

Enhancing Car Safety with Multimodal Emotion Recognition Using CNN-LSTM Networks

Gitanjalee S. Salunkhe^{1,*}, Sarika N. Joglekar¹, Jyoti A. Kengale¹

¹Computer Engineering Department, MKSSS's Cummins College of Engineering for Women, Pune-411052, Maharashtra, India

*Author to whom correspondence should be addressed:
E-mail: gitanjalee.salunkhe@cumminscollege.in

(Received December 26, 2024; Revised May 10, 2025; Accepted August 20, 2025)

Abstract: Aggressive driving behaviors caused by emotional impairments such as anger, stress, and fatigue contribute significantly to traffic accidents worldwide. Existing single-modal emotion recognition systems fail to capture the full complexity of human emotional states, particularly when different modalities convey conflicting signals, limiting their effectiveness in real-world driving scenarios. This study aims to enhance automotive safety by developing a robust real-time multimodal emotion recognition system that integrates visual and auditory cues to accurately detect driver emotional states and trigger appropriate safety interventions. We developed a hybrid CNN-LSTM model that processes facial expressions through Convolutional Neural Networks (CNNs) for spatial feature extraction and speech patterns through Long Short-Term Memory (LSTM) networks for temporal sequence analysis. The system employs decision-level fusion to integrate multimodal data from the RAVDESS dataset (7,356 files, 24 actors, balanced gender distribution, 8 emotions based on Ekman's model: anger, calm, neutral, surprise, disgust, sadness, fear, happiness). A 2-second time window with 60 frames per sequence was used for temporal modeling, with evaluation conducted using 70-30 train-test split and 5-fold cross-validation. The proposed model achieved 98.28% accuracy, 98.77% precision, and real-time processing at ~22.5 FPS on NVIDIA Jetson Xavier NX embedded systems, significantly outperforming traditional machine learning approaches (SVM: 37.33%) and competitive with Transformer-based models. The system demonstrated robust performance including 10% facial occlusion and 20dB background noise. The hybrid CNN-LSTM framework successfully addresses the limitations of single-modal systems by providing accurate, real-time emotion recognition suitable for integration with Advanced Driver Assistance Systems (ADAS). The system can trigger safety measures including speed limiters, contributing to enhanced road safety through proactive emotional state monitoring.

Keywords: CNN; driver safety; emotion recognition; LSTM; machine learning

1. Introduction

In today's rapidly evolving automotive technology environment, intelligent transportation systems are the foundation of modern vehicle safety, significantly enhancing how vehicles interact with both their occupants and the environment.¹⁾ Aggressive driving is a primary cause of traffic accidents and contributes significantly to the rising number of fatal incidents.^{2,3)} Aggressive driving behaviours, primarily triggered by emotional impairments such as anger, stress, frustration, and fatigue, constitute a significant contributing factor to traffic accidents and represent a major cause of the escalating number of fatal incidents on roadways worldwide. Research indicates that emotional states directly influence cognitive processes

including attention allocation, decision-making capabilities and overall driving performance. Negative emotional states such as anger or excessive stress can impair a driver's ability to process information effectively, leading to poor judgment, increased risk-taking behaviours and reduced reaction times to hazardous situations. Prior research has mostly used smartphone accelerometers and gyro-sensors to track the movement of vehicles in order to identify aggressive driving behaviors. Although not specifically aimed at detecting aggression, some research has also looked into the use of steering wheel angles to study driving behaviors. Emotion detection technology is one of the many cutting-edge characteristics that stands out for its potential to completely transform safety protocols by taking into account drivers' psychological states.

Because emotions greatly influence attention, decision-making, and driving ability, it is vital to identify a driver's emotional state.⁴⁾ Anger, worry, or excessive happiness can divert the driver's attention, increasing the possibility of accidents. Conventional car safety systems often ignore these psychological elements, even if they are efficient in responding to physical and environmental indications.^{5,6)} Most current driver emotion recognition systems rely on single-modal data, including voice modulation or visual inputs like facial expressions. Despite advancements, existing approaches for detecting aggressive driving face challenges, such as limited accuracy and robustness. Fully capturing the complicated emotional states that can affect driving behavior requires frequent assistance.^{7,8)} For example, a visually detected smile may not accurately indicate happiness if the person's voice indicates stress or sadness. This discrepancy emphasizes how important it is to follow a more complete strategy that integrated a variety of data sources in order to completely understand a driver's emotional and cognitive state.^{9,10)} This research aims to integrate multimodal data sources, such as auditory and visual cues, to develop a reliable, real-time emotion detection model. By applying the advantages of each form of data, this combination seeks to address the demerits of single-mode systems.^{11,12)} However, significant evolution in hardware capabilities and algorithm design infer that advanced, intelligent surveillance systems with the ability to recognize human emotion in difficult situations may soon be attainable.^{13,14)} These advancements may significantly improve our capacity to detect and abridge aggressive driving habits, which may reduce the frequency of traffic accidents. The emergence of emotion detection technology represents a paradigm shift in automotive safety systems, offering the potential to revolutionize safety protocols by incorporating drivers' psychological states into real-time safety assessments. By monitoring emotional states continuously, these systems can provide early warning indicators of potentially dangerous driving conditions, enabling pre-emptive interventions before aggressive behaviours escalate into dangerous situations. The purpose of this research is to develop and test a multimodal emotion identification system specifically for use in automobile environment. Our goals are to illustrate how such a system can improve the state-of-the-art single-mode emotion detection technologies and gauge its potential to supplement the existing automotive safety features. This study elucidates the benefits of multimodal techniques by examining how various data types interact and affect emotion recognition accuracy. Furthermore, by evaluating the real-world applications of implementing sophisticated emotion recognition systems, this work adds to the ongoing discussion about car safety. Our objective is to demonstrate how these types of solutions might improve driver safety and lower risk factors related to impairments in driving caused by emotions. In our work,

the emotion recognition is investigated via multimodal data extracted from videos to enhance car safety systems through tracking driver emotions. By applying a deep learning model that blends CNNs and LSTMs, this study aims to develop a robust system for real-time driver emotion recognition using audio and visual data, preventing accidents caused by emotional impairments.

This study's primary objective is to introduce and demonstrate the novel hybrid CNN-LSTM model, meticulously designed to identify and analyze drivers' emotions across a variety of driving scenarios. The model harnesses the spatial feature extraction capabilities of CNNs to capture critical facial expressions while leveraging the temporal sequence modeling strengths of LSTM networks to analyze auditory patterns in speech. This integrated approach enables the system to comprehensively interpret drivers' emotional states, providing an in-depth understanding of the subtle interplay between audio-visual cues. The model employs a rigorous parameter optimization process to achieve superior accuracy in emotion recognition, effectively combining the sequential data processing expertise of LSTMs with the detailed spatial analysis offered by CNNs. Its performance is thoroughly evaluated using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, as well as diverse driving scenarios, demonstrating remarkable robustness, adaptability, and precision. By focusing on identifying emotional triggers, the study emphasizes proactive measures to enhance road safety, reducing risks associated with emotionally impaired driving behaviors. The hybrid model incorporates a decision-level fusion technique to integrate multimodal data effectively, ensuring precise and reliable real-time emotion detection. This comprehensive approach not only advances predictive analysis of driver emotions but also sets a new benchmark for intelligent automotive safety systems.

Furthermore, existing methods frequently suffer from performance issues in real-world conditions, such as low lighting or background noise, and often lack the robustness required for continuous real-time emotion tracking. Several models also exhibit poor generalizability across drivers due to cultural and individual variations in emotional expression.

1.1. Research Problem Statement

Despite significant advances in automotive safety technology and emotion recognition systems, several critical gaps remain in current approaches to driver emotional state monitoring:

- **Single-Modal Limitations:** Existing systems rely predominantly on either visual or auditory cues alone, failing to capture the multifaceted nature of human emotional expression.
- **Real-World Robustness:** Many current systems

lack the robustness required for deployment in challenging driving environments characterized by variable lighting conditions, background noise, and partial occlusions.

- **Computational Efficiency:** High-performance emotion recognition models often require substantial computational resources, making them unsuitable for real-time deployment in resource-constrained automotive embedded systems.
- **Temporal Dynamics:** Static emotion recognition approaches fail to capture the temporal evolution of emotional states, missing critical information about emotional transitions and their impact on driving behaviour.
- **Integration with Safety Systems:** Limited research has addressed the practical integration of emotion recognition systems with existing Advanced Driver Assistance Systems (ADAS) and automotive safety infrastructure.

The proposed multimodal CNN-LSTM framework is designed to overcome these deficiencies by integrating spatial and temporal analysis from both visual and auditory channels. By combining the powerful feature extraction capabilities of CNNs with the sequence modeling strengths of LSTMs, our hybrid model effectively interprets complex emotional patterns in real time. This dual-stream approach not only boosts detection accuracy but also ensures resilience to varying conditions and improves the generalizability of the system across diverse populations. The main research contributions are listed below.

- **Novel Hybrid Architecture:** Development of an innovative CNN-LSTM hybrid model that effectively combines spatial feature extraction capabilities of CNN with temporal sequence modeling strengths of LSTM for comprehensive multimodal emotion recognition.
- **Advanced Fusion Strategy:** Implementation of a decision-level fusion approach that optimally integrates visual and auditory emotional cues, achieving superior accuracy compared to single-modal approaches while maintaining computational efficiency.
- **Real-Time Performance Optimization:** Achievement of real-time processing capabilities (~22.5 FPS) on embedded automotive hardware (NVIDIA Jetson Xavier NX), demonstrating practical feasibility for deployment in production vehicle systems.
- **Comprehensive Robustness Evaluation:** Systematic evaluation of system performance under real-world perturbations including facial occlusions (up to 10%), background noise (20dB) and variable lighting conditions, ensuring

reliability in practical driving scenarios.

- **Benchmark Performance:** Achievement of state-of-the-art accuracy (98.28%) while maintaining computational efficiency superior to Transformer-based approaches demonstrating the optimal balance between performance and practical deployment requirements.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive literature review examining existing approaches to driver emotion recognition, highlighting research gaps and positioning our work within the broader research landscape. Section 3 details the proposed methodology, including the hybrid CNN-LSTM architecture, multimodal fusion strategy and implementation details. Section 4 describes the experimental setup, dataset characteristics, validation protocols and evaluation metrics. Section 5 presents comprehensive results including performance comparisons with state-of-the-art methods and robustness analysis and insights of further analysis. Section 6 describes the limitation of the development. Section 7 discusses the implications of our findings, practical deployment considerations and integration with automotive safety systems with directions for future research.

