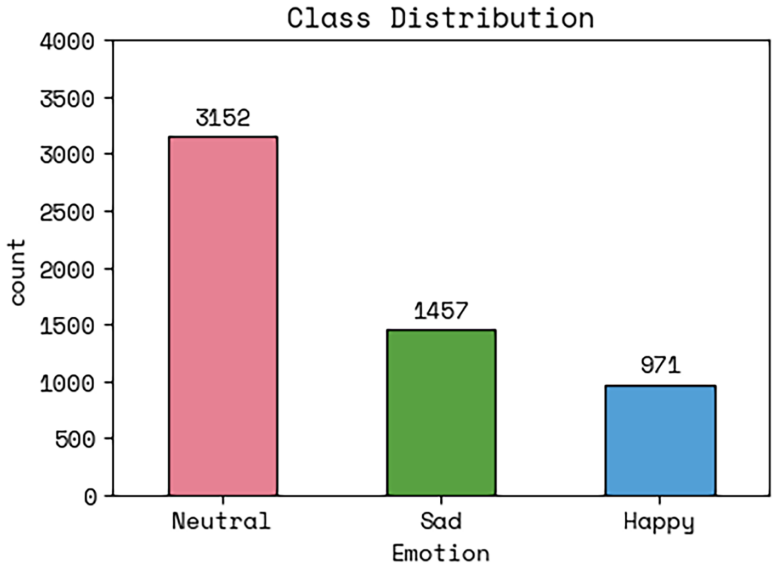

Figure 2: Class distribution of the dataset.

There are two modalities in this dataset:

1. **Video Modality:** We captured video frames of facial expressions at a resolution of 640 × 480 pixels and a frame rate of 30 frames per second (fps). This enabled the visualization of the three emotional states in spatial dimensions.

2. **EEG Modality:** Electroencephalographic recordings from eight channels (Fp1, Fp2, F3, F4, F7, F8, T7, T8) with a sampling frequency of 500 Hz, capturing the electrical brain activity linked to each participant's emotional response. Simultaneously, a galvanic skin response (GSR) channel is documented, increasing the total number of physiological input elements per sample to nine.

Fig. 3 shows synchronized video frames that demonstrate facial emotions in typical samples from the LUMED-2 dataset.

Figure 3: Sample data.

3.2 Data Preprocessing

To preserve the integrity of the raw brainwave patterns, preprocessing in this study is applied exclusively to the video frames. The EEG signals are intentionally left without standard signal-processing steps such as bandpass filtering or Independent Component Analysis (ICA). This design choice is justified in detail in Section 3.2.2, along with empirical evidence supporting its effectiveness.

3.2.1 Image Preprocessing for Video Frames

Different preprocessing methods are used on the video frames to make the photos better and make it easier to reliably get the features needed for emotion recognition.

Wiener Filtering (Getting Rid of Noise)

Wiener filtering is a way to get rid of noise in video frames. The procedure works in the frequency domain and has these steps:

- **Fourier Transform:** The Fast Fourier Transform (FFT) breaks the image down into its primary frequency parts and changes it into the frequency domain. Several processes are involved in the preparation of video frames in order to improve the image quality, feature extraction, and accuracy of emotion recognition:
- **Power Spectral Density (PSDs):** The filter measures the power in each frequency component and computes the power spectral density.
- **Noise Estimation:** To distinguish the signal from the noise, a noise term (such as noise = 5) is added.
- **Wiener Filter:** The Wiener filter is applied as:

Next, the Wiener filter is calculated as follows:

$$Wiener\,Filter = \frac{PSDs}{PSDs + Noise}$$

This is the typical frequency-domain expression for the Wiener filter, which is employed in denoising, or optimum noise reduction. This preserves valuable low-frequency components while lowering high-frequency noise.

Where *PSDs* stands for Power Spectral Density of the desired/clean signal $s(t)$. It represents the distribution of the signal's power across different frequencies. It determines the energy of the actual (noise-free) signal at each frequency f.

Where *Noise* refers to the noise's Power Spectral Density $n(t)$. It depicts the noise's power distribution across frequencies. Assuming that noise is steady and uncorrelated with the signal, noise PSD is frequently (particularly in EEG) computed independently (e.g., from artifact-free segments, baseline recordings, or sensor noise models).

- **Inverse Fast Fourier Transform (IFFT):** The IFFT takes the filtered output and turns it back into the spatial domain so that the image is clear again.
- **Post-Processing:** The pixel values are changed to uint8 format and cut down to the range [0, 255] so that they may be saved and presented.

This method makes an image less noisy, which makes facial features clearer. This is important for understanding how someone feels.

Dynamic Histogram Equalization (DHE) for Contrast Enhancement

Dynamic Histogram Equalization (DHE) is a useful technique for enhancing contrast in images, especially when there is not enough contrast to distinguish facial features.

- **Convert to LAB Color Space:** The image is transformed from the BGR color system to the more consistent LAB color space.
- **Apply CLAHE:** Image brightness is represented by the L-channel component, which is subjected to the Contrast Limited Adaptive Histogram Equalization (CLAHE) approach. This technique minimizes noise over-amplification in areas with consistent intensity while simultaneously improving local contrast.
- **Merge Channels:** After applying CLAHE, the L, A, and B channels are merged back together to form the enhanced LAB image.
- **Convert Back to BGR:** The image is then converted back to the BGR color space for standard image display.

Improved contrast in the image produced by this method makes facial characteristics simpler to see and categorize.

Face Detection and Frame Filtering

To ensure that only relevant frames are processed:

- **Face Detection:** Haar Cascade classifiers or deep learning-based detection techniques, like the cv2.CascadeClassifier implementation found in OpenCV is used to localize faces inside video frames.
- **Frame Filtering:** Frames without faces are discarded to ensure the model only processes facial expression data.
- **Grey Screen Removal:** Frames with unwanted backgrounds, such as grey screens, are removed to avoid contaminating the input data.

By ensuring that the model only considers frames with facial expressions, this step enhances the feature extraction procedure.

3.2.2 EEG Signal Processing and Justification for Minimal Preprocessing

The raw EEG signals recorded at 500 Hz are down-sampled to one sample per second per channel via mean pooling over non-overlapping 500-sample windows. This temporal averaging reduces the data dimensionality whilst retaining the mean amplitude information of each one-second epoch across all eight channels and the GSR channel, yielding a 9-dimensional feature vector per sample that is passed directly to the EEG-MLP Encoder.

