I.J. Image, Graphics and Signal Processing, **2026, 1, 163-176**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2026.01.10

Multimodal Emotion Recognition Using EEG and Facial Expressions with Potential Applications in Driver Monitoring

Ch. Raga Madhuri
Department of Computer Science and Engineering, Faculty of Engineering, Siddhartha Academy of Higher Education, Deemed to be University, Vijayawada -520007, Andhra Pradesh, India
E-mail: madhuri.chandra1209@gmail.com
ORCID iD: https://orcid.org/0000-0002-3687-0783

Anideep Seelam*
Department of Computer Science and Engineering, Student of Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada -520007, Andhra Pradesh, India
Email: anideepabhinay@gmail.com
ORCID iD: https://orcid.org/0009-0004-2289-2430
*Corresponding Author

Fatima Farheen Shaik
Department of Computer Science and Engineering, Student of Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada -520007, Andhra Pradesh, India
Email: fatimafarheen2605@gmail.com
ORCID iD: https://orcid.org/0009-0000-7919-3092

Aadi Siva Kartheek Pamarthi
Department of Computer Science and Engineering, Student of Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada -520007, Andhra Pradesh, India
Email: aadipamarthi8055@gmail.com
ORCID iD: https://orcid.org/0009-0008-6277-3662

Mohan Kireeti Krovi
Department of Computer Science and Engineering, Student of Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada -520007, Andhra Pradesh, India
Email: kireetikrovi@gmail.com
ORCID iD: https://orcid.org/0009-0008-2750-0626

Received: 10 June, 2025; Revised: 26 August, 2025; Accepted: 18 December, 2025; Published: 08 February, 2026

Abstract: Mental conditions such as fatigue, distraction, and cognitive overload are known to contribute significantly to traffic accidents. Accurate recognition of these cognitive and emotional states is therefore important for the development of intelligent monitoring systems. In this study, a multimodal emotion recognition framework using electroencephalography (EEG) signals and facial expression features is proposed, with potential applications in driver monitoring. The approach integrates Long Short-Term Memory (LSTM) networks and Transformer architectures for EEG-based temporal feature extraction, along with Vision Transformers (ViT) for facial feature representation. Feature-level fusion is employed to combine physiological and visual modalities, enabling improved emotion classification performance compared to unimodal approaches. The model is evaluated using accuracy, precision, recall, and F1-score metrics, achieving an overall accuracy of 96.38%, demonstrating the effectiveness of multimodal learning. Although the experiments are conducted on general-purpose emotion datasets, the results indicate that the proposed framework can serve as a reliable foundation for driver monitoring applications, such as fatigue, distraction, and cognitive state assessment, in intelligent transportation systems.

Index Terms: Advanced Driver Safety, Electroencephalography (EEG), Cognitive Monitoring, LSTM, CNN, Fatigue Detection, Transformers, Vision Transformers (ViT), Driver Safety

1. Introduction

Road safety continues to be an important global concern, and driver fatigue, distraction, and mental workload have been identified as major contributing factors to road accidents [1,2]. Such cognitive impairments negatively impact a driver's response time, decision-making ability, and overall performance, thereby increasing the likelihood of crashes [3]. Research indicates that a significant proportion of traffic accidents are associated with compromised cognitive states, including drowsiness and cognitive overload [4]. Conventional driver monitoring systems primarily rely on behavioral indicators such as gaze direction, head movement, and steering patterns, which may fail to detect cognitive impairment before it manifests as observable actions. This limitation highlights the need for advanced monitoring approaches capable of identifying unsafe cognitive states at an early stage [5].

Recent studies have increasingly explored sensor-based and learning-driven monitoring systems to enhance driver safety, including the use of electroencephalography (EEG) signals and facial analysis techniques [6,29–32]. EEG signals provide direct insight into brain activity and enable the detection of subtle cognitive changes related to fatigue, distraction, and mental workload, while facial expressions offer complementary behavioral cues. Motivated by these findings, this research proposes a multimodal framework that integrates EEG signals and facial expression features for cognitive state evaluation, with potential applications in driver monitoring.

To achieve this, state-of-the-art deep learning techniques are employed for multimodal feature extraction. Long Short-Term Memory (LSTM) networks and Transformer architectures are used to model temporal dependencies in EEG signals, capturing both short-term and long-term cognitive fluctuations associated with impaired driving states [7]. In parallel, Vision Transformers (ViT) are utilized for facial expression analysis, enabling the extraction of discriminative spatial features related to cognitive strain and drowsiness. The integration of these modalities supports a comprehensive evaluation of cognitive states such as fatigue, distraction, and cognitive overload. The proposed methodology is evaluated using large-scale datasets to ensure robust performance and generalizability [8,9]. Although the primary emphasis of this study is on model development and validation using emotion-related datasets, the findings contribute to the broader field of intelligent driver assistance systems (ADAS) by demonstrating the feasibility of multimodal cognitive state recognition for proactive safety applications [10].

2. Related Work

Previous studies on driver safety monitoring have investigated the use of EEG signals, facial expressions, and other physiological data to identify cognitive and emotional states such as fatigue, drowsiness, and distraction. Multimodal fusion approaches have consistently demonstrated improved performance compared to unimodal systems by leveraging complementary information from multiple sources [29–32]. Recent advances in deep learning, including convolutional neural networks (CNNs), Long Short-Term Memory (LSTM) networks, and Vision Transformers, have further enhanced the effectiveness of intelligent monitoring systems in recognizing cognitive states relevant to driver safety.

Wang et al. [11] explored the application of deep learning models for EEG-based emotion recognition with an emphasis on EEG and facial expression fusion, achieving accuracies of 96.63% for valence and 97.15% for arousal on the DEAP dataset. Their study also examined the impact of EEG preprocessing techniques, such as frequency filtering and normalization, on classification performance. Pan et al. [12] proposed Deep-Emotion, a multimodal emotion recognition framework that integrates EEG, facial expressions, and speech signals. By employing GhostNet and a tree-like LSTM (tLSTM) architecture for EEG feature extraction and decision-level fusion, their model achieved high classification accuracy on datasets including CK+ and MAHNOB-HCI.

Huang et al. [13] investigated decision-level fusion of EEG and facial expression features for emotion recognition using support vector machines and neural network classifiers, reporting an accuracy improvement to 82.75% over unimodal approaches. Their work emphasized the role of time–frequency and wavelet-based EEG features in enhancing classification robustness. Roshdy et al. [14] introduced a CNN-based multimodal emotion recognition system using EEG and facial expressions, where optimized electrode placement and brain heat-map representations improved feature extraction, resulting in an accuracy of 91.21%.

Bhatlawande et al. [15] presented a holistic fusion framework combining EEG, ECG, EMG, and facial expression data for emotion detection, using PCA for feature optimization and ensemble classifiers to achieve an accuracy of 87.5%. Their study highlighted challenges related to multimodal signal alignment and synchronization. Alotaibi et al. [16] proposed a 1D-CNN-based multimodal model that fused EEG and facial expression features using entropy-based canonical correlation analysis (ECCA), achieving an accuracy of 98.9% on the DEAP dataset. Collectively, these studies, along with recent driver-focused monitoring systems employing EEG, facial analysis, and transformer-based architectures [29–32], demonstrate the effectiveness of multimodal learning for recognizing cognitive and emotional states, motivating the feature-level fusion strategy adopted in this work.