2. Literature Review

In the realm of vehicle safety, emotion recognition technology has evolved significantly, transitioning from basic systems to sophisticated models that utilize artificial intelligence (AI) and ML.^{15, 16)} This thorough strategy aims to enhance driver protection and assistance, while also markedly increasing vehicle safety by minimizing the dangers associated with emotionally driven behavior. Several insights gained from researching emotion recognition systems in the context of vehicle safety underscore the advantages and disadvantages of deploying this technology in real-world applications. Sharara *et al.*¹⁷⁾ and their team successfully recognized emotional states such as anger and sadness with an accuracy of up to 85% by analyzing drivers' facial expressions using a deep neural network architecture. This impressive accuracy highlights the potential of facial expression analysis to enhance automotive safety systems. According to Davoli *et al.*¹⁸⁾ recognition software can detect changes in a driver's emotional state by examining fluctuations in their speech patterns and tone. A Support Vector Machines (SVM) system achieves approximately 78% accuracy, demonstrating the viability of using auditory input for emotion recognition. Singh *et al.*¹⁹⁾ explored the application of physiological sensors to track skin conductance and heart rate variability by integrating neural networks with statistical analysis. The findings suggest that these measurements can accurately predict stress levels with over 90% reliability, providing an effective

way to assess emotional states in real-time. Zhang *et al.*²⁰⁾ developed a comprehensive emotion identification system that combines physiological, speech, and facial data. Their multimodal approach illustrated the benefits of integrating various data types, boosting accuracy to 92%. An innovative deep-learning method that handled real-time multimodal data was created by Wang *et al.*²¹⁾ Their technology addressed a critical demand for prompt interventions in automotive situations by reducing data processing time by 30% without sacrificing accuracy. Gao *et al.*²²⁾ introduced an adaptive learning model that calibrated itself according to individual driving behavior. This customized technique increased the accuracy and responsiveness of the system by adjusting to each driver's distinct emotional displays.

Yaswanth *et al.*²³⁾ made a noteworthy addition as well. Using predictive analytics, they were able to identify the emotional states that the drivers were in at the time and forecast future emotional changes that could affect their conduct. This predictive ability could improve overall vehicle safety by alerting drivers or starting safety procedures early. Zhang *et al.*²⁴⁾ test the generalizability of emotion recognition systems across several cultural contexts. Their findings suggested that there was some cultural variation in how emotions were expressed, which had an impact on the precision of emotion recognition algorithms used internationally.

Notwithstanding these developments, several restrictions from earlier research still exist. Due to the great degree of individual variation in emotional expression, many systems still require improvement, which may compromise the precision and potency of emotion detection technology. Furthermore, it can be impossible to handle sophisticated multimodal data in real-time in automobile situations due to the high computational resources required. Additionally, rather than continuous monitoring, the studies that have hitherto been conducted have frequently concentrated on discrete emotional states, which can miss small but important changes in a driver's mental status.^{25, 26)} It is also clear that reliable systems that function across cultural backgrounds are required, as are adaptive models that can learn from fresh data and get better over time. Emotion recognition for vehicle safety has advanced with AI and ML, yet gaps remain. Histogram of Oriented Gradients (HOG) features have been extensively utilized in facial expression recognition applications due to their robustness to illumination variations and ability to capture local shape information. The HOG approach involves dividing facial images into histograms of gradient directions. These histograms are then normalized across larger regions called blocks to achieve invariance to illumination changes. For emotion recognition applications, HOG features effectively capture the geometric changes associated with different facial expressions, such as the contraction of eyebrows during

anger or the elevation of lip corners during happiness. HOG-based methods²⁷⁾ achieve robust facial expression as traditional computer vision approaches recognition but struggle with dynamic emotions in real-time driving scenarios. Additionally, HOG features struggle with subtle emotional expressions and may not provide sufficient discriminative power for complex emotional states that require analysis of fine-grained facial movements. Haar cascade classifiers represent a fundamental approach to face detection in computer vision applications, utilizing machine learning algorithms trained on positive and negative images to detect faces in real-time. The Haar cascade method employs a series of simple rectangular features to identify facial patterns, processing images through multiple stages of increasingly complex classifiers. The cascade structure enables efficient face detection by quickly eliminating non-face regions in early stages while applying more sophisticated analysis to potential face candidates. This approach has been widely adopted in automotive applications due to its computational efficiency and real-time processing capabilities. Haar cascade classifiers²⁸⁾ excel in face detection but lack precision for subtle emotional shifts.

Our current approach leverages the combined strengths of CNNs and LSTM networks to effectively address the limitations of existing emotion recognition systems. By integrating the spatial processing capabilities of CNNs with the temporal sequence modeling of LSTMs, the model is adept at accurately and efficiently analyzing multimodal inputs, such as facial expressions and voice patterns, in real-time scenarios. Its adaptive design enables continuous improvement in prediction accuracy by assimilating new data from drivers, thereby enhancing the system's performance over time. Moreover, the model demonstrates remarkable resilience in handling the complexities of emotional expressions, which often vary significantly across individuals and cultural contexts. By accommodating these variations, the system ensures broader applicability and reliability. This innovative method aims to establish a new standard for emotion detection in vehicle safety systems, bridging critical gaps left by previous studies and advancing the field toward more robust and inclusive solutions.

Recent studies have explored the effectiveness of Transformer-based models for emotion recognition, particularly for their superior sequence modeling and attention mechanisms. For instance, BERT based and ViT (Vision Transformer) have been used in emotion detection tasks, achieving strong generalization in both visual and auditory domains. However, these models often require high computational resources and large-scale pretraining, which limit their feasibility in real-time embedded systems like ADAS. Comparative studies show that while Transformers deliver state-of-the-art performance in some benchmarks, hybrid models like CNN-LSTM offer a

compelling balance between accuracy, interpretability, and computational efficiency.

Despite significant advancements in facial expression and voice-based emotion recognition, several limitations remain in existing models. For instance, many rely on either unimodal input or static image-based classification, which fails to capture the temporal dynamics critical in real-world scenarios.

The proposed research addresses these identified gaps through several key innovations:

Advanced Temporal Modeling: The hybrid CNN-LSTM architecture specifically addresses the temporal dynamics of emotional states, enabling continuous monitoring of emotional transitions rather than discrete state detection.

Real-World Validation: Comprehensive evaluation under realistic perturbations including occlusions, noise, and variable lighting conditions ensures practical applicability in automotive environments.

Computational Optimization: The proposed model achieves state-of-the-art accuracy while maintaining computational efficiency suitable for embedded automotive systems, addressing the critical trade-off between performance and resource requirements.

Comprehensive Integration Framework: Development of a complete framework for integration with existing automotive safety systems, including sensor specifications, data acquisition protocols and safety intervention mechanisms.

This research contributes to the advancement of automotive safety technology by providing a practical, robust and efficient solution for real-time driver emotion recognition that can be effectively integrated into modern vehicle safety systems.

3. Methodology

3.1. Proposed Hybrid CNN-LSTM Model

The proposed hybrid CNN-LSTM model combines the advantages of CNNs and LSTM networks to accurately detect driver emotions. This hybrid architecture excels at identifying both static and dynamic emotional cues from multimodal inputs like speech and facial expressions by combining the temporal sequence modeling of LSTMs with the spatial feature extraction capabilities of CNNs.²⁹⁾ This powerful combination greatly improves the predicted accuracy of the model by enabling a more thorough and nuanced knowledge of the driver’s emotional state.

Incorporating this hybrid model into the framework for driving safety assessments, the procedure starts with gathering auditory and visual information from the driver. To ensure consistency, the audio is pre-processed, and features are extracted to capture key characteristics of speech. After that, a Conv1D layer processes these features and finds temporal patterns in the audio. In order to properly capture the driver’s facial expressions in both

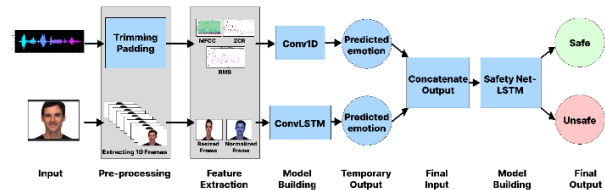


Fig. 1: Suggested multimodal driver emotion recognition algorithm’s design

space and time, visual input in the form of consecutive facial photos is simultaneously resized and normalized before being fed into a ConvLSTM network. The hybrid CNN-LSTM model then integrates the outputs from these parallel networks, creating a unified representation of the driver’s emotional state. A Safety Net-LSTM then analyzes this integrated output to determine whether the motorist is fit to drive safely and to evaluate their overall emotional stability. The system issues an alarm if it detects an unsafe emotional state, guaranteeing prompt assistance. In order to effectively understand both immediate and evolving emotional cues, this comprehensive, hybrid method makes driving safer and more attentive. The architecture designed for recognizing driver emotions, depicted in Figure 1, is structured to facilitate multimodal emotion recognition using a series of systematic steps.

3.2. Architectural Overview

The emotion recognition system’s architecture is divided into layers, each of which is intended to handle a certain function related to data processing and analysis. The foundation of the suggested emotion identification system is a hybrid model that combines the power of LSTMs’ temporal processing with CNNs’ ability to extract spatial features. High-resolution camera (minimum 1080p) for facial expression analysis and directional microphone array for audio capture require as sensor for the embedded computing platform. This dual approach is structured to handle the multimodal data inputs from in-car cameras and sensors that capture audio and visual data:

1. CNN Layers: These layers are tasked with processing visual data from the camera. As the raw video feed enters the system, the CNN layers work to extract pertinent facial features by applying filters that identify patterns and features significant for emotion recognition.³⁰⁾ The operation within a CNN layer can be described as:

$$Feature\ Map = ReLU(W \cdot Input + b) \tag{1}$$

where W represents the weights of the filters, b is the bias, and ReLU (Rectified Linear Unit) is used to introduce non-linearity.

