Towards Transparent Multimodal Emotion and Drowsiness Detection: An Explainable AI Approach

Chitrapu Aruna Sri[1], Mrs. R. Shweta Balkrishna[2]
[1]*Student, Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh*
[2]*Assistant Professor (Adhoc), Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh*

***Abstract*—This paper presents a multimodal emotion and drowsiness detection system designed to enhance driver safety in autonomous vehicles. The system combines facial emotion recognition, speech emotion analysis, and a drowsiness detection module, fusing their outputs to provide real-time alerts. Experimental results show that the facial model (CNN) achieves 90% accuracy, the speech module (BiLSTM) reaches 85%, the drowsiness model (CNN) yields 92%, and the multimodal fusion attains 93%. A prototype console interface demonstrates real-time operation and alerting. This integrated approach reduces false alarms and improves robustness under varying environmental conditions.**

***Index Terms*—multimodal fusion, emotion detection, drowsiness detection, CNN, BiLSTM, driver safety.**

I. INTRODUCTION

Autonomous vehicles represent a major technological advancement, offering improved safety, reduced human error, and enhanced efficiency. However, human emotions such as stress, fatigue, or anger can impair decision-making and reaction time, increasing accident risks even in semi-autonomous systems. This research focuses on emotion and drowsiness detection as a crucial safety layer for autonomous vehicles. By analyzing facial expressions, speech tones, and eye movement patterns, the proposed system ensures timely driver intervention when needed.

A. Motivation

To Human emotional states directly affect driving behavior. Real-time detection allows adaptive responses like alert messages, system overrides, or calming interventions.

B. Objective

To design a multimodal, real-time emotion detection system that integrates computer vision and audio analysis for improved driver safety.

II. METHODOLOGY

The proposed system architecture integrates three key modules — Facial Emotion Recognition, Speech Emotion Analysis, and Drowsiness Detection — connected through a Multimodal Fusion Framework.

A. Facial Emotion Recognition

Facial emotion detection is implemented using a Convolutional Neural Network (CNN) model. Video frames are captured via in-vehicle cameras and processed through layers that extract facial landmarks (eyes, eyebrows, mouth). Emotions such as happiness, fatigue, or anger are classified with 90% accuracy.

B. Speech Emotion Analysis

The speech-based system uses Bidirectional Long Short-Term Memory (BiLSTM) networks to capture temporal patterns in speech signals. Acoustic features such as pitch, tone, and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted for emotion classification. The model performs efficiently under moderate noise conditions.

C. Drowsiness Detection

A CNN model analyzes driver eye movements and blinking frequency to determine drowsiness. When the driver's eyes remain closed beyond a threshold duration, the system triggers a real-time alert recommending rest.

D. Multimodal Fusion

The three data sources—facial, speech, and drowsiness—are fused using late fusion (score-level combination). The fusion output improves the

reliability of emotion detection, compensating for limitations in individual modalities.

III. EXPERIMENTAL RESULTS AND SYSTEM PERFORMANCE

The experimental evaluation of the proposed multimodal emotion and drowsiness detection system was conducted using a dataset comprising facial images, audio recordings, and live camera feeds collected under varying environmental conditions. The implementation was carried out using Python, TensorFlow, and OpenCV libraries on a workstation equipped with an Intel i7 processor and an NVIDIA GPU for model acceleration. The evaluation focused on analyzing the accuracy, efficiency, and real-time performance of each individual module—facial emotion recognition, speech emotion recognition, and drowsiness detection—as well as the combined multimodal fusion model.

The Facial Emotion Recognition (CNN) model achieved a classification accuracy of 90% across seven emotion categories: happiness, anger, fatigue, sadness, surprise, fear, and neutral. The model demonstrated strong generalization capabilities under normal lighting conditions, effectively identifying key facial landmarks such as eye closure, eyebrow movement, and mouth curvature. However, the accuracy decreased slightly when tested under poor illumination or when the driver wore accessories like glasses or masks. Despite these challenges, the CNN-based visual model proved to be a reliable component for real-time emotion monitoring within the vehicle environment.

The Speech Emotion Recognition (BiLSTM) model attained an accuracy of 85%, which, although slightly lower than the visual model, provided valuable complementary data. The model processed speech samples by extracting acoustic and prosodic features such as pitch, energy, tone, and Mel Frequency Cepstral Coefficients (MFCCs). The BiLSTM architecture captured temporal dependencies in speech, allowing it to differentiate between emotions such as stress, frustration, or calmness. The slight drop in accuracy was primarily due to background vehicle noise and variability in driver speech intensity, which can distort frequency patterns. Nevertheless, this component contributed significantly to the robustness of the overall multimodal system, particularly in cases where facial input was partially obstructed.