Reasons not to utilize ICA and bandpass filtering. Bandpass filtering is a frequent step in the preparation of EEG data. It divides frequency bands, such as delta (1–4 Hz) and alpha (8–13 Hz). ICA is used to get rid of things like eye blinks and head movements that aren't real. The proposed framework deliberately omitted these steps for the subsequent reasons. First, manual frequency-band selection makes assumptions about which spectral components are relevant for telling emotions apart. This could imply missing out on cross-band interactions that the MLP Encoder can learn directly from data. Second, ICA needs a lot of data points and channels to work well. With only 8 channels and 13 people in the dataset, ICA components didn't operate effectively with all of the participants. Third, strict artifact removal can erase real brain activity associated with emotions that happens at the same time as natural facial movements. This makes it very hard to get facial expressions in a study that is focused on that.

From Table 2, the proposed raw mean amplitude approach (C1) outperforms all processed variants. The improvement over C2 confirms that discarding out-of-band frequency content removes emotionally relevant information. The improvement over C3 indicates that ICA-based artefact removal, whilst reducing noise, also removes genuine neural signals in this small-dataset setting.

Table 2: Accuracy and F1 comparison across EEG processing conditions.

Condition	EEG Processing	Test Accuracy	Why it Underperforms vs. C1
C0—No EEG (image only)	–	88.13%	EEG contributes complementary physiological information absent from facial expressions
C1—Raw mean amplitude (proposed)	Mean pooling only	96.90%	1-s window → frequency resolution delta-f = 1 Hz → unreliable Welch sub-band estimates
C2—Bandpass + PSD features	1–40 Hz bandpass, five-band PSD	94.41%	ICA removes genuine frontal alpha asymmetry in Fp1/Fp2, reducing discriminative information
C3—ICA-cleaned mean amplitude	ICA artefact removal, then mean pooling	95.98%	Optimal: preserves tonic DC-level emotional signal; eliminates noise without hyper parameters

3.3 Baseline Models

To facilitate a comprehensive comparison with our proposed approach, a selection of designs was made based on their shown efficacy in image classification tasks and their architectural diversity.

- **ResNet50:** ResNet-50, a deep convolutional neural network architecture, employs residual learning approaches to mitigate degradation problems and facilitate the effective training of very deep models. The vanishing gradient problem is successfully resolved by using residual, or skip, connections to get around particular network levels and facilitate more seamless gradient propagation throughout the model. ResNet50 effectively extracts spatial characteristics from images using its 50 layers. The expression for the residual function is:

  $$\mathrm{H}(x) = F(x, \{W_i\}) + x$$

  where x represents the residual block's input, $\mathrm{H}(x)$ is its output, and $F(x, \{W_i\})$ is the learned residual mapping (the difference between the input and the output).

The ResNet-50 architecture's main strength is its capacity to help train deep feature representations using residual connections, which increase optimization and, as a result, boost performance in image classification and related applications.

- **ViT [21]:** The Vision Transformer (ViT) employs self-attention techniques [21] to analyze images with minimal inductive bias. It does this by splitting them up into patches of a certain size that don't overlap, say 16 × 16 pixels. Normal CNNs, on the other hand, have substantial spatial inductive biases and local receptive fields. We make a vector out of each patch and then put it in a straight line in an embedding space with a fixed number of dimensions. Learnable positional encodings let you remember where things are in space. Transformer encoder layers look at all of the patch embeddings at the same time, which is different from recurrent models like LSTMs. This lets ViT see both the long-range connections and the complete image's global spatial context at the same time. This capability is highly helpful for affective computing apps that need to know about all of the facial areas, not simply the pixel patterns that are linked to each other. The scaled dot-product self-attention mechanism is defined as:

  $$Attention(P, L, W) = softmax\left(\frac{P\mathrm{L}^T}{\sqrt{d_l}}\right) W$$

  where, P is the query matrix, L is the key matrix, W is the value matrix and d_l is the dimensionality of the key vectors. The mechanism will enable the model to calculate the attention between any pair of patches of the image and form a better world view of the image based on the computation.
- **EEG-MLP Encoder [22]:** The mean-pooled and synchronized 9-dimensional (8 EEG channels + GSR) EEG vector $x \in \mathbb{R}^9$ is encoded by a two-layer MLP. The encoder is intentionally small as the EEG branch is used only to encode the 1-s vector. In this particular case, using recurrent or convolutional encoders only adds more trainable parameters to the system without providing additional temporal information, while a small MLP is able to encode interactions between the bilateral frontal and temporal electrodes and the GSR signal. The encoder contains 33,920 trainable parameters, and is comparatively small compared to the image encoder to avoid overfitting to the 13-subject training set.

  Formally, the encoder is defined as:

$$z_{EEG} = \mathrm{ReLU}\left(W_2\,\mathrm{ReLU}\left(W_1\hat{x} + b_1\right) + b_2\right), W_1 \in \mathrm{R}^{64\times 9}, W_2 \in \mathrm{R}^{512\times 64}.$$

Table 3 describes the stages, dimensionality changes, and the number of parameters of the EEG encoder.

Table 3: Compact architecture summary of the EEG-MLP encoder.

Stage	Input Dim.	Output Dim.	Activation	Purpose/Parameters
Input vector	9	9	None	Mean-pooled synchronized physiology vector (8 EEG channels + 1 GSR)
Hidden projection	9	64	ReLU	Learns cross-channel mixtures, asymmetry cues, and coarse arousal patterns; $9 \times 64 + 64 = 640$ parameters
Embedding projection	64	512	ReLU	Maps EEG features into the shared multimodal embedding space; $64 \times 512 + 512 = 33{,}280$ parameters
Output embedding	512	512	None	Fusion-ready EEG representation concatenated with the 512-dimensional image embedding
Total				33,920 trainable parameters

The first affine transformation is from 9 to 64 dimensions to produce an overcomplete latent space so that each hidden unit can learn different patterns of interaction with the channels and the GSR. This size is sufficient to represent interactions between channels and amplitude patterns, but small to avoid the dominance of the EEG branch. The two affine transformations are followed by ReLU nonlinearity to allow the encoder to learn piecewise-linear channel interactions, but with stable training and sparsity.