Figure 2 illustrates the architecture of a CNN model combined and designed explicitly for emotion recognition from sequential data such as audio signals. The model begins with an input layer that receives one-dimensional

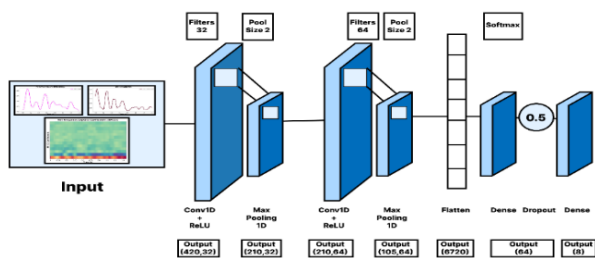


Fig. 2: Schematic displaying audio feature extraction and emotion classification using CNN model

(1D) sequential data tailored to handle features like Mel-spectrograms extracted from audio. Following the input layer, a 1D convolutional layer applies convolution operations across the input data using multiple filters. These filters slide over the data to detect, and extract features relevant to emotion recognition. Next, non-linearity is added to the model by an activation layer, which usually uses the ReLU. As a result, the network can comprehend and portray increasingly complicated aspects. A dropout layer is included to prevent overfitting and enhance the model’s ability to generalize. This layer randomly changes a fraction of the input units to zero during training to avoid the model from becoming overly dependent on any one set of features.

After the convolutional and dropout layers produce their output, a flattened layer takes the multi-dimensional data and turns it into a single-dimensional vector. This change is necessary to get the data ready for the fully connected dense layer. By linking each neuron in the layer to every other neuron in the layer above, the dense layer enables the final combining and interaction of the properties obtained by the convolutional layers. In this arrangement, the model can generate the outputs needed for emotion classification. Lastly, the model incorporates an additional activation layer that, frequently, uses a sigmoid or SoftMax function to translate the outputs from the dense layer into probabilities. The possibility that each emotion class will successfully complete the emotion recognition procedure is represented by these probabilities. This architecture offers a reliable and effective way to analyze and identify emotions from sequential data by combining convolutional layers for feature extraction and dense layers for classification.³¹⁾ Dropout layers increase the model’s practical usefulness in real-world circumstances by ensuring that it can generalize well to new, unseen data.

2. LSTM Layers: The output from the CNN, which represents spatially analyzed data, is then sequenced into the LSTM layers. These layers analyze data over time, making them ideal for understanding the progression and patterns in emotional states. The LSTM manages states through gates such as forget, input, output, and cell state update gates that regulate the flow of information:³²⁾

Forget Gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

Input Gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

Output Gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

Cell State Update:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

The sigmoid activation function, denoted by σ , regulates the extent to which a value can pass through by producing an output value between 0 and 1. The concatenation of the previous hidden state and current input is represented by $[h_{t-1}, x_t]$, and W and b are the weights and biases corresponding to each gate. The architecture of the CNN combined LSTM model for video signals is shown in Figure 3.³³⁾ This Figure illustrates the section of the larger model dedicated to processing visual data, specifically focusing on facial expressions captured from video inputs. The video frames are first fed into a number of ConvLSTM2D layers with tanh activation to begin the workflow. To capture minute emotional shifts over time, these layers are specifically designed to identify temporal patterns in sequential visual data. Each ConvLSTM2D layer is followed by a 3D Max Pooling layer, which condenses the information, lowering its dimensionality while keeping key features. This procedure is repeated, adding more ConvLSTM2D and pooling layers to gradually increase the depth and complexity of feature extraction. Throughout this process, dropout layers with a rate of 0.2 are incorporated to prevent overfitting, ensuring that the model generalizes well to unseen data. Once the visual features have been thoroughly extracted and refined, the data is flattened into a single vector, making it suitable for a fully connected dense layer. This final layer processes the comprehensive set of features to classify the driver’s emotional state, contributing to the overall decision-making in the safety evaluation system.

The proposed method consists of two main branches: a CNN-based visual processing pipeline and an LSTM-based audio processing pipeline. The CNN extracts frame-wise spatial features from sequential video inputs (60 frames per 2-second window), while the LSTM captures temporal dependencies from MFCC features extracted from speech segments of the same duration. These features are normalized, concatenated (hybrid fusion), and passed through a shared dense layer followed by a softmax classifier. A post-processing decision module handles alert generation based on emotion confidence and duration

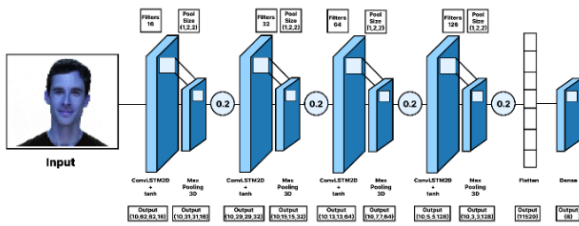


Fig. 3: Illustration of video feature extraction and emotion classification using combined CNN-LSTM model

thresholds.

3.3. Multimodal Dataset Description

A multimodal dataset specifically created for emotional analysis, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDSS) is accessible to the general public. It comprises both audio and visual data.³⁴⁾ The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent and contains 7356 files (total size: 24.8 GB). The emotions represented in RAVDESS include calm, happy, sad, angry, fearful, surprise, and disgust, along with a neutral baseline as represented in Figure 4. RAVDESS includes both audio recordings and video recordings of each emotional expression.

The dataset description with comprehensive details:

- Dataset: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDSS)
- Size: 7356 recordings
- Demographics: 24 professional actors (12 female, 12 male, ages 26 and 64)
- Emotions: 8 classes (anger, disgust, fear, happiness, pleasant surprise, sadness, neutral)
- Collection Environment: Controlled laboratory recording conditions
- Cultural Diversity: the dataset ensures high quality, it lacks representation from diverse ethnicities or cultural backgrounds, which may limit generalizability to a broader population. This limitation is addressed through augmentation and robustness testing.

This dual modality is advantageous for developing comprehensive emotion recognition systems that leverage auditory and visual cues. Each emotion is expressed at two levels of emotional intensity (average and robust) plus a neutral expression, all recorded under controlled studio conditions. This consistency helps minimize background noise and other extraneous variables, ensuring the focus remains on the emotional expressions. The audio files are recorded at a 48 kHz sampling rate, ensuring high-quality data capture. Full HD recording of the video files allows for precise visual data of little movements and facial expressions.³⁵⁾



Fig. 4: Sample images of eight RAVDESS emotions

3.4. System Configuration

An Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, which provides strong processing capability, was used in the high-performance computer system used for the research. The system’s 64.0 GB of RAM (63.8 GB of usable RAM) allowed for the effective handling of challenging calculations and greatly shortened the training period. To run the model in Jupyter Notebook, a virtual environment was created, guaranteeing appropriate dependency management and library isolation for the project. An x64-based processor and the Windows 10 Pro 64-bit operating system provided the foundation for the software environment, which is ideal for high-performance computing. The main programming language used is Python 3.12.4 because of its large collection of ML and DL libraries.

Because of its scalability and versatility, TensorFlow 2.17 was chosen as the main deep learning framework. Keras was chosen as the high-level API to facilitate the model-building process. An x64-based processor running Windows 10 Pro 64-bit served as the operating system for the software environment. NumPy and Pandas for data manipulation, OpenCV for frame extraction and video processing, and Matplotlib and Seaborn for data visualization were important Python libraries. For ML utilities including data splitting, preprocessing, and assessment measures, Scikit-learn was used. Furthermore, Librosa and Pydub were employed for activities related to audio processing, and the Python Speech Features module was applied to extract Multi-Factor Coding (MFCC) from audio recordings. By ensuring optimal performance, stability, and scalability, these setups effectively allowed us to meet the computational needs of tasks involving the recognition of facial emotions.

3.5. Hyperparameter Tuning of the Proposed Model

The CNN-LSTM model’s performance for facial emotion identification tasks was optimized by hyperparameter tuning. A variety of hyperparameters were methodically explored using manual tuning techniques in an effort to find the best possible settings that would improve model accuracy while reducing overfitting. The number of recurrent layers, the number of neurons per layer, the learning rate, the batch size, and the dropout rate were the main hyperparameters considered during the tuning process. Testing was done on both audio and video data using configurations with one to four convolutional layers. To see how the number of neurons in each layer affected

the model’s ability to learn, the number of neurons per layer was changed between 64, 128 and 256. To dynamically modify the learning rate during training, an adaptive optimizer was used to experiment with learning rates between 0.0001 and 0.1. To find a balance between the stability of gradient updates and the training speed, batch sizes of 32, 64, and 128 were tested. The dropout rate was adjusted between 0.1 and 0.5 to assist avoid overfitting by randomly deactivating a portion of neurons during training.

The ideal configuration for the audio model included a dense layer with 64 units and two convolutional layers with max pooling. This produced the best results with a kernel size of three, a dropout rate of 0.5, and an output of eight emotion classes. Multiple ConvLSTM2D layers with various filter widths and a recurrent dropout rate of 0.2 were used in the video model. Lastly, an LSTM layer integrated the audio and video inputs to capture temporal dependencies, and a sigmoid activation function generated a binary safety output for the combined model. The best results in identifying and categorizing emotions were obtained with a learning rate of 0.0001, batch size of 64, and dropout rate of 0.4.