3. Methodology

The suggested approach combines facial emotion recognition and EEG data into a single multimodal driver safety system. Images of facial expressions and EEG data are gathered, pre-processed, and converted into feature-rich representations that can be used with deep learning models. While facial images are standardized, encoded, and processed using sophisticated vision models, the EEG stream is subjected to noise removal, normalization, and segmentation in order to capture both temporal and frequency-domain characteristics. Long Short-Term Memory (LSTM) networks and Transformer architectures are used in conjunction for EEG feature extraction in order to capture both long-range temporal patterns and short-term dependencies. A Vision Transformer (ViT) is used to extract facial emotion features in order to detect subtle spatial cues. To improve classification accuracy for driver cognitive states such as fatigue, distraction, and cognitive load, the features extracted from both modalities are fused at the feature-level fusion using concatenation. The entire pipeline is built to function effectively, allowing for early alert generation of hazardous driving conditions [17,18].

Fig. 1 represents the architecture of the proposed model. Our multimodal driver safety system processes EEG and facial expression data independently before they are combined for analysis. The preprocessed EEG data is then converted into 3D tensors to support temporal dependencies. Facial data undergoes a similar preprocessing strategy, involving face detection, alignment, and normalization. This makes the facial features extracted comparable in effectiveness with the EEG features. It is important to have proper synchronization among these modalities in order to have consistency in the multimodal framework.

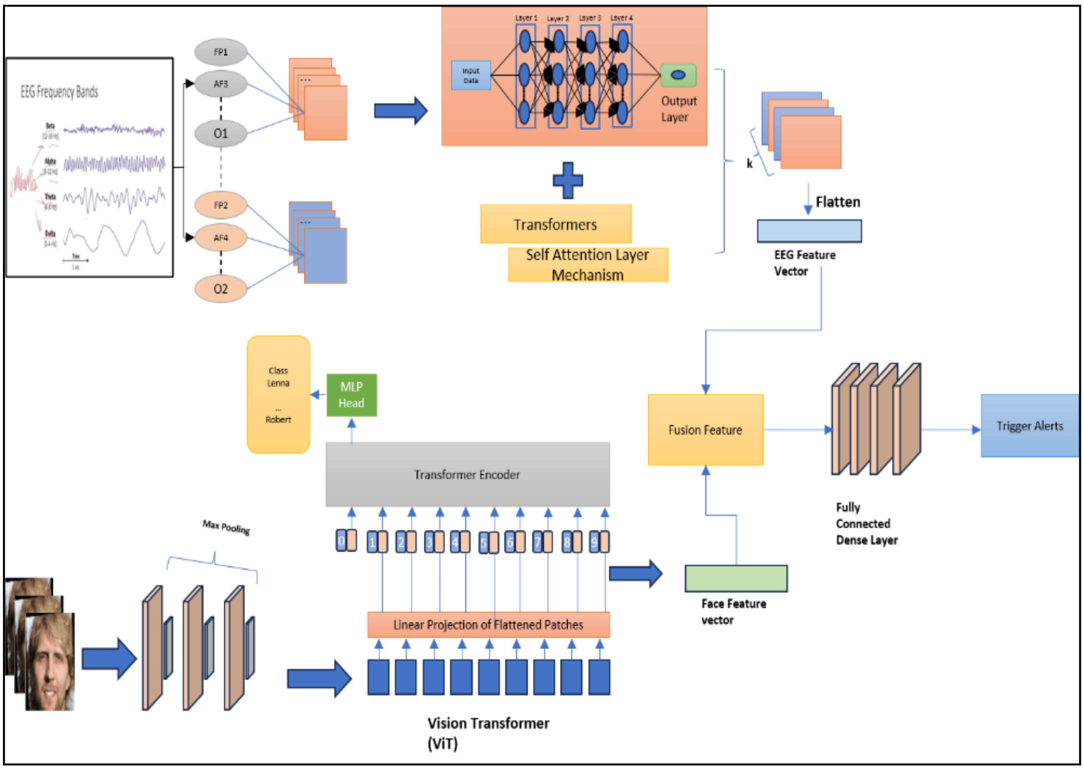

Fig. 1. Architecture of proposed model

Fig. 2 represents the LSTM layer. LSTM layer is applied to capture the local time patterns in the EEG signal. EEG data that is reshaped into sequences where each timestep represents a specific frequency band or channel of EEG is taken as input and fed to the input layer. Then the LSTM layer is built with return sequences with which the LSTM layer produces a sequence of outputs for each input step, which is essential for integration with the transformer layer. Therefore, this step provides an initial feature map of the EEG data, capturing short-term dependencies and local temporal features within the data.

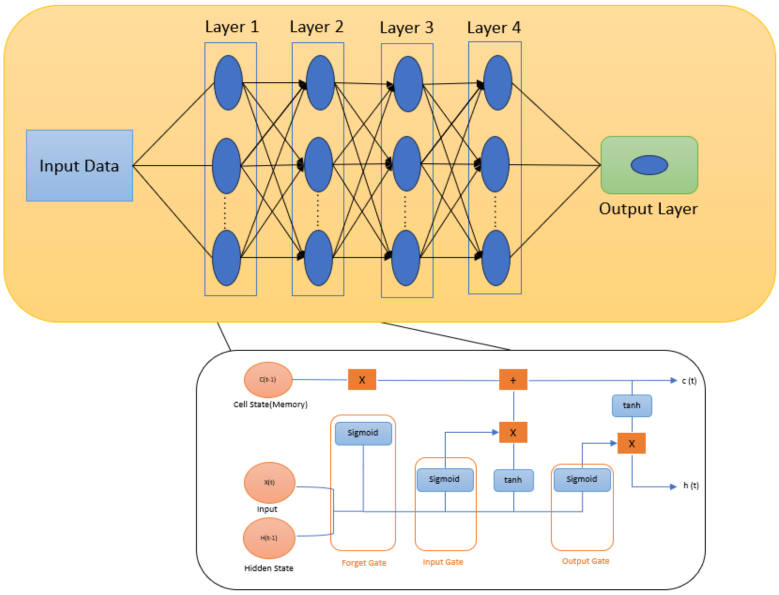

Fig. 2. LSTM Layer Architecture

3.1 Transformer Block:

The LSTM layer followed by the transformer block helps the model in improving the model's ability to detect long-term dependencies and complex patterns in EEG data. Fig.3. represents the structure of transformer block. Self-Attention Mechanism: The multi-head self-attention mechanism allows the model to be attentive to different parts of the sequence, allowing it to focus on key EEG signal features in a larger context. The sequence of outputs from the LSTM layer is taken as input for this block. The attention keys, Keys K, Queries Q, and Values V are taken from the inputs, and attention scores are calculated to determine the significance of various sequence elements. To avoid the block from attending unnecessary parts of the sequence masking is applied, then the attention scores are scaled and normalized.