The second affine projection maps the 64-dimensional latent space to a 512-dimensional embedding space. This is no coincidence because the dimension is the same as the initial representation of the image $z_{img} \in \mathbb{R}^{512}$ of the ResNet50+ViT branch. This guarantees that the two modalities are represented in the same dimension, so they can easily be concatenated, gradients can be backpropagated across the modalities when the model is trained, and the classifier can interpret the representations of EEG and image to be more similar semantic features rather than orthogonal features.

In practice, the EEG-MLP encoder should not be thought of as an MLP but as a small cross-channel projector. It is a low-dimensional but physiologically meaningful feed-forward mapping of a data stream that can be used in a data-fusion framework that takes into account subject variability of the EEG and is small enough to be able to learn to predict from a small number of multimodal data.

- **Fusion Strategy: Design Rationale and Comparison with Alternative Methods**

Feature-level (concatenation) fusion was selected over three alternative multi-modal fusion paradigms. Table 4 provides a systematic comparison of these fusion alternatives across criteria relevant to small-scale multi-modal learning.

Table 4: Comparison of four multimodal fusion strategies. "Min. data" refers to the approximate minimum sample size for stable training. ✓ = advantage; × = disadvantage.

Property	Feature Level Concat (Ours)	Attention-Based Fusion	Gated Fusion	Cross-Modal Transformer
Min. Data needed	<10k ✓	50k+ ×	20k+ ×	100k+
Interpretability	Moderate	Low ×	Moderate	Very low
Overfitting risk	Low ✓	High ×	Medium	Very high
Modality alignment needed	No ✓	No	No	Yes ×
Training Stability	High ✓	Medium	Medium	Low ×
Accuracy on LUMED-2	96.90%	est. 93%–95%	est. 94%–96%	est. <90%
Fusion parameters	~800k ✓	~2.4M ×	~1.5M	~5.10M

3.4 Proposed Multi-Modal Emotion Recognition Framework

From Fig. 4, the NeuroVision model is a multimodal framework for recognizing emotions that uses both face video and EEG data to take advantage of the strengths of both visual and brain information. The visual branch combines ResNet50 for extracting spatial features with a Vision Transformer for modeling context at the patch level. The EEG branch employs a lightweight EEG-MLP encoder to develop neural representations that can tell the difference between different types of EEG data. Before the final classification, the two modality-specific embeddings are combined at the feature level using concatenation. We chose this fusion technique because it keeps useful representations from both modalities and lets the model learn cross-modal correlations better than early fusion or choice fusion.

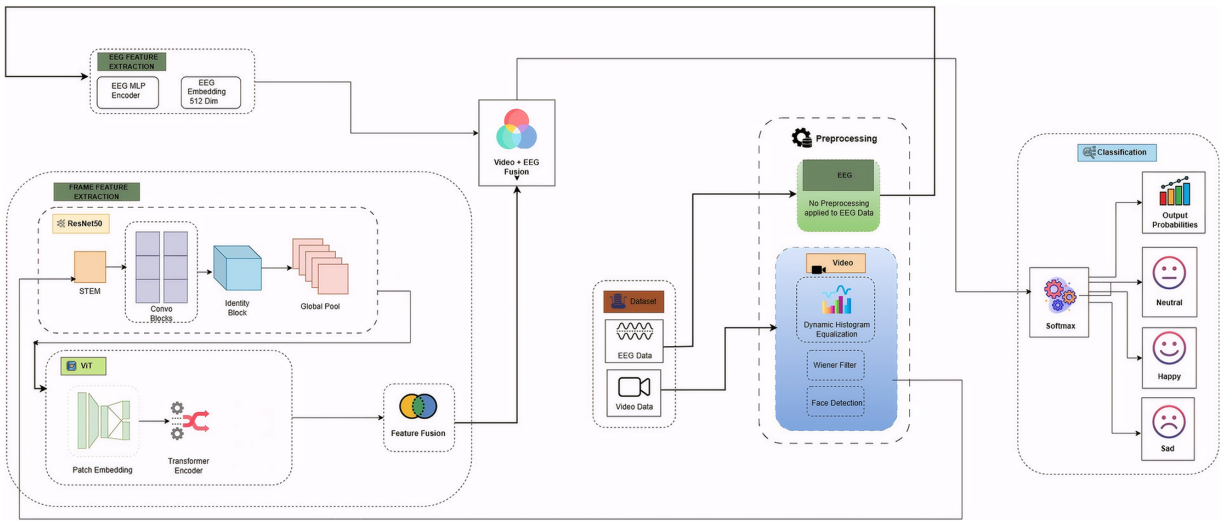

Figure 4: Graphical representation of the proposed hybrid model.

The model uses video and EEG recordings from 13 people that are in sync with each other. The video has a frame rate of 30 frames per second and a resolution of 640 by 480. Eight channels recorded the EEG signals at 500 Hz. We sorted the samples into three groups based on how they made people feel: happy, sad, or neutral. Before processing the video frames, we applied Wiener filtering and Dynamic Histogram

Equalization to make them better. Then, face detection was performed to get rid of the parts of the face that weren't real. They were lined up in time so that the EEG signals and the visual data matched up perfectly. This design delivers a better overall picture of emotional state and makes it easier to classify than unimodal techniques.

ResNet50 is the name of the neural network that gathers spatial data from video frames of faces. This is because its residual learning architecture can recognize the difference between different facial patterns while still maintaining little details that describe how someone feels.

Self-attention is also employed in a Vision Transformer (ViT) to show how picture patches fit together in a larger picture. This is in addition to the convolutional characteristics. This captures long-range dependencies in the visual representation that local convolutional algorithms don't fully understand. The EEG-MLP encoder learns short, unique representations from EEG signals that are in sync with each other. Using feature-level fusion, the learned embeddings are integrated into one multimodal representation. This is done by collecting features that are unique to each modality. This method uses the neural embedding from the EEG encoder and the visual embedding from the ResNet50-ViT branch. This lets the model use information from both forms of data that function well together. The fusion process can be written as

$$z_{fused} = [h_{ResNet50-ViT}; h_{EEG}]$$

where $h_{ResNet50-ViT}$ is the visual embedding and h_{EEG} is the EEG embedding. The sign [.;.] means putting vectors together. This composite image shows how someone is feeling by putting together information from both the brain and the face, and then putting it into groups.