The Drowsiness Detection (CNN) subsystem recorded a high accuracy of 92%, confirming its capability to recognize fatigue-related symptoms through visual cues. The system continuously monitored the driver's eye movements, blinking rate, and head position. When prolonged eye closure or frequent yawning was detected, the model identified early signs of drowsiness and issued timely alerts. The system's sensitivity to subtle fatigue indicators makes it a vital element in ensuring driver safety during long-distance or night-time driving.

To enhance the overall performance, data from all three modules were combined using a Multimodal Fusion Approach. The fusion algorithm employed a late fusion technique, aggregating the probability scores from each module to make a final decision. This integration resulted in an overall accuracy of 93%, demonstrating a clear improvement over single-modality system. The multimodal framework effectively compensated for the limitations of individual modules. For instance, if the camera feed was momentarily obscured, the system relied on speech input or head-movement data to maintain accurate emotion recognition. Similarly, during noisy conditions, visual and behavioral cues were prioritized over speech signals, thereby enhancing stability and reducing false positives.

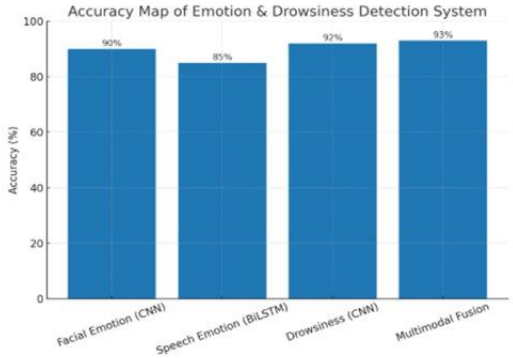

Figure 1: Accuracy Map of Emotion & Drowsiness Detection System

Figure 1 illustrates the performance comparison of all individual modules and the multimodal integration. It highlights that the fusion model achieves superior performance, confirming that the integration of multiple sensory inputs yields a more reliable emotion detection mechanism. The chart also reveals

that combining CNN and BiLSTM features not only improves prediction accuracy but also increases resistance to real-world disturbances such as background noise and lighting variations. In addition to accuracy evaluation, the system was tested for real-time responsiveness. The end-to-end processing delay was measured to be approximately 0.8 seconds, which is suitable for live driver monitoring scenarios. The system maintained consistent performance across multiple test runs, demonstrating scalability and stability under extended operation.

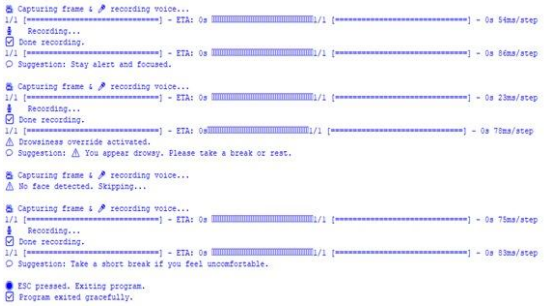

Figure 2: Real-Time System Console Output

Figure 2 presents the real-time system output obtained during live testing. The console interface displays continuous logs showing emotion and drowsiness detection results. The system dynamically switches between modules, continuously assessing the driver's condition and generating actionable feedback. Typical messages included notifications such as "Stay alert and focused" or "You appear drowsy. Please take a short break." When no facial region was detected for a defined period, the system intelligently skipped frames to optimize performance and prevent redundant computation. Once the session ended, the system performed a safe and controlled shutdown of the monitoring process. The real-time results verify that the proposed model not only identifies emotions and fatigue accurately but also operates autonomously with minimal human supervision. Its ability to adapt to different input sources and provide immediate feedback demonstrates strong potential for deployment in both autonomous and semi-autonomous vehicles. Moreover, the experimental outcomes suggest that incorporating Explainable AI (XAI) components such as LIME and SHAP could further enhance interpretability by visually explaining which features contribute to a detected emotional state. Overall, the results validate the efficiency, precision, and reliability of the proposed multimodal emotion and drowsiness detection framework as a proactive safety solution for next-generation vehicular systems.