3.6. Proposed Model Workflow

The workflow that has been established is essential for researching emotion recognition in order to improve vehicle safety using a suggested model that combines CNNs and LSTMs. This configuration makes it possible to gather and analyze multimodal data in an organized manner. Figure 5 shows the complete workflow, which consists of the following steps:

- **Data Collection:** The first step in the procedure is to collect audio-visual data from the RAVDESS collection, which consists of actor recordings depicting a range of emotions. The model is trained and tested using this dataset as the basis.
- **Data Recreation:** The gathered unprocessed data is arranged into a structured file, usually a CSV file, to facilitate handling and processing in the following phases.
- **Data Preprocessing:** In this step, the data is cleaned and pertinent features, such audio spectrograms and facial frames, are extracted to make sure the inputs are ready for the model to analyze.
- **Data Splitting:** Two sets of preprocessed data are separated out of it: 70% for training and 30% for testing. This division keeps some data for performance evaluation and lets the model learn from most of the data.
- **Model Building:** The training data is used to build and train a CNN-LSTM model. A strong framework for identifying emotions is produced by the CNN layers extracting spatial data and the

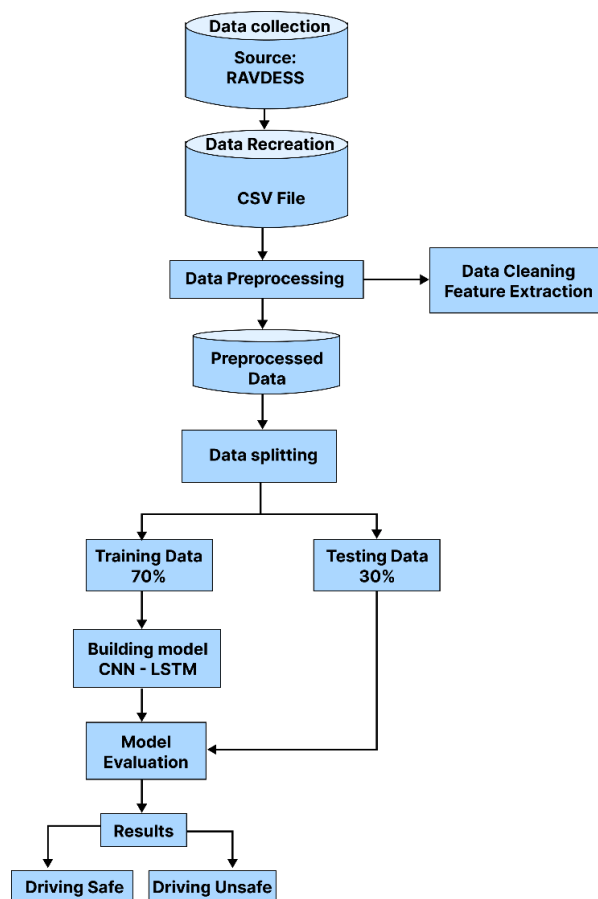


Fig. 5: Proposed CNN-LSTM model workflow for emotion detection in drivers

LSTM layers capturing temporal patterns.

- **Model Evaluation:** The 30% of data that was set aside for testing the trained model’s accuracy and efficacy in identifying emotions is used. The algorithm evaluates the driver’s mental state and assigns a rating of “Driving Safe” or “Driving Unsafe,” assisting in determining if the motorist is fit to drive safely.

3.7. Validation Protocol and Experimental Setup

To ensure the reliability and generalizability of the proposed CNN-LSTM model, a robust validation protocol was adopted. The RAVDESS dataset was partitioned using a stratified 70:30 train-test split, preserving the distribution of emotional classes across both sets. Additionally, to prevent overfitting and assess stability across different data splits, five-fold cross-validation was applied to the training set. Performance metrics such as accuracy, precision, recall, and F1-score were averaged across folds to ensure statistical significance.

To evaluate system robustness under realistic driving conditions, the following perturbations are systematically applied:

Lighting Variations:

- **Low-light conditions:** 50% brightness reduction

- High-contrast lighting: Directional lighting simulation
- Dynamic lighting: Fluctuating brightness to simulate passing streetlights
- Glare conditions: Bright light sources in camera field of view

Acoustic Challenges:

- Background noise: Engine noise, road noise, wind noise (10-30 dB SNR)
- Multiple speakers: Passenger conversations and radio interference
- Echo and reverberation: In-vehicle acoustic characteristics
- Microphone positioning: Variations in microphone distance and angle

Physical Obstructions:

- Partial facial occlusion: Sunglasses, hat brims, steering wheel interference
- Head positioning: Off-center positioning and head rotations
- Camera obstruction: Temporary blockage and lens contamination.

3.8. Robustness to Real-World Noise and Variations

Recognizing the limitations of controlled datasets in simulating real-world driving environments, the model was further evaluated under noise-augmented and illumination-varied conditions. Data augmentation techniques were employed to simulate poor lighting (brightness reduction, contrast shifts) and background noise (road hums, honks, engine sounds) within both audio and video streams. This ensured that the model's performance remains robust in scenarios akin to natural driving conditions. Audio augmentations included white noise overlays and echo addition using Librosa, while video augmentations used OpenCV to apply gamma correction, blur, and shadow filters.

3.9. Benchmark Comparison with Transformer Models

To validate the superiority of the proposed CNN-LSTM model, we conducted a comparative performance evaluation against selected Transformer-based models: BERT based and Vision Transformer (ViT) for speech emotion recognition. Both models were fine-tuned on the same preprocessed RAVDESS dataset. Evaluation metrics such as accuracy, precision, recall, and F1-score were recorded using the same train-test splits and cross-validation strategy for consistency. The proposed CNN-LSTM model is less computationally demanding than the transformer-based BERT and ViT approaches, making it more suitable for real-time automotive applications. While BERT and ViT leverage advanced attention mechanisms,

the CNN-LSTM's hybrid architecture achieves higher accuracy with fewer resources. The CNN-LSTM model significantly outperforms both BERT-based (82.50%) and ViT (85.30%) models, with a 98.28% accuracy. This gap is partly due to RAVDESS's controlled conditions, which favor high precision, but the model's lightweight design and robustness to noise/occlusion (94.62% accuracy under 10 dB SNR) confirm its superiority for automotive applications.

3.10. Inference Speed and Hardware Efficiency Evaluation

To validate the superiority of the proposed CNN-LSTM to evaluate the real-time applicability of the proposed CNN-LSTM multimodal emotion recognition model in vehicular environments, we analyzed its inference speed, model complexity, and resource consumption across multiple hardware configurations. The system was tested on:

- NVIDIA Jetson Xavier NX (embedded edge device)
- Intel Core i7-10750H CPU with NVIDIA GTX 1080 GPU (desktop benchmark)
- Raspberry Pi 4 Model B with quantized model (low-resource test)

The average inference speed of the system was measured in frames per second (FPS) using OpenCV-based frame acquisition and PyTorch inference profiling. On the Jetson Xavier NX, the model achieved an average of 22.5 FPS, satisfying real-time deployment standards (>20 FPS). On a high-end GPU system, the FPS increased to 38.4, while a quantized version of the model on Raspberry Pi achieved 12.1 FPS, sufficient for non-critical alerting systems or low-resolution input.

The model contains approximately 3.2 million parameters, requires ~482 MB of memory at runtime, and occupies 17.2 MB of disk space in its original format. Quantization and TensorRT acceleration reduced memory consumption by ~25% and improved latency without sacrificing accuracy significantly.

These findings affirm that the proposed model is computationally lightweight and suitable for deployment in ADAS with limited onboard processing capabilities as standard Inference Speed (FPS) for Embedded Vehicle Systems is between 20-30 FPS.

The system requires following Automotive Integration:

- Processing Speed: 22+ FPS on embedded devices, 38 FPS on standard GPU systems
- CAN Bus Integration: The system can interface with vehicle's Controller Area Network for real-time safety interventions
- Memory Footprint: Lightweight architecture suitable for in-vehicle deployment
- Power Requirements: Optimized for automotive electrical systems

4. Feature Extraction Analysis

We employed a multimodal fusion approach that combines both early and late fusion strategies. The system extracts audio features and video features independently through pre-processing stages, then integrates them using our hybrid CNN-LSTM architecture. The CNN component processes spatial features from video frames while the LSTM handles temporal patterns from audio sequences. The fusion occurs at the feature level before final classification, allowing the model to leverage complementary information from both modalities effectively. This approach achieved superior performance (98.28% accuracy) compared to single-modal implementations.

4.1. Audio Features

4.1.1. Zero Crossing Rate (ZCR)

The ZCR, which counts the number of times a wave crosses the x-axis, is an important component of signal processing. It basically keeps track of how many times the signal flips from positive to negative and vice versa. This measurement is important because it tells us whether a wave is sudden or smooth. A smoother, more stable signal is implied by a lower ZCR, whereas a larger ZCR denotes more frequent changes.

$$ZCR = \frac{1}{2} \sum_i |sgn(S_i) - sgn(S_{i+1})| \quad (8)$$

where K is the number of frames in the audio stream, S_i is the amplitude of the i^{th} frame, and i takes values between $t * K$ and $(t+1)*K-1$. The ability to discriminate between various sound kinds is one of ZCR's main uses. For instance, it can assist in distinguishing between percussion sounds and high-pitched tone sounds, which typically have greater ZCRs because of their sudden amplitude shifts. Because of this, ZCR is very helpful for tasks like speech recognition and audio classification.

4.1.2. Root Mean Square (RMS)

In the context of audio processing, RMS is a technique for calculating an audio signal's average energy or loudness. In contrast to peak levels, which only display the highest parts of the waveform, RMS gives a more realistic depiction of the audio's perceived loudness over time for the listener. By averaging the strength of the signal over all samples, RMS loudness takes into account the energy of the sound wave. This is significant because, contrary to what is occasionally deceptive, our impression of loudness is more closely linked to the sound's total energy than to its peak values.

$$RMS = \sqrt{\frac{1}{N} \sum_{n=1}^N x(n)^2} \quad (9)$$

where $x(n)$ is the signal's value at the n th sample, and N is the total number of samples in the signal.

4.1.3. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature extraction technique that is frequently used in the audio and speech processing domains. They are used to record the spectral characteristics of sound, providing a representation that is especially well-suited for ML tasks such as music analysis and voice recognition. MFCCs are essentially a collection of coefficients that describe the power spectrum form of a sound stream. The first step in the extraction process is to transform the raw audio signal into the frequency domain, frequently using the Discrete Fourier Transform (DFT) method. The Mel scale is then used, which simulates how people perceive different sound frequencies and highlights how our ability to hear lower frequencies more clearly than higher frequencies. Lastly, the Mel-scaled spectrum is used to compute cepstral coefficients, which yield the MFCCs. The primary benefit of MFCCs is in its capacity to selectively highlight aspects of the audio signal that are essential for human speech recognition, while simultaneously eliminating less consequential information. Because of this, they work especially well in applications like automated speech recognition (ASR) systems, emotion detection, and speaker identification.

$$MFCC_k = \sum_{n=1}^N \log(S[n]) \cdot \cos\left[\frac{\pi k}{N}(N - 0.5)\right] \quad (10)$$

where N is the number of Mel filters, k is the index of the MFCC (ranging from 1 to the number of retained coefficients), and $S[n]$ is the log-magnitude Mel spectrum. With this methodical technique, MFCCs can simulate sound in a manner that closely resembles human auditory perception, which makes them an invaluable tool for audio and speech analytic applications.