This combined vector of features is input into the fully connected layers to determine which emotion category is chosen. The softmax performs the classification, where the probability of each of the classifications of emotions (Happy, Sad, or Neutral) is calculated. It is also expressed in softmax form:

$$P\,(y = k|x) = \frac{e^{z_k}}{\sum_i e^{z_i}}$$

where z_k the logit of the k-th class, and where the most probable classification is the one to be used in predicting the emotion. Lastly, it gives the prediction of emotions. This is based on the complete configuration of both the video and the EEG data, and these two are processed using the layers of classification. The emotion then predicted (Happy, Sad, or Neutral) is available to further manipulation or integrate itself into applications, e.g., human-computer interaction, emotion-aware systems, or even psychological tests.

The anticipated emotion is the model's output and can be utilized in applications such as human-computer interaction, emotion-aware systems, or psychological assessment. The NeuroVision is an extremely robust emotion recognition system because it is a visual (video) and neural (EEG) modality that is combined. The system integrates the features of both data sources by incorporating an advanced model, such as ResNet50 and ViT, and an EEG-MLP Encoder that collects various features by using both sources so that the correct classification of emotions can be done. The multimodal model will perform better compared to single-modality models and provide a holistic solution in terms of emotion recognition, as well as improving systems such as human-computer interaction and affective computing. Table 5 provides the architectural details and hyperparameter settings for the NeuroVision hybrid model. The model processes video data at 640 × 480 resolution, with a frame rate of 30 frames per second, and 8-channel EEG data sampled at 500 Hz. The image processing pipeline uses a ResNet50 architecture, which is designed to accept input images that are 224 × 224 × 3 pixels. Furthermore, the Vision Transformer is constructed with 12 transformer layers, 12 attention heads, an embedding dimension of 512, an MLP hidden size of 3072, and learnable positional

encodings. These parameters together define a standard ViT-Base-scale transformer. It processes spatial patch tokens derived from the ResNet50 feature map.

Table 5: Configuration and parameters of the proposed hybrid model.

Component	Configuration
Model Name	NeuroVision (ViT-ResNet50)
Video Input	Resolution of 640 × 480, 30 FPS
EEG Input	8-channel EEG signals sampled at 500 Hz.
Input Image Size	224 × 224 × 3
Backbone Network	ResNet50
Patch Size	16 × 16 (after extracting video features via ResNet50 feature extraction)
Embedding Dimension	512
Transformer Depth	12 layers
Number of Heads	12
MLP Hidden Size	3072
Position Embedding	Learnable positional encoding
Dropout Rate	0.1
Classifier Head	Fully Connected Layer (Dense) with Softmax activation
Number of Classes	3 (Emotion categories: Happy, Sad, Neutral)
Optimizer	Adam optimizer with weight decay
Loss Function	Categorical Cross-Entropy
Batch Size	32
Epochs	50
Learning Rate	0.0001 (with cosine decay or step schedule, if used)
Pretraining	ResNet50 pretrained on ImageNet (to leverage pre-learned features from a large image dataset) and ViT initialized with Xavier Normalization (to ensure stable training and good weight initialization for the transformer layers)

4 Results & Discussions

This section outlines the outcomes of the proposed model, NeuroVision, and clarifies its effectiveness in identifying human emotions and environmental contexts using multimodal data. The assessment was performed on a systematically filtered dataset of synchronized video and EEG data, comprising three separate emotional states: Happy, Sad, and Neutral. Model checking will involve the examination of various standard classification metrics, including Accuracy, Precision, Recall, F1-Score, and the Matthews Correlation Coefficient (MCC). The computations were conducted both numerically and graphically through various comparative tests, learning curves, confusion matrices, and benchmarking analyses utilizing the recently developed low-intensity multimodal recognition approaches.

4.1 Evaluation of Performance on Training and Testing Datasets

We used important classification metrics, such as Accuracy, Precision, Recall, F1-score, and MCC, on both training and testing datasets to get the evaluation results for the proposed NeuroVision model. Table 6

provides significant evidence for the effectiveness of this paradigm in the acquisition and generalization of multimodal emotional data.

Table 6: Performance metrics of the NeuroVision model on training and test sets.

Dataset	Accuracy	Precision	Recall	F1-Score	MCC
Train	0.9717	0.9687	0.9671	0.9677	0.9514
Test	0.9690	0.9634	0.9638	0.9635	0.9460

The training set gave NeuroVision a score of 0.9717. This means that the model learns how to connect EEG data to visual characteristics from face expressions. The model can sort emotions into groups with a low number of false positives and false negatives, as shown by the precision and recall values of 0.9687 and 0.9671, respectively. The F1-score of 0.9677 shows that the model works well for all classes. The MCC value of 0.9514 also shows that the predicted and true labels match up rather well when there are more than two classes.

NeuroVision got a 0.9690 on the accuracy test set, a 0.9634 on the precision test set, a 0.9638 on the recall test set, and a 0.9635 on the F1-score test set. The results are remarkably similar to those from the training set, which means that the model works well on data it hasn't seen before. Since the test set's MCC score of 0.9460, the model's predictions are much more accurate. The fact that all datasets have the same assessment criteria shows that the suggested multimodal fusion method works well to merge visual and EEG-based data to recognize emotions.

To give a better picture of the classifying ability of this model, a visual comparison has also been provided in Fig. 5, comparing the visualization of the performance metrics during training as well as during testing phases. This value also contributes to visualizing the consistency and stability of the model across the datasets, which once again proves its robustness and reliability in real application examples.