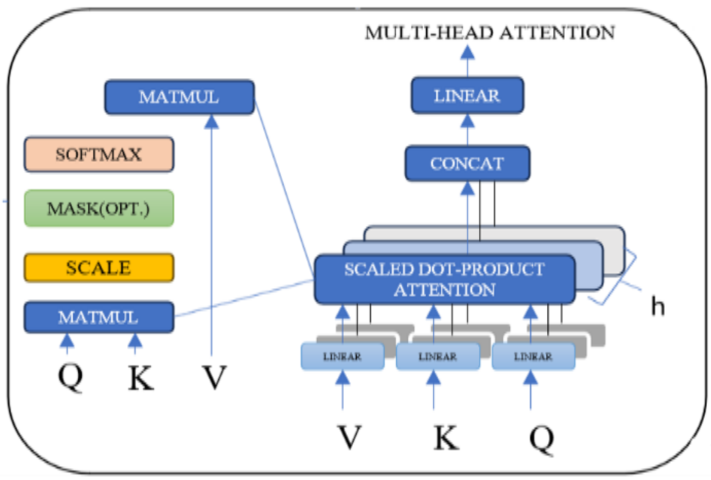

Fig. 3. Structure of Transformer block

Feed Forward Neural Network: The output of the attention process is further refined by the Feed Forward Neural Network (FFN) using a sequence of dense layers with ReLU activation. This allows the model to capture nonlinear correlations in the data, thereby increasing the likelihood of capturing complex patterns in the EEG signals associated with various emotions.

The contribution of the Feed Forward Neural Network (FFN) is important as it adjusts the output of attention, which results the model to be more responsive for the emotion classification tasks.

Residual Connections and Layer Normalization: Adding residual connections to both attention and FFN sub-layers helps in stabilizing training processes and even supports deeper architectures, enabling gradients to backpropagate smoothly, reducing the risk of vanishing gradient problem in deep networks. Moreover, applying layer normalization after each sub-layer which normalizes the output and maintains stable activations, thus aiding in efficient training. Therefore, the residual connection and layer normalization together create a much more resilient architecture

3.2 Power Spectrum Analysis of EEG Signals:

Power Spectrum analysis is one of the most established techniques for the analysis of EEG signals. The power spectrum of an electroencephalogram is a waveform which represents the distribution of signal power over frequencies. A signal is decomposed into its frequency components to obtain a power spectrum, which can be used for classifying the EEG by frequency [7].

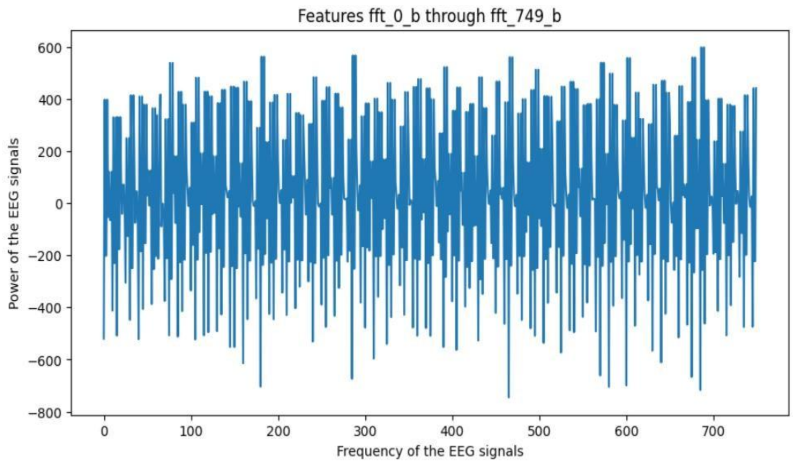

Fig. 4. Power Spectrum of EEG Signals

Fig. 4. represents the power spectrum of the EEG signal in the first trial of the dataset. The frequency of the EEG signal plotted along the x-axis and the power of the EEG signal at every frequency plotted along the y-axis. This graph depicts that the highest power of the EEG signal falls between the frequencies 4 to 8 Hz. These are the frequencies of alpha waves that are usually associated with a relaxed but attentive mind. Additionally, at frequencies ranging between 12 and 30 Hz, the graph indicates that there is a little amount of power. Beta waves are generally related to alert and attentive states. This suggests that the power spectrum of the EEG signal varies over time. The electrical activity in the brain changes due to varied stimuli and emotions.

3.3 Data Collection

This work uses the Kaggle EEG perceptual dataset(created from public Deap dataset) containing 2,132 samples, each representing EEG signals recorded from 32 subjects in different emotional states. Each subject's brain activity was recorded for 60 seconds across six emotional film clips and six minutes of neutral stimuli, totaling 36 minutes of EEG data [19]. Down-sampling to 150 Hz produced 324,000 data points, from which 2,548 statistical and frequency-domain features were extracted using a 1-second sliding window with 0.5-second overlap. Synchronized facial expression data was collected for each EEG sample, resulting in 2,132 frames labeled into seven emotions (neutral, sad, happy, angry, disgust, fear, and surprise). Preprocessing included face detection, alignment, grayscale normalization, and resizing to 224×224 pixels for Vision Transformer (ViT) compatibility, with feature vectors extracted via the [CLS] token [20]. This synchronized multimodal dataset ensures subject- and trial-level consistency, enhancing the reliability of fused EEG–facial emotion recognition models for applications such as intelligent driver support systems.

3.4 Data Preprocessing

Effective data preprocessing is essential for improving model accuracy and reliability by converting raw EEG and facial data into clean, structured formats. In this work, preprocessing techniques were applied to remove noise, handle missing values, normalize inputs, and prepare data for efficient training using LSTM, Transformers, and Vision Transformers (ViT) [21]. For EEG signals, preprocessing included handling missing values through deletion or statistical imputation, label encoding of emotion categories, one-hot encoding for multi-class classification, and reshaping data into model-compatible sequences. Feature extraction was performed in both time and frequency domains—computing statistical measures and applying FFT to extract power spectral density features across delta, theta, alpha, beta, and gamma bands. For facial data, preprocessing involved face detection, alignment, grayscale normalization, and resizing images to 224×224 pixels to match ViT requirements. Pixel values were normalized to the range [-1,1] for consistent feature distribution. Emotion labels were numerically encoded, and expressive facial features were extracted from the [CLS] token of ViT's final hidden state. Dimensionality reduction with PCA compressed these high-dimensional vectors to 128 dimensions, aligning them with EEG feature sizes for multimodal fusion.