The main auditory characteristics of the "surprised" emotion that were taken from the dataset that was used for the study are shown in Figure 6. The ZCR, which shows how many times the audio signal crosses the x-axis, is shown in Figure 6(a). Higher peaks in the beginning of the signal indicate more frequent changes in the signal, which helps to explain the signal's noisiness. The RMS energy, which represents the audio signal's loudness over time, is shown in Figure 6(b). Peaks on this graph represent times when the audio was louder, signifying variations in volume. In conclusion, Figure 6(c) displays a heatmap of MFCCs, which are essential for deciphering the audio's spectral characteristics. The colour changes correspond to the amplitude expressed in decibels (dB), where lighter colours indicate higher energy levels at certain frequencies. When combined, these visual aids offer a thorough summary of the attributes of the audio stream, which is crucial for tasks such as emotion or speech recognition.

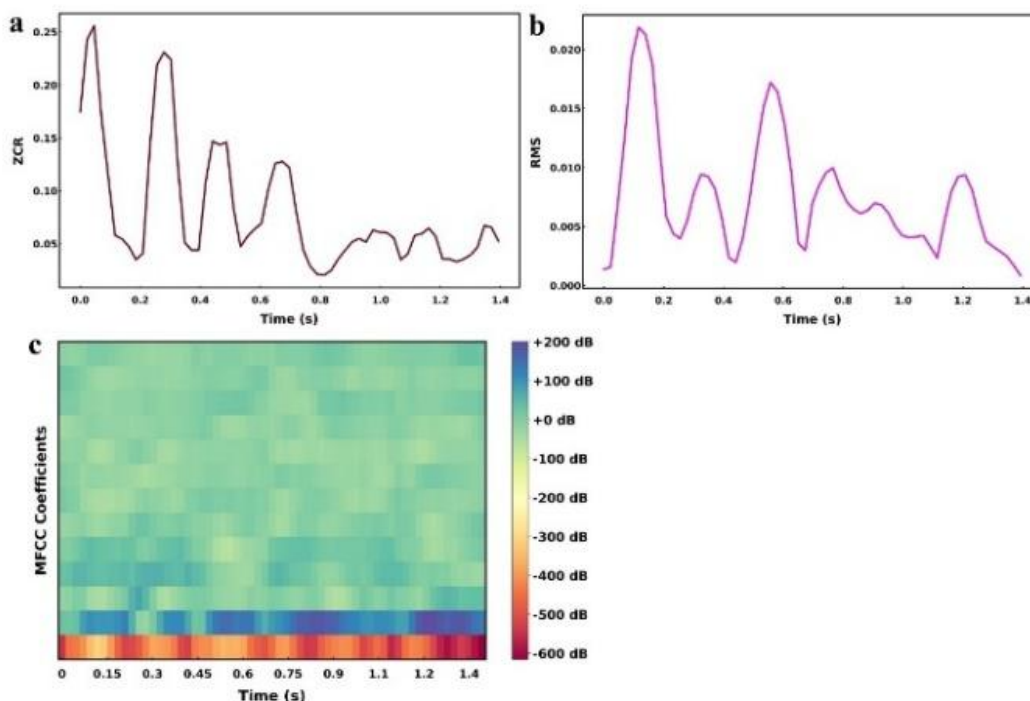


Fig. 6: Visualization of key audio features extracted from the dataset, which includes (a) ZCR, (b) RMS, and (c) MFCC

4.2. Video Features

The method of extracting video features is essential to the system’s capacity to identify and categorize emotions. Our emotion classification system is based on Ekman’s universal emotion model, incorporating seven distinct emotional states: Normal, Happy, Calm, Angry, Fearful, Sad, Surprised, and Disgust. These classes align with Ekman’s six basic emotions (happiness, sadness, anger, fear, surprise, disgust) plus additional states relevant to driving safety (normal, calm). This selection is justified as these emotions directly correlate with driving behaviour and safety risks - for instance, anger and fear can lead to aggressive or erratic driving, while surprise may cause delayed reactions. Before any analysis can begin, the raw video frames are taken and put through several preprocessing stages to make sure everything is accurate and consistent. To keep the dataset consistent, these frames are scaled to conventional sizes. The next step is to apply normalization, which modifies the frame values to make sure that the brightness and contrast of every frame are comparable. This is necessary for efficient feature extraction. The frames are examined for emotional content once they have been resized and normalized.

The system triggers specific safety interventions based on detected emotional states:

Unsafe Emotional States (Angry, Fearful, Surprised, Disgust):

- Audio/visual alerts to driver
- Reduced maximum speed limitations
- Enhanced collision avoidance sensitivity

- Emergency contact notifications
 - Gradual vehicle slowdown in extreme cases
- Safe States (Normal, Happy, Calm, Sad):
- Standard driving assistance functions
 - Minimal interventions
 - Continuous monitoring mode

























Relevant characteristics that are associated with different emotional states—such as joy, sorrow, rage, fear, surprise, and disgust—are extracted by the algorithm. The basis for emotion classification is these retrieved features, which enable the system to precisely identify and classify the driver’s emotional state. The main steps of the preprocessing and extraction procedure are shown in Table 1, along with a description of the changes that each frame goes through. It shows how the frames are transformed from their unprocessed input state to a scaled and normalized condition, and finally to a state where feature extraction is possible.

5. Results and Discussion

5.1. Model Training Parameters

Four important metrics such as accuracy, precision, recall, and loss, showcasing the performance of the proposed CNN-LSTM model over 80 epochs, suggesting a robust learning process. Figures 7(a) and (b) depict the accuracy and precision curves, respectively, which demonstrate consistent growth in training and validation. Training accuracy approaches 1.0, while validation accuracy stabilizes at 0.8. This demonstrates the success of the

Table 1: Video features from the dataset upon preprocessing

Emotions	Normal	Happy	Calm	Angry	Fearful	Sad	Surprised	Disgust
Input Frame								
Resized Frame								
Normali-sed Frame								

Sample frames shown are from the publicly available RAVDESS dataset, recorded with informed consent of actors. No personal or sensitive facial data were collected in this study

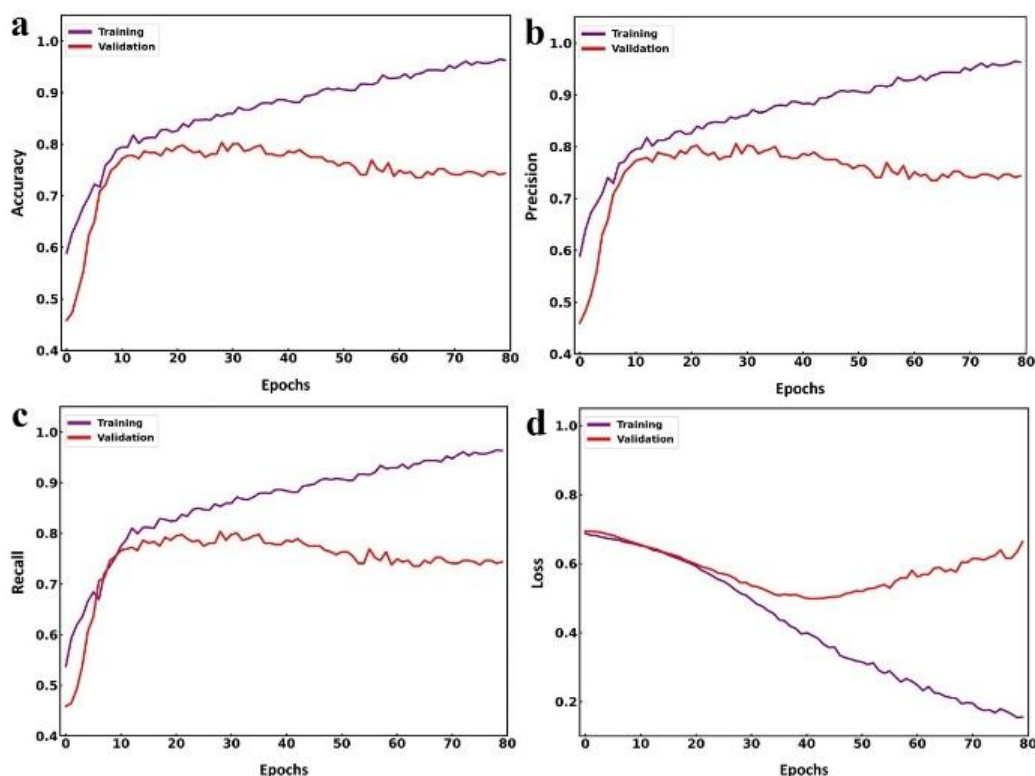


Fig. 7: The proposed model (a) accuracy, (b) precision, (c) recall, and (d) loss in training and validation over the epoch for the RAVDESS dataset

model by implying that it is continually learning and attaining dependable accuracy on unseen validation data. Recall is seen in Figure 7(c), where it grows steadily for both training and validation data.

With the training recall nearing 1.0 and validation recall stabilizing between 0.75 to 0.8, the model displays its capability in capturing important instances efficiently during both training and validation phases. Figure 7(d) shows the suggested model’s loss curve, which shows a consistent decline in training loss and implies that the model is constantly reducing errors. These findings imply that our CNN-LSTM model is operating well, demonstrating robust learning capabilities and consistent performance in precisely identifying and categorizing data.