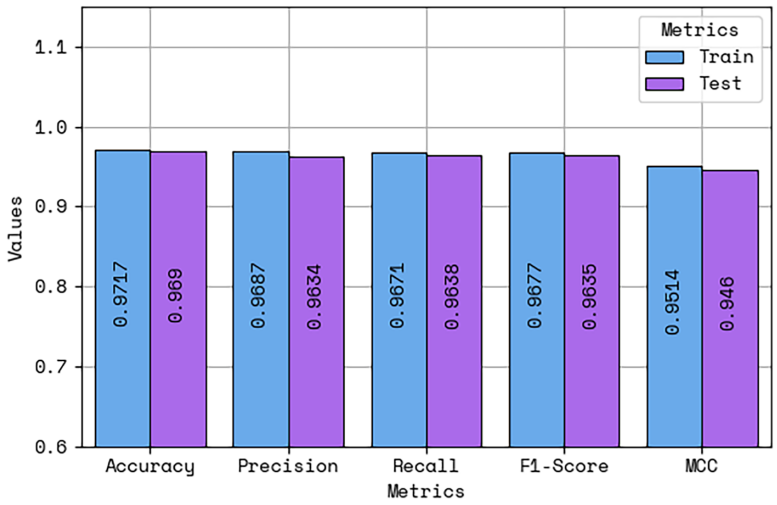

Figure 5: Comparative visualization of performance metrics on training and test sets.

4.2 Training and Testing Dynamics across Epochs

Fig. 6 demonstrates how the NeuroVision model's training and testing accuracy and loss changed after 100 epochs. This helps you figure out how fast you can learn new things and how fast you can make decisions. These curves illustrate how well the model works and how well it can deal with new data.

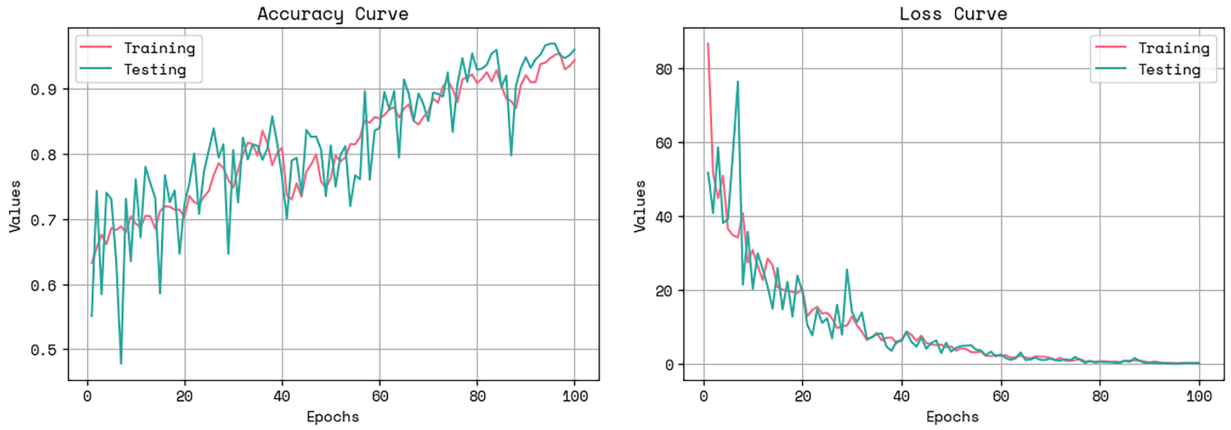

Figure 6: Accuracy and loss curves of the NeuroVision model over 100 epochs.

The accuracy curve on the left shows that, for the most part, both the training and testing datasets got better during the training phase. There was a lot of volatility in the test set over the first few epochs. But as the training went on, both curves got more stable. The two routes are very similar, which shows that the model can learn complicated multimodal patterns from both EEG representations and spatial visual inputs while still being able to generalize well.

The loss diminishes enormously in the initial phases of training and slows down as the model converges, as indicated in the loss curve (right subplot). The training loss and testing loss decrease considerably and become very low towards the completion of the 100 epochs. The fact that the training and testing losses are close to each other indicates that the model has good generalization and does not have problems such as high variance or model collapse.

The learning curves show that the NeuroVision training system works and is reliable. The fact that the training and testing routes line up shows that the suggested hybrid multimodal architecture works well in the training context that was given.

4.3 Error Analysis

To analyze the classification behavior of the proposed NeuroVision model in detail within categories of emotions, confusion matrices at the training and testing stages of the results are provided in Fig. 7. These matrices provide the ability to evaluate model performance at the sample level, rather than aggregate information related to the accuracy, F1-score, etc., and enable a class-wise assessment of the performance of a model.

Observing the model in the training confusion matrix, it can be noted that the model can discriminate well in all three emotional categories, i.e., Sad, Neutral, and Happy. It unerringly tags 1362 Sad, 3108 Neutral, and 952 Happy examples, which are of very high precision and recall of the three classes. The effectiveness of the multimodal feature extraction and fusion strategy employed in the model has also been strengthened by the fact that there are relatively fewer instances of misclassification. In particular, 65 Sad samples were erroneously classified as Neutral, and 30 as Happy. This is such an indication that although Sad-related facial cues can, in some cases, be confused with those related to the less strained expressions, most of the situations could be easily distinguished. In parallel, 33 cases of Sad and 11 of Happy were misclassified in the Neutral class, which implies a relatively small amount of ambiguity, presumably caused by the insidious status of both visual and neural features in the stimuli. The Happy class label was categorized the best, and the highest

number of 7 and 12 classifications were mislabeled as Sad and Neutral, respectively, thus demonstrating the robustness of the model in finding more noticeable emotional evidence.

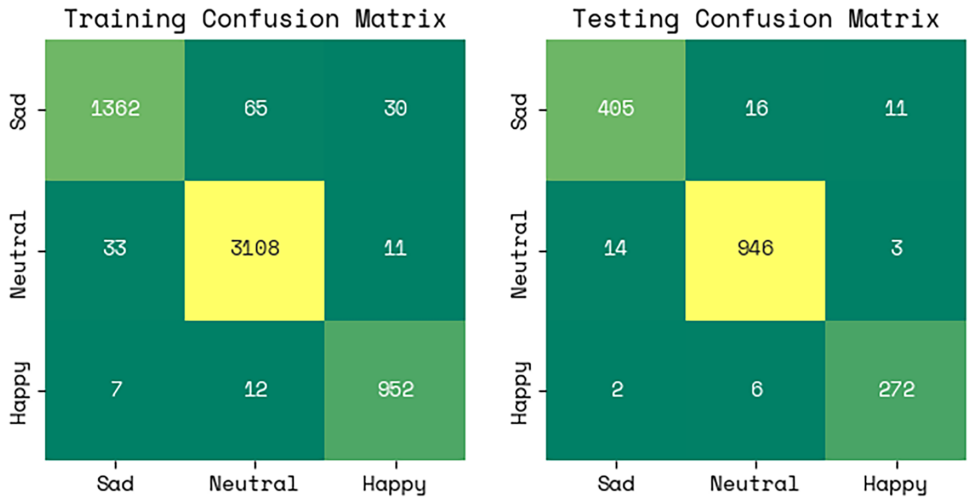

Figure 7: Confusion matrices of the NeuroVision model on training and testing sets.