3.5 Feature Extraction

i. For EEG

After preprocessing, EEG data is input into a deep learning pipeline in the form of an LSTM layer followed by a Transformer layer. The LSTM layer is used to extract short-term dependencies from the EEG signals, while the Transformer layer is used to further boost long-term dependencies that LSTM cannot effectively capture. To set the stage for multimodal fusion, feature extraction of the EEG data is done with a Global Average 1D Pooling layer, which summarizes the information into a significant representation.

ii. For Facial Feature Extraction via Vision Transformers

For facial expression analysis, we utilize Google's Vision Transformer (ViT), a pre-trained deep learning facial emotion recognition model [28]. The preprocessed facial information is input to the ViT model, with the last hidden state of the classification ([CLS]) token utilized as the extracted feature vector. For compatibility with EEG feature vectors, dimensionality reduction methods are applied to the extracted facial features. This ensures that both modalities have aligned feature dimensions, enabling seamless multimodal fusion.

iii. Multimodal Feature Fusion and Training

The feature vectors obtained from EEG and face data are concatenated and passed into a Multilayer Perceptron (MLP) or a fully connected layer for classification. The Adam optimizer and ReLU activation functions are employed to improve model performance.

Algorithm 1: Multimodal Fusion of EEG and Facial Features for Driver State Detection

```
Input:
        EEG signals X_eeg ∈ ℝ^(N×T×C)
        Facial images X_face ∈ ℝ^(N×H×W×3)
Output:
        Driver state y ∈ {Fatigue, Distraction, Cognitive Load}
Initialization:
        N ← number of samples
        T ← number of time frames
        C ← number of EEG channels
        K ← number of driver state classes

1: Preprocess EEG signals
        - Remove noise and missing values
        - Normalize signals
        - Segment into time windows
2: Preprocess facial images
        - Face detection and alignment
        - Resize images to 224×224
        - Normalize pixel values
3: Extract EEG temporal features:
    h_eeg ← LSTM(X_eeg)
    z_eeg ← Transformer(h_eeg)
    f_eeg ← GlobalAveragePooling(z_eeg)
4: Extract facial features:
    f_face ← ViT(X_face)
    f_face ← PCA(f_face)    // dimensional alignment
5: Temporal alignment:
    if length(f_eeg) ≠ length(f_face) then
       Align features using time index T
    end if
6: Feature-level fusion:
    f_fused ← Concatenate(f_eeg, f_face)
7: Driver state classification
    ŷ ← MLP(f_fused)
    ŷ ← Softmax(ŷ)
8: Model evaluation
    Compute Accuracy, Precision, Recall, F1-score
9: Driver state decision
    if ŷ ∈ {Fatigue, Distraction, Cognitive Load} then
       Trigger alert
    end if
10: return ŷ
end
```

Algorithm 1 represents the fusion-based driver state detection approach using EEG signals and facial features. It begins by extracting and normalizing features from both modalities, then aligns them across time. Temporal dependencies in EEG data are captured using LSTM, while facial expressions are interpreted via Vision Transformer (ViT). Feature-level fusion using concatenation combines the EEG and facial feature representations for final classification.into fatigue, distraction, or cognitive load. Alerts are triggered based on the detected driver state.
Emotion Prediction - Based on EEG and facial expressions, the model predicts emotional states using multimodal data.
Cognitive Load, Fatigue, and Distraction Estimation - EEG and facial data are then processed further to calculate cognitive load, level of fatigue, and driver distraction.

- ***Cognitive Load Estimation***

Cognitive load is estimated by examining EEG frequency bands, especially theta, beta, and alpha. The Theta/Beta Ratio (TBR) is calculated using a special formula and categorizes cognitive load into Low, Moderate, and High levels. The cut-off values for the categories are established via statistical processing of EEG data from experimentally controlled setups.
This is done as per (1):

- ***Cognitive Load Index (CLI)***

$$CLI = \frac{Mean\ Theta\ Power}{Mean\ Beta\ Power} \quad (1)$$

Thresholds:

1. Low Load: CLI < 1.5
2. Moderate Load: 1.5 ≤ CLI < 3.0
3. High Load: CLI ≥ 3.0

Previous studies have used EEG-based workload detection models with machine learning to achieve accurate cognitive load classification. This is done as per (2):

- ***Fatigue Detection***

$$Fatigue\ Score = 1 - Mean(Eye\ Aspect\ Ration(EAR)) \quad (2)$$

$$EAR = \frac{\|p2 - p6\| + \|p3 - p5\|}{2 * \ \|p1 - p4\|}$$

A threshold of 70 is used to detect high fatigue **(Fatigue Score ≥ 70).** Fatigue levels are estimated based on eye-tracking facial expression features. Two parameters are employed:

- **Eye Aspect Ratio (EAR):** Defined as: $EAR = \frac{|p2-p6|+|p3-p5|}{2*|p1-p4|}$ where p1, p2, … p6 are facial landmarks surrounding the eyes. A lower EAR indicates prolonged eye closure, a common sign of fatigue. CNN-based systems have also been used to detect drowsiness based on facial landmarks in real-time [23].
- **Blink Rate:** Measured as the number of blinks within a given time period, where higher blink rates correlate with increased fatigue levels. Combining facial features such as EAR and blink rate has shown success in fatigue detection [15]. Advanced classifiers like GRUs on facial landmarks have further improved accuracy in fatigue prediction [24].

- ***Distraction Estimation***

$$Distraction\ Score = \sqrt{}(var(yaw) + var(pitch) + var(roll)) \quad (3)$$

Distraction is assessed using head pose analysis. In (3) shows the Distraction score. Two critical parameters are used:

- **Head Pose Variability:** Calculated as: $Variance(yaw) + Variance(pitch) + Variance(roll)$ Higher variance suggests excessive head movement, which may indicate distraction.
- **Head Movement Frequency:** Defined as: $\frac{Number\ of\ Head\ Movements}{Time\ Period}$ Frequent head movements indicate reduced focus and increased distraction risk.

Head pose and facial cues have previously been used to predict unsafe driving behaviors [25]. Following the training of the model for 15 epochs, the system automatically tracks the cognitive load, fatigue, and distraction of the driver in real time. In case any of these states surpass specific thresholds, an alert is triggered to maximize road safety and avert accidents. If any of those pass the threshold then these alerts High Cognitive Load detected!, High Fatigue detected!, High Distraction! Are generated.

3.6 From Emotions to Driver States

Table 1. Represents Emotional states recognized from EEG and facial expressions are mapped to higher-level driver states based on established cognitive and behavioral research. Prolonged sadness, neutral affect, and reduced facial expressiveness are associated with driver fatigue, while increased blink rate and low arousal EEG patterns further reinforce fatigue detection. Distraction is characterized by frequent head movements, elevated surprise or fear responses, and unstable EEG beta activity. Cognitive load is inferred from increased theta and beta power ratios in EEG signals, indicating mental overload during complex driving scenarios. This hierarchical mapping enables the proposed system to translate low-level emotional cues into actionable driver safety states.

Table 1. Emotions to Driver States

Emotion Pattern	EEG Indicator	Driver State
Sad + Neutral	High Theta	Fatigue
Surprise + Fear	Beta Spikes	Distraction
Mixed Emotions	High TBR	Cognitive Load

3.7 Comparison with baselines

The Table 2. multimodal fusion model consistently outperformed both EEG-only and face-only baselines across all evaluation metrics, demonstrating that integrating physiological EEG signals with facial expression features provides complementary information for more reliable driver state detection.