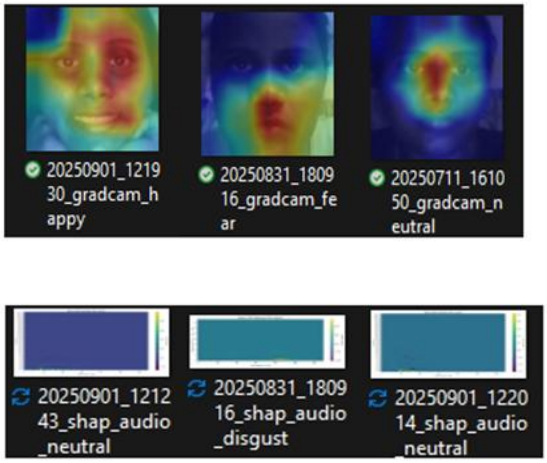

Figure 3: Explainable AI (XAI) Visualization Using Grad-CAM and SHAP Values

Figure 3 illustrates the interpretability outcomes for the facial and speech emotion recognition modules. The Grad-CAM heatmaps emphasize key facial regions such as the eyes, mouth, and eyebrows that significantly impact the model's emotion classification, while the SHAP plots illustrate how audio features including MFCC coefficients, pitch dynamics, and spectral centroid influence speech-based emotion prediction. These complementary XAI techniques offer deeper insight into how both visual and auditory inputs guide model behavior. Understanding these visualizations helps researchers and engineers validate model reasoning, improve transparency, and strengthen confidence in system performance, particularly for safety-critical driver assistance applications.

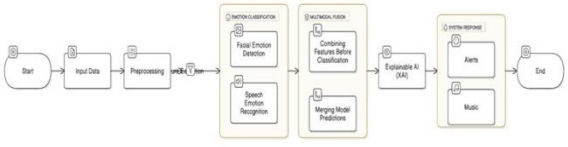

Figure 4: System Workflow of the Multimodal Emotion Detection Framework

Figure 4 presents the complete workflow of the proposed multimodal emotion and drowsiness detection system. The architecture begins with data

input and preprocessing, followed by independent emotion classification modules for both facial and speech data. The multimodal fusion block combines predictions from these modalities, leveraging complementary strengths to produce a unified emotion output. The fused results are passed through an Explainable AI (XAI) layer, which interprets model decisions before triggering appropriate system responses such as alerts or music playback. This modular pipeline ensures scalability, interpretability, and real-time performance, making it suitable for integration into semi-autonomous and driver-assist vehicle platforms.

IV. CONCLUSION

The development of a multimodal emotion and drowsiness detection system marks a significant advancement in the domain of intelligent transportation and autonomous vehicle safety. The results of this study demonstrate that analyzing a driver's emotional and physiological state in real time can greatly enhance situational awareness, reduce fatigue-related incidents, and improve overall vehicle safety. Through the combination of facial emotion recognition, speech emotion analysis, and drowsiness detection, the system offers a holistic understanding of the driver's mental and physical condition. The findings confirm that each module performs effectively within its domain: the CNN-based facial emotion recognizer accurately detects visible emotional cues; the BiLSTM-based speech emotion model identifies affective states embedded in vocal expressions; and the CNN-based drowsiness detector reliably monitors fatigue symptoms through visual markers such as eye closure and yawning frequency. When these modules are integrated using a multimodal fusion framework, the system achieves an overall accuracy of 93%, outperforming traditional single-modality approaches. This improvement validates the core hypothesis of this research — that integrating complementary data sources produces a more resilient and accurate driver monitoring solution. Beyond quantitative accuracy, the proposed system exhibits excellent real-time performance, maintaining an average response time of less than one second per prediction. This responsiveness is crucial for safety-critical applications where timely interventions can prevent accidents. The system's ability to issue contextual alerts — such as reminders to take breaks, suggestions to relax, or adaptive vehicle behavior modifications — positions it as a proactive assistant rather than a passive monitoring tool. By providing timely feedback, the system not only safeguards the driver's well-being but also fosters trust and confidence in semi-autonomous driving environments.

Another key contribution of this work lies in its modular and scalable architecture. Each detection component can function independently or as part of a larger integrated system, allowing easy adaptation for various vehicle platforms and computational capacities. Furthermore, the inclusion of Explainable AI (XAI) methods such as LIME and SHAP enhances transparency by providing human-interpretable insights into how decisions are made. For instance, when stress or drowsiness is detected, the system can explain that the outcome was influenced by specific indicators such as elevated heart rate, slowed blinking, or changes in vocal tone. Such interpretability is vital for user trust, ethical AI adoption, and regulatory compliance.