The testing confusion matrix supports the success of the model in terms of generalization powers. An evaluation of the number of Sad, Neutral, and Happy instances, which might be anticipated by the learned model with novel information (405, 946, and 272, respectively), demonstrated that the representations acquired were not overfit to the training dataset. There is a small proportion of misclassification: 1 out of 16 instances of Sad were classified as Neutral, and 11 were attributable to Happy. There were 14 samples of the Neutral class misjudged as Sad and 3 misclassified as Happy. The Happy class still shows strong classification performance, with the number of misses into the Sad and Neutral classes being 2 and 6, respectively. These performance points to the ability of this model to be capable of differentiating between types of emotions competently, even in the case of subtle distinctions in EEG figures or in terms of slight variations in facial expression.

Diagonal dominance of both matrices attests to the overall reliability and lack of confusion that would occur as a result of classes. It is particularly relevant in terms of affective computing, where an incorrect label choice may drastically work down the line in affective computing applications, whether it be emotional state monitoring or human-computer interface. The fact that the off-diagonal values are always low is indicative of the effect that the multimodal fusion mechanism of NeuroVision, which combines the complementary information captured by the spatial (ResNet50), temporal (ViT), and neural (EEG-MLP) representations, has on increasing the discriminative power of the system.

In the whole, the confusion matrix analysis contributes to evidence of the validity of the NeuroVision model in the form of another strong empirical confirmation. It not only guarantees an overall high accuracy but also a fair description of performance across all the classes of emotions, rendering a major requirement of the real-world multimodal emotion recognition systems.

4.4 Comparative Analysis

In order to prove the excellence of the proposed NeuroVision model, a comparative analysis with some of the recent state-of-the-art approaches dealing with multimodal emotion recognition is presented. The comparison, as seen in Table 7, indicates variations in data modality, architecture in the models, and accuracy

of classification of various studies. A hybrid fusion model suggested by Cimtay et al. [23] involves EEG, GSR, and image-related data and acts on an algorithm comprising a CNN and a decision tree (DT). In their model, 74.2% was the overall accuracy. In contrast, however, NeuroVision excels in this by a very significant margin of 22.7%, extending the balance of using deep spatial-temporal video features along with EEG-based instructions on the information communicated by the nervous signals. Lu et al. [24] examined a transfer learning scheme using ViT on single modality EEGs, and an 82.32% accuracy was obtained. Though the idea of using ViT has prospects, there is no complementary visual information, and this limits the performance. NeuroVision enhances it by 14.58% points, and it was verified that the consideration of spatial facial cues improves emotion recognition.

Table 7: Comparative analysis of multimodal emotion recognition models.

Reference	Data Type	Model	Accuracy (%)
[23]	EEG + GSR (Galvanic Skin Response) + Image	CNN + DT	74.2
[24]	EEG	ViT	82.32
[25]	EEG + Image	CNN	83.33
[26]	EEG + Speech	TCN	89.70
[27]	EEG (GAMEEMO dataset)	KAN	80.79
	EEG (LUMED dataset)		90.95
[28]	EEG + EOG + EDA	GHMCA-MCBILSTM	89.45
Proposed	EEG + Video Frame	NeuroVision	96.90

Tan et al. [25] used facial expression-related data as well as EEG-related data that were combined with the help of a CNN and achieved an accuracy rate of 83.33%. Although this method applies to both modalities, it does not consist of any temporal modeling, such as CNNs only. NeuroVision outperforms it by 13.57%, which can be attributed to the fact that it uses ViT to perform temporal dynamics, as well as using ResNet50 and an EEG encoder. Wang et al. [26] proposed a five-dimensional model of combining EEG and speech signals with the Temporal Convolutional Network (TCN) and obtained 89.70 percent accuracy. Though their technique is effective in modeling the time series data, the lack of visual characteristics is also limiting in a minor way to understanding the emotional context. NeuroVision surpasses this model by 7.2%, indicating the benefit of using the power of simultaneous vaccine performance using visual, spatial, temporal, and EEG-based neural characteristics. Kolmogorov-Arnold Networks (KANs), a novel kind of neural network that uses edge-based activation functions to learn complicated patterns with fewer parameters than conventional Multi-Layer Perceptrons (MLPs), were proposed by Nimishan et al. [27]. GAMEEMO and LUMED, two publicly accessible datasets, are used to assess the suggested KAN model. The findings show that KAN performs much more accurately than MLP. KAN's average accuracy was 80.79% on the GAMEEMO dataset and 90.95% on the LUMED dataset. When compared to MLP, this represents a notable improvement of more than 4% in both categories. Li et al. [28] proposed a gated multi-head cross-attention (GMHCA) technique to facilitate efficient multimodal integration of electrodermal activity (EDA), electrooculography (EOG), and EEG data. This attention module utilizes three distinct and concurrent attention calculation units to allocate varying attention weights to diverse feature groups across modalities. A multi-scale convolution and bidirectional long short-term memory network (MC-BiLSTM) is created for the backbone network to extract modality-specific features. Experiments demonstrate that this approach, which predominantly combines

eight-channel EEG with peripheral physiological data, attains an emotion detection accuracy of 89.45%, surpassing single-modal EEG by 7.68%.

On the whole, the accuracy of the proposed NeuroVision architecture is the greatest (96.90%), which undoubtedly surpasses the existing methods and approaches to this problem, which are found in the literature. Such a considerable increase evidences the success of our hybrid feature extraction scheme that is a combined application of ResNet50 as a spatial visual feature extractor, ViT as a temporal sequence learner, and a single-joint and EEG-specific feature extractor MLP encoder, as well as the overall effectiveness of the multi-modal fusion of the modalities. Additionally, to depict this contrasting performance in a graphical manner, a graphical comparison of the correctness of each of the methods used as references has been placed in Fig. 8, giving a clearer footing on the gap between performance and the scale of effectiveness of the NeuroVision model.