Table 2. Comaprison Baselines

Model Configuration	Accuracy	Precision	Recall	F1-Score
EEG Data	92.3%	89.07%	91%	90.03%
Facial Data	93.97%	90.29%	90.89%	92.87%
EEG +Face (Multimodal Fusion)	96.38%	95.97%	95.80%	96.21%

3.8 Proposed Architecture and Hyperparameters

Fig. 5. the EEG branch processes multichannel EEG signals using LSTM and Transformer encoders to capture short- and long-term temporal patterns, followed by global average pooling to obtain EEG feature vectors. In parallel,

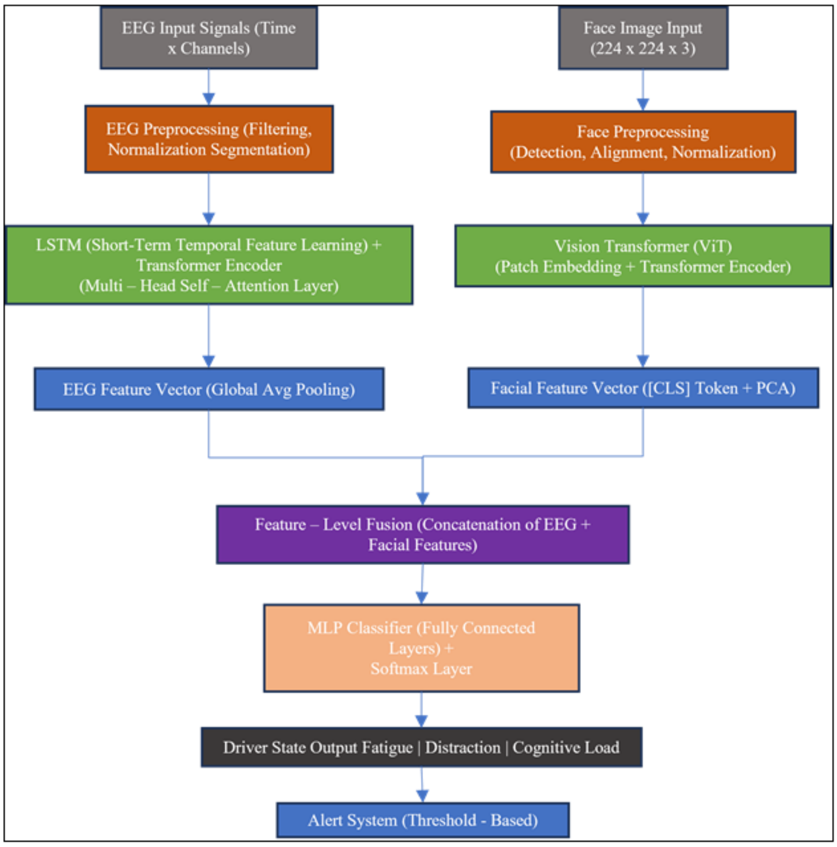

Fig. 5. Architecture Block Model

facial images are analyzed using a Vision Transformer (ViT), and discriminative facial features are extracted from the [CLS] token and reduced using PCA. The EEG and facial features are fused at the feature level through concatenation and passed to an MLP classifier with a softmax layer to predict driver states, namely fatigue, distraction, and cognitive load. A threshold-based alert mechanism is finally used to trigger safety warnings.

The Table 3. details the configuration of the LSTM and Transformer components for EEG feature modeling, the Vision Transformer (ViT) architecture for facial feature extraction, the feature-level fusion strategy, and the training settings of the MLP classifier. These parameters were selected to ensure stable training, effective multimodal feature learning, and reproducibility of the reported results.

Table 3. Hyperparameters

Component	Hyperparameter	Value
LSTM	Hidden Units	128
Transformers	Attention Heads	4
ViT	Layers	12
Training	Batch Size	32
Training	Learning Rate	0.001
Fusion	Type	Feature Concatenation
MLP Classifier	Dropout Rate	0.3
Training Configuration	Epochs	50
Training Configuration	Optimizer	Adam
Training Configuration	Loss Function	Categorical Cross-Entropy

4. Results and Discussions

The efficacy of the suggested multimodal driver safety system in identifying cognitive states like fatigue, distraction, and cognitive load was assessed using accuracy, precision, recall, and F1-score. The results are intended to demonstrate how precise and successful our methods are. The models' confusion matrices were used to calculate True positives (TP), False positives (FP), True negatives (TN), and False negatives (FN). When a model predicts a positive outcome and the actual result is also positive, the TP value rises. FP refers to a situation in which the model predicted a positive outcome but the actual result was a negative one. When the model predicts a negative outcome and the actual result is also negative, TN rises. FN refers to a situation in which the model predicted a negative outcome but the actual outcome was positive. Accuracy is a performance metric that is defined as the ratio of the number of correct predictions made by the proposed model to the total number of predictions made using it. The mathematical expression for accuracy is as follows (4):

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{4}$$

The ratio of true positives among all of the model's positive predictions is measured using the precision metric. It can be expressed mathematically as in (5):

$$Precision = \frac{TP}{(TP+FP)} \tag{5}$$

Recall is used to calculate the percentage of TPs found among all positive cases that the model has successfully identified. It can be mathematically expressed as follows (6):

$$Recall = \frac{TP}{(TP+FN)} \tag{6}$$

The F1-score, which is derived from the balanced average of recall and precision, is used to evaluate a model's performance in binary classification tasks. The F1-score can range from 0 to 1, with 1 representing the highest possible score. It can be mathematically expressed as follows (7):

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

The benefit of combining physiological and visual cues was demonstrated by the significant improvement in performance compared to single-modal approaches achieved by fusing EEG and facial emotion features. Low training and validation losses demonstrated the model's high classification accuracy and stable convergence, while the confusion matrix and classification report demonstrated balanced performance across all emotion classes [25]. The combination of LSTM, Transformer architectures, and Vision Transformers improved the system's robustness and generalizability, making it appropriate for intelligent driver assistance applications in the real world, according to a comparative analysis with traditional techniques. The developed multimodal model that combined EEG and facial features attained a test accuracy of 96.38%, proving its efficiency in emotion recognition. The model suffered from a test loss of 0.1335, which is a sign of consistent convergence and low overfitting. Performance assessment by different measures further confirmed the stability of the model, with a precision of 0.9557, recall of 0.9323, and F1-score of 0.9721

Fig. 6. shows training and validation accuracy of the model over 12 epochs. The blue line represents training accuracy which increases steadily and nearly reaches the perfect accuracy for later epochs. The Red line is for validation accuracy, which is slightly below training accuracy as the model generalizes to new, unseen data.

Fig. 7. represents the training and validation loss over 12 epochs throughout the training process. The training loss, shown by the blue line is constantly decreasing, which means it is learning from the training data. The validation loss, represented by the red line, also decreases in this similar pattern but with minor oscillations. By the last epochs, training as well as validation loss reduces to low values, which shows that it fits very well with almost no overfitting. Thus, these results will have good generalization ability as the validation loss nearly mirrors the training loss in later epochs.