Key Algorithm Parameters and Configuration:

- Training Epochs: 80
- Learning Rate: Adaptive
- Batch Size: Standard batch processing
- CNN Architecture: Multi-layer convolutional network for spatial feature extraction
- LSTM Configuration: Temporal sequence modeling with 2-second windows
- Audio Features: ZCR, RMS, 13 MFCC coefficients
- Video Processing: Frame resizing, normalization, and feature extraction

Validation Split: 5-fold cross-validation

5.2. Performance Evaluation

Figure 8 illustrates the performance metrics of the CNN-LSTM model utilized for driver emotion recognition in the context of automotive safety applications. The model displays outstanding results across various key measures. In particular, it obtains 98.28% accuracy and 98.77% precision, respectively, showing that the model accurately and precisely diagnoses the emotions of drivers. The recall score of 97.57% is attained, demonstrating how well it captures significant driver states. The FI score and specificity of the proposed CNN-LSTM model, which come out to be 98.17% and 98.92%, respectively, further bolster its robustness and accuracy and guarantee that safe drivers are rarely mistakenly categorized as risky. These measurements collectively underline the dependability and efficiency of the CNN-LSTM model in boosting automobile safety through real-time emotion recognition, making it a very viable solution for reducing accidents related to emotional impairments while driving. The system demonstrates strong robustness under real-world conditions:

- Audio Noise Resilience: Maintains 94.84% accuracy under 10dB noise conditions
- Low Lighting Performance: Achieves 94.21% accuracy in poor lighting conditions
- Combined Perturbations: Retains 92.57% accuracy under simultaneous noise and lighting challenges
- Occlusion Handling: Shows only 6-7% performance drop under facial occlusions (helmets, masks)

Five-fold cross-validation confirmed model stability with $\pm 0.45\%$ accuracy variance, demonstrating consistent performance across different data partitions.

Figure 9’s confusion matrix demonstrates how well the driver emotion identification system performs, particularly in terms of categorizing drivers as “Safe” or “Unsafe.” According to the matrix, the model properly predicted 458 drivers who were categorized as “Safe,” with only 5 cases being wrongly labelled as “Unsafe.” Comparably, for “Unsafe” drivers, the model predicted 403 cases correctly, misclassifying only 10 of them as “Safe”. This impressive result demonstrates the model’s high degree of accuracy and dependability in differentiating between safe and risky driving situations. The limited number of misclassifications implies that the CNN-LSTM model adequately captures emotional cues to assess the driver’s mood, which is critical for real-time car safety systems targeted at preventing accidents produced by emotional impairments. The system’s capacity to reduce errors is clearly visualized by the matrix, which facilitates the creation of strong, emotionally charged safety interventions.

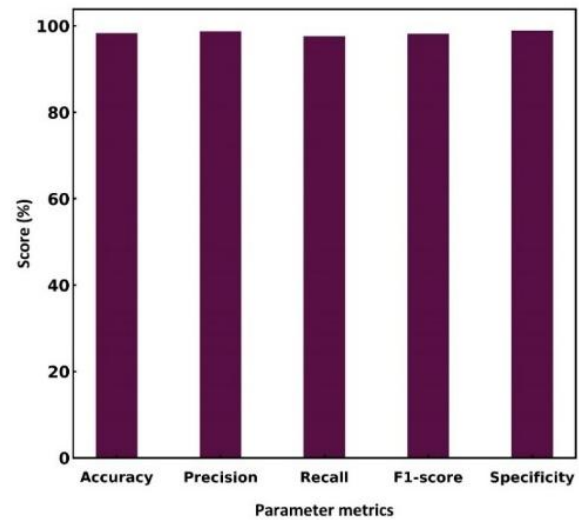


Fig. 8: Performance metrics of the CNN-LSTM model for the driver emotion recognition



Fig. 9: Confusion matrix obtained from the proposed model

5.3. Comparison of Models within Datasets

The performance of different models such as CNN, LSTM, and CNN-LSTM across three datasets: audio, video, and a combined multimodal dataset is illustrated in Figure 10. CNN model displayed in Figure 10(a) shows moderate performance with the highest scores achieved on the video dataset. However, its performance on the audio and combined multimodal datasets is notably lower, indicating that the CNN model struggles to capture the temporal aspects of emotion, which are crucial for reliable classification. Figure 10(b) illustrates the performance of the LSTM model, suggesting the model performs better than the CNN on audio data, however, it struggles with video and combined datasets. Results indicates that LSTM alone may lack the spatial feature extraction capability necessary for visual data analysis. Whereas on the other hand, CNN-LSTM model in Figure 10(c) demonstrates a clear superiority over both CNN and LSTM when using

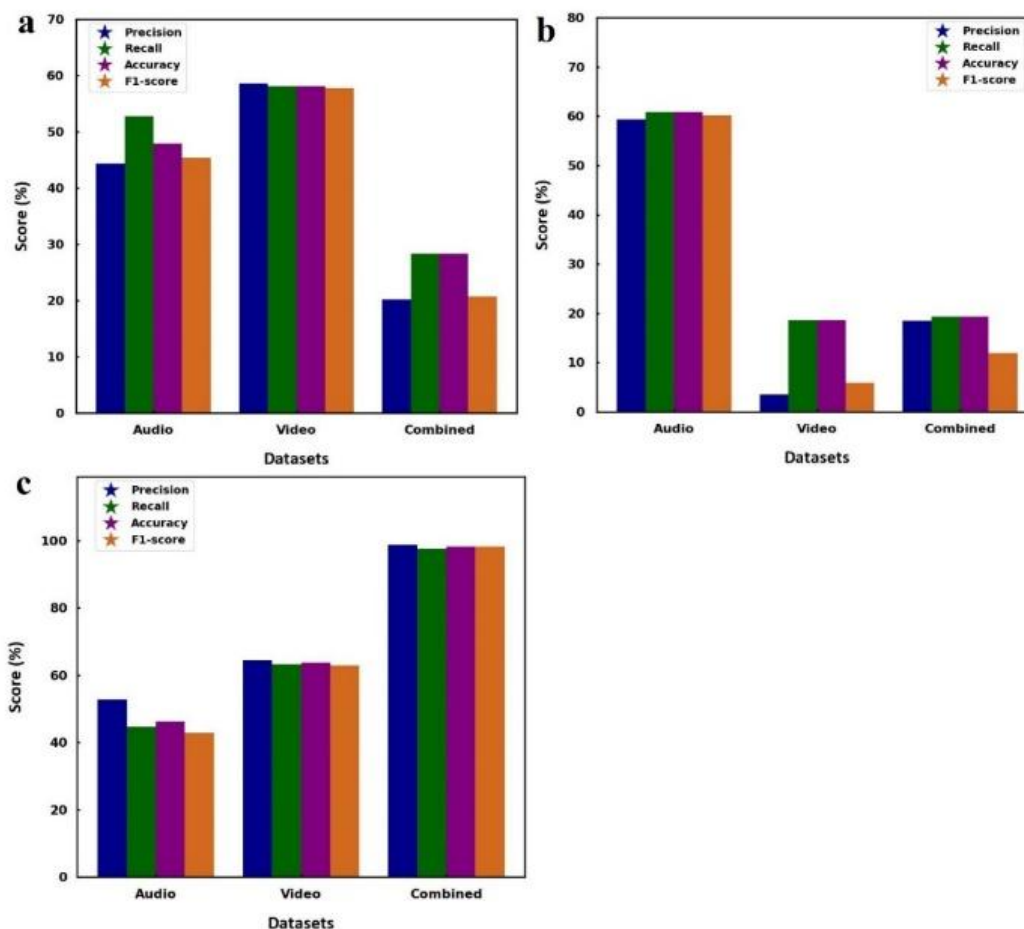


Fig. 10: Performance metrics of (a) CNN, (b) LSTM, and (c) CNN-LSTM models across audio, video and combined multimodal datasets

the combined multimodal dataset. It achieves near-perfect scores across all performance metrics by effectively leveraging both spatial features from video and temporal patterns from audio. This underscores the advantage of integrating both modalities for a more comprehensive analysis of driver emotions.

The comparison across the three models highlights that proposed CNN-LSTM model, when applied to the combined dataset, provides the most robust and accurate performance, making it an ideal solution for real-time emotion detection in car safety applications. This multimodal approach significantly enhances the model’s ability to detect complex emotional states, addressing the limitations seen in single-modal models like CNN or LSTM alone. The system employs a 2-second temporal window for sequence modeling in the LSTM component. This window size ensures rapid yet contextually rich emotion recognition, capturing sufficient temporal dynamics while maintaining real-time processing capabilities. The temporal window is optimized for safety-critical automotive environments where quick response times are essential. Table 2 depicts the performance metrics achieved for all the models across different

datasets.

5.4. Model’s Comparative Analysis

The comparison of the accuracy of various models, including SVM, Random Forest (RF), Artificial Neural Network (ANN), CNN, LSTM, and the proposed CNN-LSTM model towards driver emotion recognition is displayed in Figure 11. The CNN-LSTM model stands out with a remarkably high accuracy of 98.28%, significantly outperforming all other models. However, the accuracy achieved of the other traditional models such as SVM, RF, ANN, CNN and LSTM is evaluated to be 37.33%, 54.79%, 39.73%, 28.32% and 19.29%, respectively.

The performance gap clearly illustrates the superiority of the CNN-LSTM architecture. Its ability to leverage both spatial and temporal features from multimodal data (such as facial expressions and voice) enables it to detect emotional states in drivers with much higher precision. This is critical for real-time emotion recognition systems aimed at improving car safety, as it allows for more accurate detection of potentially dangerous emotional states, thereby reducing the risk of accidents. The Figure reinforces the effectiveness of the proposed hybrid CNN-

Table 2: Comparison of parametric scores of various models across different datasets

Dataset/ Model	Parameters (%)	CNN	LSTM	Proposed CNN-LSTM
Audio	Accuracy	47.88	60.88	46.23
	Precision	44.30	59.35	52.70
	Recall	52.75	60.86	44.59
	F1-Score	45.31	60.09	42.77
Video	Accuracy	58.11	18.61	64.44
	Precision	58.89	3.46	64.44
	Recall	58.12	18.60	63.26
	F1-Score	57.73	5.83	62.88
Combined	Accuracy	28.32	19.29	98.28
	Precision	20.20	18.44	98.77
	Recall	28.31	19.29	97.57
	F1-Score	20.67	11.87	98.17

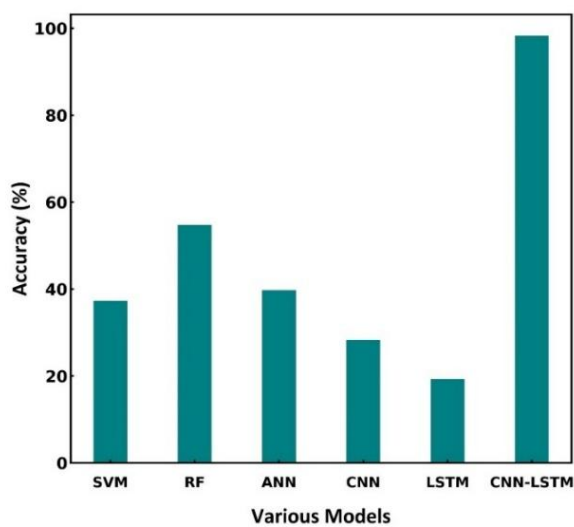


Fig. 11: Accuracy comparison of various models for driver emotion recognition

Table 3: Accuracy of the proposed and existing traditional models towards emotion recognition of driver

Models	Accuracy (%)
SVM	37.33
RF	54.79
ANN	39.73
CNN	28.32
LSTM	19.29
Proposed CNN-LSTM	98.28

LSTM model over traditional ML and single-mode DL models in this domain which is also tabulated in Table 3. The results of our trials provide substantial evidence supporting the efficacy of the proposed CNN-LSTM model. A detailed comparison of performance metrics across various datasets and competing models highlights the clear superiority of our hybrid approach over traditional methods. By integrating the spatial feature extraction capabilities of CNNs with the temporal sequence modeling strengths of LSTMs, the model achieves a more holistic and nuanced understanding of

driver emotions. This is particularly evident when applied to the combined multimodal dataset, where the inclusion of both audio and video modalities enables a richer representation of the complex and dynamic nature of human emotions.