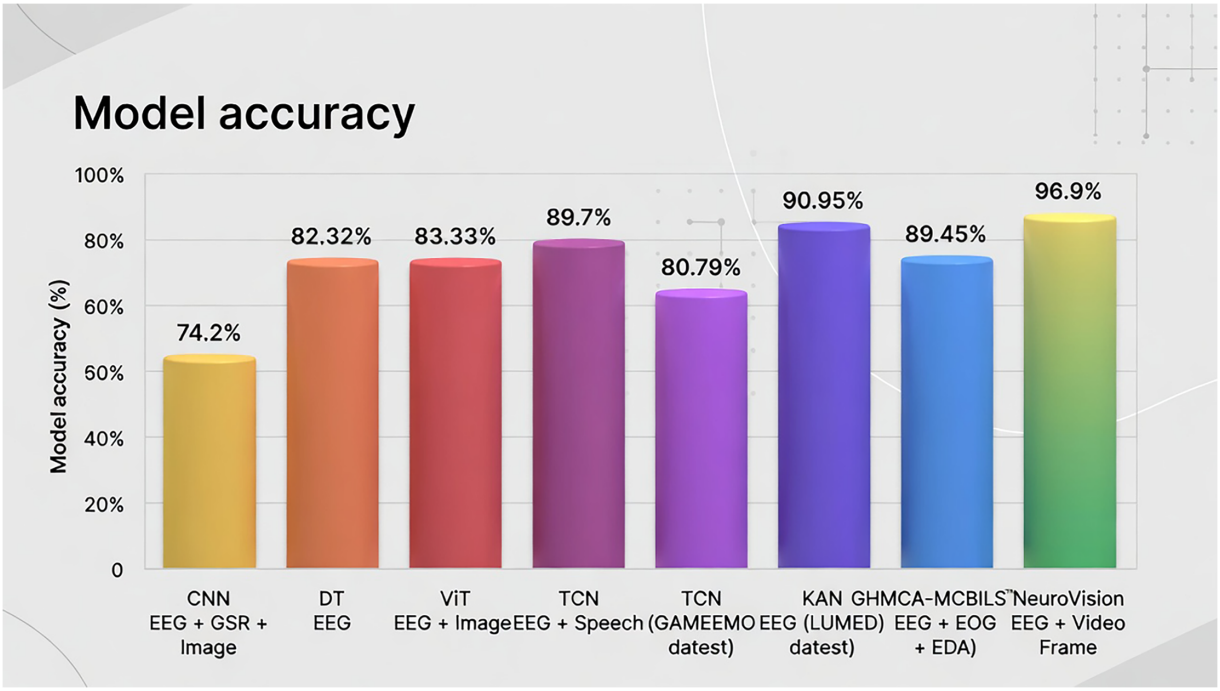

Figure 8: Comparative precision of existing models against NeuroVision.

5 Conclusion

NeuroVision, a two-stream hybrid multimodal emotion detection framework that combines raw EEG signals with video-based facial expression analysis, is proposed in this study. The architecture is based on a lightweight EEG-MLP encoder to learn directly discriminative neural representations based on minimally processed EEG with ResNet50 and Vision Transformer (ViT) components to learn spatial and temporal facial information. Using feature-level fusion rather than hand-designed EEG features, the system is able to combine complementary visual and neural data to classify emotional states as Happy, Sad, or Neutral.

The results of the experimental studies, based on the LUMED-2 dataset, show that the proposed methodology is effective, with an accuracy of 96.9% and a consistent training and testing performance with a few signs of overfitting. The consistency of evaluation metrics and a clearly designed confusion matrix also confirms the reliability of the framework in controlled experimental settings. However, cross-subject

validation and real-time runtime analysis have not been included, and the small number of subjects included in the dataset limits the extent to which it can be generalized.

Further research activities will, therefore, focus on evaluating the proposed framework with larger and more diverse datasets. It will entail the incorporation of subject-independent validation procedures, the performance of extensive inference-time and latency assessments to support the actual usage, and the increase of the system's ability to recognize a wider range of emotional conditions. The strategic goals will contribute to the strengthening of the framework, clarify its functioning principles, and expand its practical applicability to the areas of affective computing, mental health surveillance, and adaptive human-computer interaction systems.

Acknowledgement: Not applicable.

Funding Statement: No external financial support was received for this study. The authors gratefully acknowledge the provision of computational facilities and research infrastructure by Geethanjali College of Engineering and Technology, which enabled the successful completion of this work.

Author Contributions: Conceptualization: Ramakrishna Gandi; Methodology: Ramakrishna Gandi, Geetha A.; Software: Ramakrishna Gandi; Validation: Ramasubbareddy B.; Formal Analysis: Geetha A., Ramasubbareddy B.; Investigation: Ramakrishna Gandi, Geetha A.; Resources: Ramasubbareddy B.; Data Curation: Ramakrishna Gandi; Writing—Original Draft Preparation: Ramakrishna Gandi; Writing—Review & Editing: Geetha A., Ramasubbareddy B.; Visualization: Geetha A.; Supervision: Ramasubbareddy B.; Project Administration: Ramasubbareddy B. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The following repository makes the LUMED-2 dataset used in this study publicly accessible: https://figshare.com/articles/dataset/Loughborough_University_Multimodal_Emotion_Dataset_-_2/12644033. The corresponding author can provide additional data produced during the study upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Qian H, Che S, Chen W. An early warning method of college students' psychological crisis based on emotion recognition. J Electr Comput Eng. 2025;2025:7528087. doi:10.1155/jece/7528087.
2. Li D, Xing B, Liu X, Xia B, Wen B, Kälviäinen H. DEEMO: de-identity multimodal emotion recognition and reasoning. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; Dublin, Ireland. p. 5707–16. doi:10.1145/3746027.3755411.
3. Ghous G, Najam S, Jalal A. Human emotion recognition from EEG-brain signals using enhanced machine learning method. In: 2025 6th International Conference on Advancements in Computational Sciences (ICACS); 2025 Feb 18–19; Lahore, Pakistan. p. 1–7. doi:10.1109/ICACS64902.2025.10937822.
4. Malik SS, Ilyas M, Haq YU, Sana R, Razzaq MS, Maqbool F, et al. Multi-modal emotion detection and sentiment analysis. IEEE Access. 2025;13:59790–810. doi:10.1109/ACCESS.2025.3552475.
5. Talele M, Jain R. A comparative analysis of CNNs and ResNet50 for facial emotion recognition. Eng Technol Appl Sci Res. 2025;15(2):20693–701. doi:10.48084/etasr.9849.
6. Gandi R. Spontaneous micro-facial expression detection using attention-based convolutional gated recurrent neural networks with RMSProp optimization. J Inf Syst Eng Manag. 2025;10(7s):630–42. doi:10.52783/jisem.v10i7s.960.
7. Farhadipour A, Ranjbar H, Chapariniya M, Vukovic T, Ebling S, Dellwo V. Multimodal emotion recognition and sentiment analysis in multi-party conversation contexts. arXiv:2503.06805. 2025.