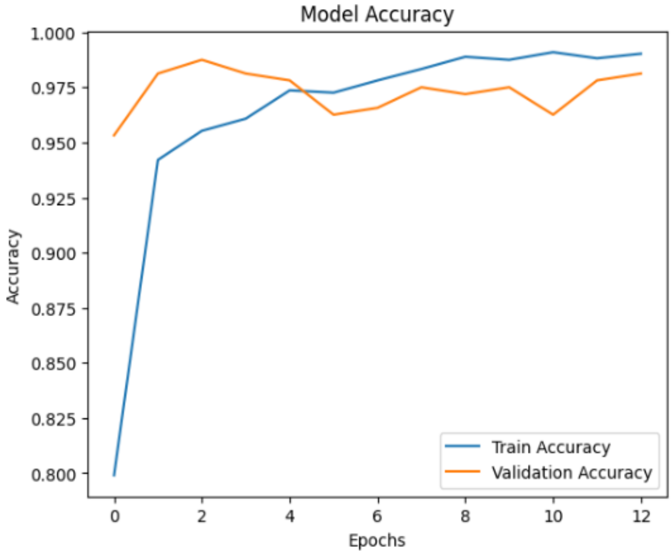

Fig. 6. Proposed model accuracy graph

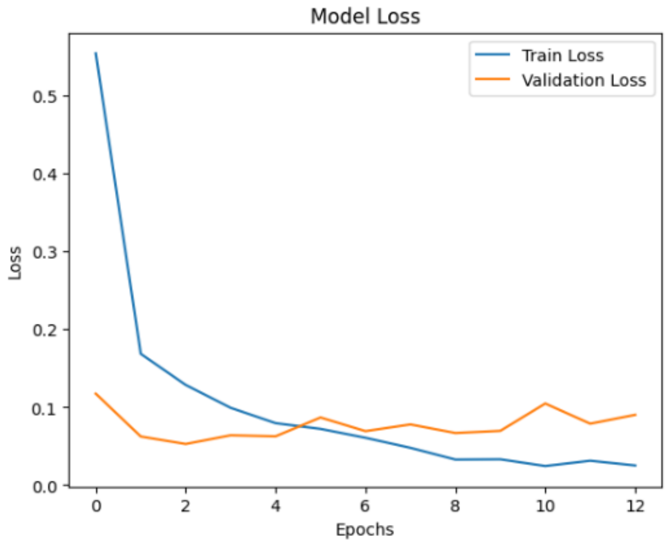

Fig. 7. Proposed model loss graph

Fig. 8. represents the confusion matrix constructed from the classification results. The model demonstrates strong classification performance across all seven emotional classes, with most predictions concentrated along the diagonal. Minimal off-diagonal entries indicate low misclassification rates, confirming the robustness and generalizability of the proposed system. This further supports the high accuracy level of 96.38% of the hybrid LSTM and multi-head self-attention transformer model in capturing short and long-range dependencies within the EEG signals. The performance is represented in the confusion matrix, emphasizing the reliability of the model in the task of emotion classification.

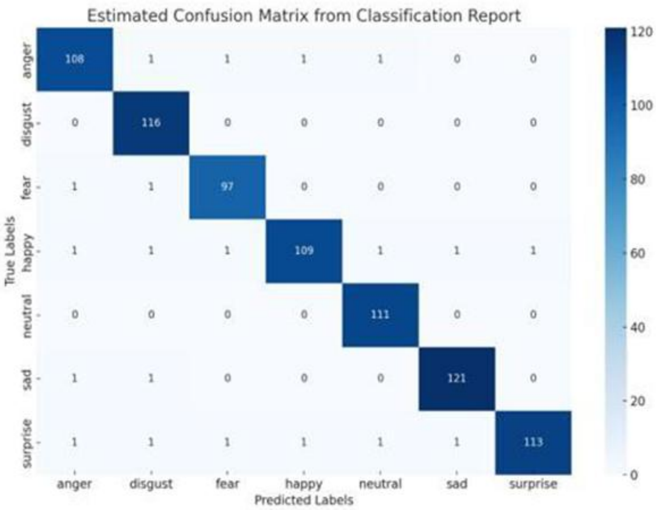

Fig. 8. Confusion Matrix

In comparison to conventional single-modal methods, the integration of EEG-based cognitive load analysis with facial emotion recognition considerably enhanced model generalization. Extraction of both short-term facial expressions and long-term EEG-based cognitive states guaranteed a more complete emotional response understanding. This integration not only increased classification accuracy but also enhanced the model's ability to identify cognitive load conditions. Other works also show that fusion of time-series EEG with facial features increases model robustness [13].

The robust performances on various evaluation measures confirm the validity of this method for driver observation, cognitive state estimation, and emotion-driven interventions. ECG and video fusion techniques have also shown promise in fatigue detection [17]. The fusion of LSTM and Transformer-based structures enhances the adaptability and strength of the system in processing EEG and facial cues simultaneously. The proposed framework is more interpretable and usable for real-world deployment, making it applicable to human-computer interaction, mental illness monitoring, and driver safety applications.

```
Precision: 0.9642
Recall: 0.9638
F1-Score: 0.9639

Classification Report:
              precision    recall  f1-score   support

       anger       1.00      0.96      0.98       112
     disgust       0.96      1.00      0.98       116
        fear       0.98      0.98      0.98        99
       happy       0.93      0.91      0.92       120
     neutral       1.00      1.00      1.00       111
         sad       0.98      0.98      0.98       123
    surprise       0.90      0.93      0.91       121

    accuracy                           0.96       802
   macro avg       0.97      0.96      0.96       802
weighted avg       0.96      0.96      0.96       802
```

Fig. 9. Classification Report

Fig. 9. represents the classification report points out the model's well-balanced performance across various emotion classes. The highest accuracy of classification was seen for neutral (99%), followed by sad (98%) and disgust (98%), whereas emotions such as happy (91%) and surprise (91%) were slightly lower, reflecting the intrinsic difficulty in differentiating these expressions. In comparison to conventional single-modal methods, the integration of EEG-based cognitive load analysis with facial emotion recognition considerably enhanced model generalization. Extraction of both short-term facial expressions and long-term EEG-based cognitive states guaranteed a more complete emotional response understanding. This integration not only increased classification accuracy but also enhanced the model's ability to identify cognitive load conditions. Other works also show that fusion of time-series EEG with facial features increases model robustness [26]. The robust performances on various evaluation measures confirm the validity of this

method for driver observation, cognitive state estimation, and emotion-driven interventions. EEG and video fusion techniques have also shown promise in fatigue detection [27].

The fusion of LSTM and Transformer-based structures enhances the adaptability and strength of the system in processing EEG and facial cues simultaneously. The proposed framework is more interpretable and usable for real-world deployment, making it applicable to human-computer interaction, mental illness monitoring, and driver safety applications. The accuracy of the suggested work was compared to other related works. Table 4. lists the task completed, the dataset used, and the accuracy attained. It can be concluded that the combination of models used in this method improved the accuracy of classification.