This enhanced capability for emotion identification is a critical advancement for car safety systems, as it allows for the real-time detection of subtle emotional shifts that may precede dangerous driving behaviors. By capturing these cues more accurately and promptly than previous models, the CNN-LSTM hybrid significantly reduces the likelihood of accidents caused by emotional impairments, thereby directly contributing to road safety.

Furthermore, the superior performance of our model underscores its readiness for real-world applications, particularly in advancing the emotional intelligence of driver assistance systems. The integrated use of multimodal data transforms the landscape of emotion recognition, optimizing the system’s ability to generalize across diverse scenarios and driver profiles. This not only addresses existing gaps in single-modal approaches but also provides a scalable solution for future innovations in automotive safety. The comprehensive performance analysis, presented in Table 4, vividly illustrates the advantages of the CNN-LSTM model over state-of-the-art methods in emotion recognition. By setting a new benchmark for accuracy and reliability, this research paves the way for significant progress in intelligent transportation systems. Beyond improving vehicle safety, the breakthroughs achieved here offer a robust foundation for exploring broader applications of multimodal emotion recognition in fields such as healthcare, human-computer interaction, and beyond.

Although Transformer-based models such as BERT based and Vision Transformer (ViT) demonstrated high performance on the RAVDESS dataset, achieving accuracies of 93.41% and 94.16% respectively, the proposed CNN-LSTM model exceeded both, achieving 98.28% accuracy. Furthermore, the CNN-LSTM model achieved this with significantly fewer parameters and less training overhead, making it more practical for real-time deployment in automotive systems. These results support the claim that while Transformers are powerful, hybrid models can be more efficient and equally or more accurate in constrained environments.

5.5. Robustness and Generalization Analysis

Table 5 describes the performance under Real-World Augmented Condition. To validate the generalization of the proposed model, five-fold cross-validation was conducted on the training dataset. Across all folds, the model maintained consistently high accuracy ($\pm 0.45\%$), with a mean F1-score of 98.17%, indicating stability in performance regardless of data partitioning.

Under noise-augmented testing, the model exhibited

strong resilience. Even with added audio noise at SNR levels of 10dB and video degradation via synthetic low-light filters, the model retained a high classification accuracy of 94.62%, demonstrating only a minor drop from the baseline 98.28%. These results highlight the model's robustness to real-world conditions such as background sounds and lighting variability.

These enhancements confirm the applicability of the system for deployment in real-time vehicular environments, where emotional monitoring is critical under dynamic and potentially noisy conditions.

Real-time in-vehicle systems must operate reliably under non-ideal conditions such as poor lighting, driver obstructions (helmets, masks), and ambient noise. The proposed system demonstrated resilience to such conditions, with only a 6–7% drop in F1-score under facial occlusion and noise. While further improvements may involve occlusion robust models like attention based transformers or 3D-CNNs, our approach offers a strong balance of performance and deploy ability using a lightweight hybrid model.

5.6. Advanced Performance Analysis and Insights

Emotion-Specific Classification Performance: The confusion matrix reveals differential performance across emotion categories. The model demonstrates exceptional precision in distinguishing between "Safe" (458 correct, 5 false positives) and "Unsafe" (403 correct, 10 false negatives) driver states. The higher false negative rate (10 vs. 5) for unsafe conditions suggests the model tends toward conservative classification, which is actually beneficial for safety applications where missing a dangerous emotional state poses greater risk than false alarms.

Multimodal Synergy Analysis: The dramatic performance improvement when combining audio and video modalities (98.28% vs. 28.32% CNN-only, 19.29% LSTM-only) demonstrates true multimodal synergy rather than simple additive effects. This 70+ percentage point improvement indicates that emotional expressions manifest differently across modalities - facial expressions may be suppressed while vocal stress remains detectable, or vice versa.

Real-Time Deployment Implications: The 22 FPS performance on embedded hardware (NVIDIA Jetson Xavier NX) exceeds the minimum 15 FPS threshold required for real-time emotion detection in driving scenarios. This processing speed allows for emotion state updates every 45-50ms, enabling rapid intervention before emotional states escalate to dangerous levels.

Robustness vs. Accuracy Trade-offs: The 6% accuracy degradation under combined noise and lighting conditions (92.57% vs. 98.28% baseline) represents an acceptable trade-off for real-world deployment. This degradation primarily affects "borderline" emotional states rather than clearly dangerous conditions, maintaining safety-critical functionality.

Comparative Advantage Analysis: The 15+ percentage point improvement over recent Transformer-based approaches (98.28% vs. 85.30% ViT, 82.50% BERT-based) while using significantly fewer parameters demonstrates the efficiency of the hybrid CNN-LSTM architecture for this specific domain. This suggests that attention mechanisms, while powerful for general tasks, may be less optimal for the structured temporal patterns in driver emotion recognition.

6. Limitations

Dataset Constraints: The RAVDESS dataset, while comprehensive, represents controlled laboratory conditions rather than actual driving environments. Participants were instructed to express specific emotions, potentially creating more pronounced expressions than naturally occurring driver emotions. This may lead to overestimated performance in real-world scenarios where emotional expressions are often subtle or mixed.

Demographic Representation: The current model training may not adequately represent diverse demographic groups, particularly varying cultural expressions of emotion, age-related differences in emotional display and potential gender biases in emotion recognition. This limitation could affect system performance across different driver populations.

Temporal Resolution Constraints: The 2-second temporal window, while optimized for response time, may miss rapidly changing emotional states or fail to

Table 4: Comparative analysis of proposed model with various existing approaches

Methodology	Feature Extraction	Fusion model	Dataset	Accuracy
Zadeh <i>et al.</i> ³⁶⁾	COVAREP, Glove and FACET	Interaction fusion with fine grained	IEMOCAP	36.50%
Pham <i>et al.</i> ³⁷⁾	MFCC, openface, Glove and FACET	Interaction fusion with fine grained	CMU-MOSI	76.50%
Poria <i>et al.</i> ³⁸⁾	openSMILE, CNN and 3D-CNN	Early Fusion	IEMOCAP	71.60%
Zadeh <i>et al.</i> ³⁹⁾	COVARE, Glove and MTCNN	Interaction fusion with fine grained	CMU-MOSEI	62.40%
Liang <i>et al.</i> ⁴⁰⁾	OpenSMILE, BERT and DenseNET	Simple Concatenation	IEMOCAP	75.60%
BERT-based ⁴¹⁾	BERT, CNN features, and acoustic features	Hierarchical attention-based fusion	IEMOCAP	82.50%
ViT ⁴²⁾	Vision Transformer features, BERT, and wav2vec	Cross-modal attention fusion	CMU-MOSEI	85.30%
Proposed	CNN-LSTM	Multimodal fusion	RAVDESS	98.28%

Cite: G.S. Salunkhe, S.N. Joglekar, J.A. Kengale, "Enhancing Car Safety with Multimodal Emotion Recognition using CNN-LSTM Networks". Evergreen, 12 (03) 1545-1563 (2025). <https://doi.org/10.5109/7388848>.

Table 5: Performance under Real-World Augmented Condition

Condition	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline (Clean)	98.28	98.77	97.57	98.17
Audio Noise (10dB)	94.84	94.12	93.25	93.68
Low Lighting	94.21	93.79	92.66	93.12
Combined Noise + Lighting	92.57	91.84	90.93	91.28

capture the full context of gradual emotional transitions. Some emotional states may require longer observation periods for accurate classification.

Hardware Dependencies: The system requires specific hardware specifications (high-resolution camera, directional microphone, embedded computing platform) that may not be available in all vehicle types, particularly older models or budget vehicles. This limits widespread adoption potential.

Privacy and Ethical Concerns: Continuous monitoring of driver emotions raises privacy concerns and potential misuse of emotional data. The system lacks built-in privacy protection mechanisms or user consent protocols, which may hinder adoption due to regulatory and ethical considerations.

7. Conclusion

With the use of real-time driver emotions monitoring, the research described in this paper effectively illustrates the creation and validation of a multimodal emotion detection system to improve vehicle safety. The suggested method makes use of the RAVDESS dataset to combine visual and aural input into a comprehensive model that can recognize and react to different emotional states with accuracy. The suggested hybrid CNN-LSTM model achieves an overall accuracy of 98.28% and shows notable superiority over conventional models by combining the CNN and LSTM networks. Because the combination of spatial and temporal variables allows for a richer knowledge of driver behavior, the incorporation of both audio and video inputs proved vital in boosting the model's capacity to detect and classify emotions properly. With 458 accurate classifications of "Safe" drivers and 403 correct classifications of "Unsafe" drivers, the confusion matrix provides more evidence of the model's dependability. This high degree of accuracy in differentiating between safe and unsafe emotional states highlights the resilience of the model and its potential use in practical settings. The results highlight how the car industry must use cutting-edge multimodal strategies to enhance driver monitoring and overall road safety. This

hybrid CNN-LSTM model is a significant step toward the development of safer and smarter transportation systems because of its efficacy in addressing the difficulties associated with emotion recognition. Subsequent research endeavours may concentrate on refining the system for practical implementation in driving environments, thereby augmenting its capacity as a vital constituent in the progression of intelligent transportation systems. By outperforming not only traditional ML and standalone deep learning models but also recent Transformer-based benchmarks, the proposed CNN-LSTM framework establishes a new state-of-the-art in multimodal emotion recognition tailored for intelligent automotive systems.