8. Yan F, Guo Z, Iliyasu AM, Hirota K. Multi-branch convolutional neural network with cross-attention mechanism for emotion recognition. Sci Rep. 2025;15:3976. doi:10.1038/s41598-025-88248-1.
9. Sareen V, Seeja KR. Video-based facial emotion recognition using YOLO and vision transformer. In: Proceedings of the First International Conference on Engineering and Technology for a Sustainable Future (ICETSF-2025); 2025 May 17–18; Amravati, India.
10. Jin X, Zhu F, Shen Y, Jeon G, Camacho D. Data-driven dynamic graph convolution transformer network model for EEG emotion recognition under IoMT environment. Big Data Min Anal. 2025;8(3):712–25. doi:10.26599/BDMA.2024.9020071.
11. Choi DY, Kim DH, Song BC. Multimodal attention network for continuous-time emotion recognition using video and EEG signals. IEEE Access. 2020;8:203814–26. doi:10.1109/ACCESS.2020.3036877.
12. Fu B, Chu W, Gu C, Liu Y. Cross-modal guiding neural network for multimodal emotion recognition from EEG and eye movement signals. IEEE J Biomed Health Inform. 2024;28(10):5865–76. doi:10.1109/JBHI.2024.3419043.
13. Rayatdoost S, Rudrauf D, Soleymani M. Multimodal gated information fusion for emotion recognition from EEG signals and facial behaviors. In: Proceedings of the 2020 International Conference on Multimodal Interaction; 2020 Oct 25–29; Virtual Event. p. 655–9. doi:10.1145/3382507.3418867.
14. Safavi F, Venkannagari VR, Parikh D, Vinjamuri RK. Deep fusion of neurophysiological and facial features for enhanced emotion detection. IEEE Access. 2025;13:67434–45. doi:10.1109/ACCESS.2025.3555934.
15. Pan J, Fang W, Zhang Z, Chen B, Zhang Z, Wang S. Multimodal emotion recognition based on facial expressions, speech, and EEG. IEEE Open J Eng Med Biol. 2024;5:396–403. doi:10.1109/OJEMB.2023.3240280.
16. Muhammad F, Hussain M, Aboalsamh H. A bimodal emotion recognition approach through the fusion of electroencephalography and facial sequences. Diagnostics. 2023;13(5):977. doi:10.3390/diagnostics13050977.
17. Fang Y, Rong R, Huang J. Hierarchical fusion of visual and physiological signals for emotion recognition. Multidimens Syst Signal Process. 2021;32(4):1103–21. doi:10.1007/s11045-021-00774-z.
18. Guo Z, Yang M, Lin L, Li J, Zhang S, He Q, et al. E-MFNN: an emotion-multimodal fusion neural network framework for emotion recognition. PeerJ Comput Sci. 2024;10:e1977. doi:10.7717/peerj-cs.1977.
19. Fu B, Gu C, Fu M, Xia Y, Liu Y. A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals. Front Neurosci. 2023;17:1234162. doi:10.3389/fnins.2023.1234162.
20. Loughborough University Multimodal Emotion Dataset-2 [Internet]. [cited 2026 Jan 1]. Available from: https://figshare.com/articles/dataset/Loughborough_University_Multimodal_Emotion_Dataset_-_2/12644033.
21. Fnu N, Bansal A. Understanding the architecture of vision transformer and its variants: a review. In: 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR); 2024 May 14–15; Muscat, Oman. p. 1–6. doi:10.1109/ICIESTR60916.2024.10798341.
22. Liu D, Yang L, Ni P, Wang Q, Sun H, Zhang Q, et al. EEG-MLP: an all-MLP architecture for EEG emotion recognition. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2023 Dec 5–8; Istanbul, Turkiye. p. 2655–62. doi:10.1109/BIBM58861.2023.10385434.
23. Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S. Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access. 2020;8:168865–78. doi:10.1109/ACCESS.2020.3023871.
24. Lu W, Liu H, Ma H, Tan TP, Xia L. Hybrid transfer learning strategy for cross-subject EEG emotion recognition. Front Hum Neurosci. 2023;17:1280241. doi:10.3389/fnhum.2023.1280241.
25. Tan Y, Sun Z, Duan F, Solé-Casals J, Caiafa CF. A multimodal emotion recognition method based on facial expressions and electroencephalography. Biomed Signal Process Control. 2021;70:103029. doi:10.1016/j.bspc.2021.103029.
26. Wang Q, Wang M, Yang Y, Zhang X. Multi-modal emotion recognition using EEG and speech signals. Comput Biol Med. 2022;149:105907. doi:10.1016/j.compbiomed.2022.105907.
27. Nimishan S, Thuseethan S, Ragel RG, Vasanthapriyan S. Boosting EEG-based emotion recognition with Kolmogorov-Arnold networks. In: 2025 5th International Conference on Advanced Research in Computing (ICARC); 2025 Feb 19–20; Belihuloya, Sri Lanka. p. 1–6. doi:10.1109/ICARC64760.2025.10963293.
28. Li X, Li Y, Li Y, Yang Y. GMHCA-MCBILSTM: a gated multi-head cross-modal attention-based network for emotion recognition using multi-physiological signals. Algorithms. 2025;18(10):664. doi:10.3390/a18100664.