Table 4. Comparison of the proposed work with earlier related implementations

Related Work	Task	Dataset	Accuracy (%)
Saffaryazdi et al. [2]	Facial expressions with EEG and emotion recognition	SEED	79.51
Mutawa et al. [3]	Facial expressions with emotion recognition and EEG	DEAP	90
Hashemi et al. [5]	Driver drowsiness detection using CNN	ZJU	95.30
Shang et al. [8]	Driver emotion and Fatigue detection	FER2013	73.32
Bhatlawande et al. [18]	Multimodal emotion recognition based on EEG, ECG and EMG	CUSTOM	87.51
Our Proposed Work	Driver safety using EEG and Facial Emotions	CUSTOM	96.38

5. Conclusion and Future Work

This study presents a multimodal emotion recognition framework that integrates EEG signal processing with facial expression analysis to infer cognitive and emotional states related to fatigue, distraction, and cognitive load, with potential applications in driver monitoring. By employing a combination of Long Short-Term Memory (LSTM) networks for temporal EEG feature extraction and Vision Transformers (ViT) for facial feature representation, the proposed approach achieved an overall accuracy of 96.38% in emotion classification tasks. The results demonstrate that fusing physiological EEG signals with visual facial cues enhances feature representation and classification performance compared to unimodal methods. Although the evaluation was conducted using general-purpose emotion datasets, the findings indicate that the proposed multimodal framework provides a strong foundation for cognitive state assessment in intelligent monitoring systems. A key limitation of this study is that the experiments were performed in controlled, laboratory-based settings with limited head movement and environmental variability, which may not fully capture real-world driving conditions.

Future work will focus on extending the proposed framework toward practical driver monitoring scenarios. Incorporating additional sensing modalities, such as eye-tracking, heart rate variability (HRV), and galvanic skin response (GSR), may provide deeper insights into driver attention, fatigue, and cognitive load. Furthermore, improving generalizability through transfer learning and validation on larger, cross-population datasets or simulated driving environments will enhance robustness under real-world conditions. Optimizing computational efficiency for real-time inference and exploring edge-computing-based deployment strategies will further support the practical applicability of the framework in intelligent transportation and human-centric monitoring systems.

References

[1] Y. Alotaibi and V. A. Vuyyuru, “Electroencephalogram-based face emotion recognition using multimodal fusion and 1-D convolution neural network (ID-CNN) classifier,” *AIMS Mathematics*, vol. 8, no. 10, pp. 22984–23002, 2023, doi: 10.3934/math.20231169.

[2] N. Saffaryazdi *et al.*, “Using facial micro-expressions in combination with EEG and physiological signals for emotion recognition,” *Frontiers in Psychology*, vol. 13, p. 864047, 2022, doi: 10.3389/fpsyg.2022.864047.

[3] A. M. Mutawa and A. Hassouneh, “Multimodal real-time patient emotion recognition system using facial expressions and brain EEG signals based on machine learning and log-sync methods,” *Biomedical Signal Processing and Control*, vol. 91, p. 105942, 2024, doi: 10.1016/j.bspc.2023.105942.

[4] H. Gao, A. Yüce, and J.-P. Thiran, “Detecting emotional stress from facial expressions for driving safety,” in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Oct. 2014, pp. 5961–5965, doi: 10.1109/ICIP.2014.7026203.

[5] M. Hashemi, A. Mirrashid, and A. Beheshti Shirazi, “Driver safety development: Real-time driver drowsiness detection system based on convolutional neural network,” *SN Computer Science*, vol. 1, no. 5, p. 289, 2020, doi: 10.1007/s42979-020-00306-9.

[6] Y. Zhou *et al.*, “Cognitive workload recognition using EEG signals and machine learning: A review,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 799–818, 2021, doi: 10.1109/TCDS.2021.3090217.

[7] M. Jabon, J. Bailenson, E. Pontikakis, L. Takayama, and C. Nass, “Facial expression analysis for predicting unsafe driving behavior,” *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 84–95, 2010, doi: 10.1109/MPRV.2010.46

[8] Y. Shang *et al.*, “Driver emotion and fatigue state detection based on time series fusion,” *Electronics*, vol. 12, no. 1, p. 26, 2022, doi: 10.3390/electronics12010026.

[9] M. Dua, Shakshi, R. Singla, S. Raj, and A. Jangra, “Deep CNN models-based ensemble approach to driver drowsiness detection,” *Neural Computing and Applications*, vol. 33, pp. 3155–3168, 2021, doi: 10.1007/s00521-020-05209-7.

[10] Y. Yi, H. Zhang, W. Zhang, Y. Yuan, and C. Li, “Fatigue working detection based on facial multifeature fusion,” *IEEE Sensors Journal*, vol. 23, no. 6, pp. 5956–5961, 2023, doi: 10.1109/JSEN.2023.3239029.

[11] 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, doi: 10.1109/ACCESS.2023.3263670.

[12] J. Pan *et al.*, “Multimodal emotion recognition based on facial expressions, speech, and EEG,” *IEEE Open Journal of Engineering in Medicine and Biology*, 2023, doi: 10.1109/OJEMB.2023.3240280.

[13] Y. Huang, J. Yang, P. Liao, and J. Pan, “Fusion of facial expressions and EEG for multimodal emotion recognition,” *Computational Intelligence and Neuroscience*, vol. 2017, p. 2107451, 2017, doi: 10.1155/2017/2107451.

[14] A. Roshdy, A. Karar, S. A. Kork, T. Beyrouthy, and A. Nait-Ali, “Advancements in EEG emotion recognition: Leveraging multi-modal database integration,” *Applied Sciences*, vol. 14, no. 6, p. 2487, 2024, doi: 10.3390/app14062487.

[15] Y. Peng *et al.*, “The application of electroencephalogram in driving safety: Current status and future prospects,” *Frontiers in Psychology*, vol. 13, p. 919695, 2022, doi: 10.3389/fpsyg.2022.919695.

[16] G. S. Thirunavukkarasu, H. Abdi, and N. Mohajer, “A smart HMI for driving safety using emotion prediction of EEG signals,” in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, Oct. 2016, pp. 4148–4153, doi: 10.1109/SMC.2016.7844882.

[17] M. Yousaf *et al.*, “Enhancing driver attention and road safety through EEG-informed deep reinforcement learning and soft computing,” *Applied Soft Computing*, vol. 167, p. 112320, 2024, doi: 10.1016/j.asoc.2024.112320.

[18] S. Bhatlawande, S. Shilaskar, S. Pramanik, and S. Sole, “Multimodal emotion recognition based on the fusion of vision, EEG, ECG, and EMG signals,” *International Journal of Electrical and Computer Engineering Systems*, vol. 15, no. 1, pp. 41–58, 2024, doi: 10.32985/ijeces.15.1.5.