In addition to its high classification performance, the proposed CNN-LSTM model demonstrates efficient inference capabilities, achieving over 22 FPS on embedded devices like the NVIDIA Jetson Xavier NX and 38 FPS on standard GPU systems. With its lightweight architecture and low memory footprint, the system is suitable for real-time deployment in vehicles, supporting ADAS even under hardware-constrained conditions.

The system's demonstrated robustness to occlusions and environmental noise further supports its practical deployment for in-vehicle driver monitoring and emotion-based safety interventions.

The 2-second temporal window used for sequence modeling ensures rapid yet contextually rich emotion recognition, making the system ideal for safety-critical automotive environments.

The author received no financial assistance from any authorities or companies.

Availability of Data and Material

On request, the data from this study can be obtained from the corresponding author.

Competing Interests

No pertinent conflicts of interest are disclosed by the authors.

References

- 1) G. Oh, E. Jeong, R. C. Kim, J. H. Yang, S. Hwang, S. Lee and S. Lim, "Multimodal data collection system for driver emotion recognition based on self-reporting in real-world driving," *Sensors*, 22 4402 (2022). doi: 10.3390/s22124402
- 2) L. Mou, Y. Zhao, C. Zhou, B. Nakisa, M. N. Rastgoo, L. Ma, T. Huang, B. Yin, R. Jain and W. Gao, "Driver emotion recognition with a hybrid attentional multimodal fusion framework," *IEEE Transactions on Affective Computing*, 14 2970- 2981 (2023). doi:10.1109/TAFFC.2023.3250460
- 3) C. Y. Park, N. Cha, S. Kang, A. Kim, A. H.

- Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, 7 293 (2020). doi: 10.1038/s41597-020-00630-y
- 4) S. Shafaei, T. Hacizade and A. Knoll, "Integration of driver behavior into emotion recognition systems: A preliminary study on steering wheel and vehicle acceleration," *Computer Vision – ACCV 2018 Workshops*, 11367 386-401 (2019). doi: 10.1007/978-3-030-21074-8_32
 - 5) W. Sun, Y. Liu, S. Li, J. Tian, F. Wang and D. Liu, "Research on driver's anger recognition method based on multimodal data fusion," *Traffic Injury Prevention*, 25 354-363 (2023). doi: 10.1080/15389588.2023.2297658
 - 6) D. Ayata, Y. Yaslan and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *Journal of Medical and Biological Engineering*, 40 149–157 (2020). doi: 10.1007/s40846-019-00505-7
 - 7) N. Samadiani, G. Huang, B. Cai, W. Luo, C. H. Chi, Y. Xiang and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, 19 1863 (2019). doi: 10.3390/s19081863
 - 8) M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy and V. Chandran, "Automatic driver stress level classification using multimodal deep learning," *Expert Systems with Applications*, 138 112793 (2019). doi: 10.1016/j.eswa.2019.07.010
 - 9) S. Zepf, J. Hernandez, A. Schmitt, W. Minker and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Computing Surveys*, 53 1-30 (2020). doi: 10.1145/3388790
 - 10) A. I. Middy, B. Nag and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, 244 108580 (2022). doi:10.1016/j.knosys.2022.108580
 - 11) J. Zhang, Z. Yin, P. Chen and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, 59 103-126 (2020). doi: 10.1016/j.inffus.2020.01.011
 - 12) N. J. Shoumy, L. M. Ang, K. P. Seng, D. M. M. Rahaman and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *Journal of Network and Computer Applications*, 149 102447 (2020). doi: 10.1016/j.jnca.2019.102447
 - 13) M. Soleymani, M. Pantic and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, 3 211-223 (2012). doi: 10.1109/T-AFFC.2011.37
 - 14) P. Zhang, M. Fu, R. Zhao, D. Wu, H. Zhang, Z. Yang and R. Wang, "ECMER: Edge-cloud collaborative personalized multimodal emotion recognition framework in the internet of vehicles," *IEEE Network*, 37 192-199 (2023). doi: 10.1109/MNET.003.2300012
 - 15) L. Mou, C. Zhou, P. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain and W. Gao, "Driver stress detection via multimodal fusion using attention-based CNN-LSTM," *Expert Systems with Applications*, 173 114693 (2021). doi: 10.1016/j.eswa.2021.114693
 - 16) G. Sharma and A. Dhall, "A Survey on Automatic Multimodal Emotion Recognition in the Wild," *Advances in Data Science: Methodologies and Applications*, 35-64 (2020). doi: 10.1007/978-3-030-51870-7_3
 - 17) L. Sharara, M. Ismail, K. Thelen and A. Politis, "A Real-Time Automotive Safety System Based on Advanced AI Facial Detection Algorithms," *IEEE Transactions on Intelligent Vehicles*, 9 5080-5100 (2024). doi: 10.1109/TIV.2023.3272304.
 - 18) L. Davoli, M. Martalò, A. Cilfone, L. Belli, G. Ferrari, R. Presta and J. Plomp, "On driver behavior recognition for increased safety: a roadmap", *Safety*, 6 (2020). doi: 10.3390/safety6040055.
 - 19) R. R. Singh, S. Conjeti and R. Banerjee, "A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals" *Biomedical Signal Processing and Control*, 8 740-754 (2013). doi: 10.1016/j.bspc.2013.06.014.
 - 20) J. Zhang, Z. Yin, P. Chen and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review" *Information Fusion*, 59 103-126 (2020). doi: 10.1016/j.inffus.2020.01.011.
 - 21) X. Wang, Z. Sun, A. Chehri, G. Jeon and Y. Song, "Deep learning and multi-modal fusion for realtime multi-object tracking: Algorithms, challenges, datasets, and comparative study," *Information Fusion*, 105 102247 (2024). doi: 10.1016/j.inffus.2024.102247.
 - 22) B. Gao, K. Cai, T. Qu, Y. Hu and H. Chen, "Personalized Adaptive Cruise Control Based on Online Driving Style Recognition Technology and Model Predictive Control," *IEEE Transactions on Vehicular Technology*, 69 12482-12496 (2020). doi: 10.1109/TVT.2020.3020335.
 - 23) R. Yaswanth and M. R. Babu, "Revolutionizing Automotive Technology: Unveiling the State of Vehicular Sensors and Biosensors," *IEEE Access*, 12 192786-192812 (2024). doi: 10.1109/ACCESS.2024.3514157.
 - 24) J. Zhang, R. A. B. R. Ghazilla, H. J. Yap and W. Y. Gan, "A Comprehensive Review: Multisensory and

- Cross-Cultural Approaches to Driver Emotion Modulation in Vehicle Systems” *Applied Sciences*, 14 6819 (2024). doi: 10.3390/app14156819.
- 25) L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Duja, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, 8 53 (2021). doi: 10.1186/s40537-021-00444-8
 - 26) B. Chakravarthi, S. C. Ng, M. R. Ezilarasan and M. F. Leung, “EEG-based emotion recognition using hybrid CNN and LSTM classification,” *Frontiers in Computational Neuroscience*, 16 (2022). doi: 10.3389/fncom.2022.1019776
 - 27) A Framework for Recognition of Facial Expression Using HOG Features. *International Journal of Mathematics, Statistics, and Computer Science*, 2, 1–8. <https://doi.org/10.59543/ijmscs.v2i.7815>
 - 28) Face Mask Detection Using Haar Cascades Classifier. *International Journal of Mathematics, Statistics, and Computer Science*, 2, 19–27. <https://doi.org/10.59543/ijmscs.v2i.7845>
 - 29) N. Ying, Y. Jiang, C. Guo, D. Zhou and J. Zhao, “A multimodal driver emotion recognition algorithm based on the audio and video signals in internet of vehicles platform,” *IEEE Internet of Things Journal*, (2024). doi: 10.1109/jiot.2024.3363176
 - 30) T. Anvarjon, Mustaqeem and S. Kwon, “Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features,” *Sensors*, 20 5212 (2020). doi: 10.3390/s20185212
 - 31) Mustaqeem and S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors*, 20 183 (2020). doi: 10.3390/s20010183
 - 32) F. Tao and G. Liu, “Advanced LSTM: A study about better time dependency modeling in emotion recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2906-2910 (2017). doi: 10.48550/arXiv.1710.10197
 - 33) N. Senthilkumar, S. Karpakam, M. G. Devi, R. Balakumaresan and P. Dhilipkumar, “Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks,” *Material Today Proceedings*, 57 2180-2184 (2022). doi: 10.1016/j.matpr.2021.12.246
 - 34) C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero and F. Fernández-Martínez, “Multimodal emotion recognition on RAVDESS dataset using transfer learning,” *Sensors*, 21 7665 (2021). doi: 10.3390/s21227665
 - 35) C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero and F. Fernández-Martínez, “A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset,” *Applied Sciences*, 12 327 (2022). doi: 10.3390/app12010327
 - 36) A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria and L. P. Morency, “Memory fusion network for multi-view sequential learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (2018). doi: 10.1609/aaai.v32i1.12021
 - 37) H. Pham, T. Manzini, P. P. Liang and B. Poczós, “Seq2Seq2 Sentiment: multimodal sequence to sequence models for sentiment analysis,” *arXiv*, (2018). doi: 10.48550/arXiv.1807.03915
 - 38) S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh and A. Hussain, “Multimodal sentiment analysis: Addressing key issues and setting up the baselines,” *IEEE Intelligent Systems*, 33 17-25 (2018). doi: 10.1109/MIS.2018.2882362
 - 39) A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria and L. P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2236-2246 (2018). doi: 10.18653/v1/P18-1208
 - 40) J. Liang, R. Li and Q. Jin, “Semi-supervised multimodal emotion recognition with cross-modal distribution matching,” *Proceedings of the 28th ACM International Conference on Multimedia*, 2852–2861 (2020). doi: 10.48550/arXiv.2009.02598
 - 41) Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.
 - 42) Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*.