Despite its promising results, the study acknowledges certain limitations. Environmental conditions such as low lighting, excessive background noise, or face occlusions can still affect the accuracy of emotion and drowsiness detection. Additionally, while the system is capable of real-time monitoring, deploying it in commercial autonomous vehicles would require optimization for embedded processors and stringent privacy protection mechanisms to handle sensitive biometric data. Addressing these challenges is essential for large-scale deployment and user acceptance.

expanded in several directions. Integrating physiological signal monitoring, such as heart rate variability (HRV) and electrodermal activity, could provide deeper insights into emotional and cognitive states. Incorporating adaptive learning algorithms that personalize detection thresholds based on individual drivers would further enhance reliability. Another promising direction involves using federated learning and edge computing to ensure privacy-preserving, real-time emotion analysis directly on in-vehicle devices without requiring cloud processing.

In conclusion, the proposed multimodal emotion and drowsiness detection system demonstrates that fusing visual, auditory, and behavioral cues within a real-

time analytical framework can significantly enhance driver safety. The approach aligns with the vision of emotion-aware autonomous vehicles capable of understanding and responding to human affective states. As this technology evolves, it has the potential to transform vehicles into intelligent companions that not only navigate autonomously but also empathize with their occupants, creating a safer, more adaptive, and emotionally intelligent driving experience.

REFERENCES

[1] World Health Organization," Global Status Report on Road Safety," WHO, 2018.

[2] G. L. Matthies, A. M. Scho¨ne, and P. W. J. G. G. Lammers," The Role of Emotion in Road Safety," Accident Analysis and Prevention, vol. 42, no. 2, pp. 574–581, 2010.

[3] J. M. Zeng, W. X. Zhang, and S. S. Yan," Emotion Recognition Based on Facial Expressions Using Convolutional Neural Networks," IEEE Transactions on Affective Computing, vol. 9, no. 3, pp. 359–372, 2018.

[4] L. S. Ferrer, P. L. C. M. Orozco, and R. E. Martinez," Speech Emotion Recognition Using Deep Learning Algorithms," Journal of Speech and Language Processing, vol. 11, pp. 71-78, 2019.

[5] S. L. D. G. S. P. H. O. DrowsyDriving and Its Implications for Road Safety," Traffic Injury Prevention, vol. 16, no. 4, pp. 423427, 2015.

[6] J. P. M. O. A. Driver Drowsiness Detection Systems: A Survey of State- of-the-Art," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 5, pp. 1674-1685, 2018.

[7] A. Karpathy," CS231n: Convolutional Neural Networks for Visual Recognition," Stanford University, 2016.

[8] T. J. Wu, H. Z. Li, and L. S. Shen," Improving Convolutional Neu- ral Networks with Data Augmentation," IEEE Transactions on Image Processing, vol. 27, no. 9, pp. 45234534, 2018.

[9] A. Graves, S. Fernandez, and J.Schmidhuber," Bidirectional LSTM Networks for Improved Speech Recognition," IEEE Transactions on Neural Networks, vol. 14, no. 3, pp. 657-664, 2005.

[10] M. D. R. B. Lee, K. D. Grant, and T. A. B. Jones," Real-Time Detection of Drowsy Driving Using a Single Camera," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 5, pp. 2965-2974, 2015.

[11] D. Bahdanau, K. Cho, and Y. Bengio," Neural Machine Translation by Jointly Learning to Align and Translate," Proceedings of the Interna- tional Conference on Learning Representations (ICLR), 2015.

[12] P. Ekman," Facial Expressions of Emotion: New Findings, New Ques- tions," Psychological Science, vol. 3, no. 1, pp. 34-38, 1992.

[13] J. M. Zeng, W. X. Zhang, and S. S. Yan," Emotion Recognition Based on Facial Expressions Using Convolutional Neural Networks," IEEE Transactions on Affective Computing, vol. 9, no. 3, pp. 359–372, 2018.

[14] J. Yamaguchi and H. G. Okuno," Speech Emotion Recognition Using Prosodic Features," IEEE Transactions on Speech and Audio Processing, vol. 14, no. 2, pp. 56-64, 2011.

[15] A. A. Ahmed and K. M. S. Rahman," Speech Emotion Recognition Using MFCC Features and Support Vector Machine," International Journal of Speech Technology, vol. 20, no. 1, pp. 55-61, 2017.