[19] D. Li, X. Zhang, X. Liu, Z. Ma, and B. Zhang, “Driver fatigue detection based on comprehensive facial features and gated recurrent unit,” *Journal of Real-Time Image Processing*, vol. 20, no. 2, pp. 19–29, 2023, doi: 10.1016/j.heliyon.2024.e39479.

[20] L. Zhao *et al.*, “Data-driven learning fatigue detection system: A multimodal fusion approach of ECG and video signals,” *Measurement*, vol. 201, p. 111648, 2022, doi: 10.1016/j.measurement.2022.111648.

[21] R. M. Chandra, G. S. Neelaiahgari, and S. S. Vanapalli, “Enhancing driver safety through sensor-based detection and mitigation of health risks in vehicles,” in *Proc. Int. Conf. Algorithms and Computational Theory for Engineering Applications*, Cham, Switzerland: Springer Nature, Feb. 2024, pp. 199–204, doi: 10.1007/978-3-031-72747-4_30.

[22] X. Lin, Z. Huang, W. Ma, and W. Tang, “EEG-based driver drowsiness detection based on simulated driving environment,” *Neurocomputing*, vol. 616, p. 128961, 2025, doi: 10.1016/j.neucom.2024.128961.

[23] H. Jia, Z. Xiao, and P. Ji, “End-to-end fatigue driving EEG signal detection model based on improved temporal-graph convolution network,” *Computers in Biology and Medicine*, vol. 152, p. 106431, 2023, doi: 10.1016/j.compbiomed.2022.106431.

[24] A. Topic and M. Russo, “Emotion recognition based on EEG feature maps through deep learning network,” *Engineering Science and Technology, an International Journal*, vol. 24, no. 6, pp. 1442–1454, 2021, doi: 10.1016/j.jestch.2021.03.012.

[25] T. Ergin, M. A. Ozdemir, and A. Akan, “Emotion recognition with multi-channel EEG signals using visual stimulus,” in *Proc. Medical Technologies Congress (TIPTEKNO)*, Oct. 2019, pp. 1–4, doi: 10.1109/TIPTEKNO.2019.8895242.

[26] J. Y. Kim, C. H. Jeong, M. J. Jung, J. H. Park, and D. H. Jung, “Highly reliable driving workload analysis using driver electroencephalogram (EEG) activities during driving,” *International Journal of Automotive Technology*, vol. 14, no. 6, pp. 965–970, 2013, doi: 10.1007/s12239-013-0106-z.

[27] W. Deng and R. Wu, “Real-time driver-drowsiness detection system using facial features,” *IEEE Access*, vol. 7, pp. 118727–118738, 2019, doi: 10.1109/ACCESS.2019.2936663.

[28] A. Salbi, M. A. Gadi, T. Bouganssa, A. E. Hassani, and A. Lasfar, “Design and implementation of a driving safety assistant system based on driver behavior,” *IAES International Journal of Artificial Intelligence*, vol. 13, no. 3, pp. 2603–2613, 2024, doi: 10.11591/ijai.v13.i3.pp2603-2613.

[29] P. Neeraja, R. G. Kumar, M. S. Kumar, K. K. S. Liyakat, and M. S. Vani, “DL-Based Somnolence Detection for Improved Driver Safety and Alertness Monitoring,” in Proc. IEEE Int. Conf. on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 589–594, doi: 10.1109/IC2PCT60090.2024.10486714

[30] D. Kim, H. Park, T. Kim, et al., “Real-time driver monitoring system with facial landmark-based eye closure detection and head pose recognition,” Scientific Reports, vol. 13, p. 18264, 2023, doi: 10.1038/s41598-023-44955-1.

[31] J. Chen, Y. Cui, C. Wei, K. Polat, and F. Alenezi, “Driver fatigue detection using EEG-based graph attention convolutional neural networks: An end-to-end learning approach with mutual information-driven connectivity,” Applied Soft Computing, vol. 186, Part A, p. 114097, 2026, doi: 10.1016/j.asoc.2025.114097

[32] O. F. Hassan, A. F. Ibrahim, A. Gomaa, et al., “Real-time driver drowsiness detection using transformer architectures: A novel deep learning approach,” Scientific Reports, vol. 15, p. 17493, 2025, doi: 10.1038/s41598-025-02111-x.

Authors’ Profiles

Ch. Raga Madhuri is an accomplished Assistant Professor at the Department of Computer Science and Engineering at Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, having held this position since April 2015. Alongside her teaching responsibilities, she is currently pursuing a Ph.D. degree from JNTUK, Kakinada. With over 10 years of teaching experience, she has gained extensive knowledge and expertise in Artificial Intelligence, Machine Learning, Data Analytics, cloud computing and IoT fields. She has a distinguished research record, having published 14 papers in many national and international conferences supported by IEEE and Scopus. In addition, she has published 5 research articles in SCI,Scopus and UGC Indexed journals. She also holds 3 patents in her name.

Anideep Seelam is a student researcher in the Artificial Intelligence and Data Science program at Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, Andhra Pradesh, where he is graduated in 2025. His research interests span deep learning, machine learning, Image Processing, natural language processing. He published a paper in IEEE Conference Research Paper - (ISML), 2024.

Shaik Fatima Farheen is a student researcher in the Artificial Intelligence and Data Science program at Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, Andhra Pradesh, where she is graduated in 2025. Her research interests lie in Machine Learning, Deep Learning and artificial intelligence. Her work has been published at the IEEE 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), where she presented her paper on Internet of Things.

Aadi Siva Kartheek Pamarthi is a student researcher in the Artificial Intelligence and Data Science program at Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, Andhra Pradesh, where he is graduated in 2025. His research interests span deep learning, machine learning, artificial intelligence, and natural language processing. Notably, he has worked on projects such as "Sleep Apnea Prediction Using Real-Time ECG Data," which aims to provide early detection through ECG analysis. This approach offers a cost-effective and scalable alternative to traditional Polysomnography, enabling early intervention. His research has been published at the IEEE 2024 Conference on Intelligent Systems and Machine Learning (ISML).

Krovi Mohan Kireeti is currently a Student researcher in the Artificial Intelligence and Data Science program at Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, Andhra Pradesh, where he is graduated in 2025, His research interests focus on Internet of things, Artificial Intelligence, Deep Learning and Natural Language Processing. During his academic career, he has been involved in projects such as Enhanced Mental Health Prediction through Emotion Cause Extraction in Client-Therapist using NLP which aids the Therapist to detect the mental health of the people and also another project Spy camera Detection using IR sensor which helps to detect the Hidden cameras. His work has been published at the IEEE 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), where he presented his paper on Internet of Things. He is also a member of the VRS ACM SIGKDD Student chapter and the GDSC student chapter.

How to cite this paper: Ch. Raga Madhuri, Anideep Seelam, Fatima Farheen Shaik, Aadi Siva Kartheek Pamarthi, Mohan Kireeti Krovi, "Multimodal Emotion Recognition Using EEG and Facial Expressions with Potential Applications in Driver Monitoring", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.1, pp. 163-176, 2026. DOI:10.5815/ijigsp.2026.01.10