Multimodal Driver Stress Detection in Real-World Driving Scenarios

Heonjun Lee, and Seoyoung Ko, and Youngtae Noh
School of Data Science, Hanyang University

Research Motivation

❖ The Challenge

- **Real-time detection difficulty:** Single-modality signals are prone to noise and motion artifacts
- **User constraints:** Intrusive sensors often cause driver discomfort, affecting data validity

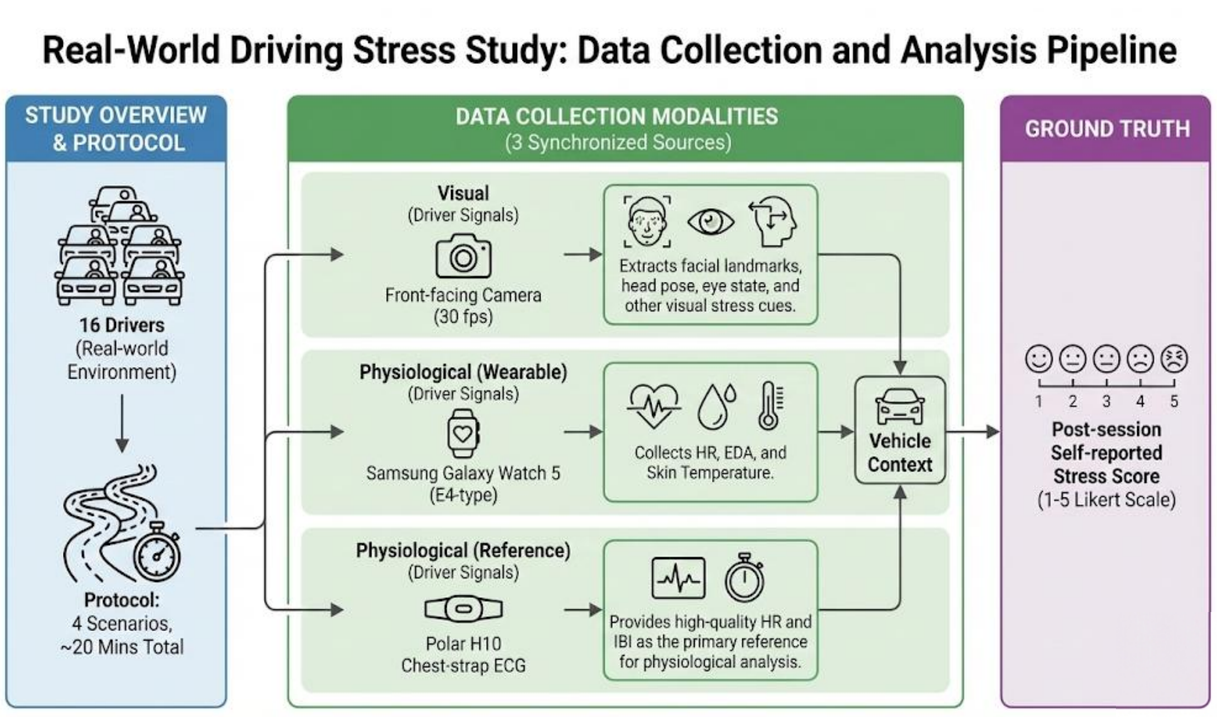
❖ Our Approach: Multimodal Fusion

- **Visual Cues:** Captures in-cabin behavioral data(Head Pose, Eye State)
- **Physiological Signals:** Integrates biometric data(HR, EDA, Temperature) collected during real-driving scenarios

❖ Key Contribution

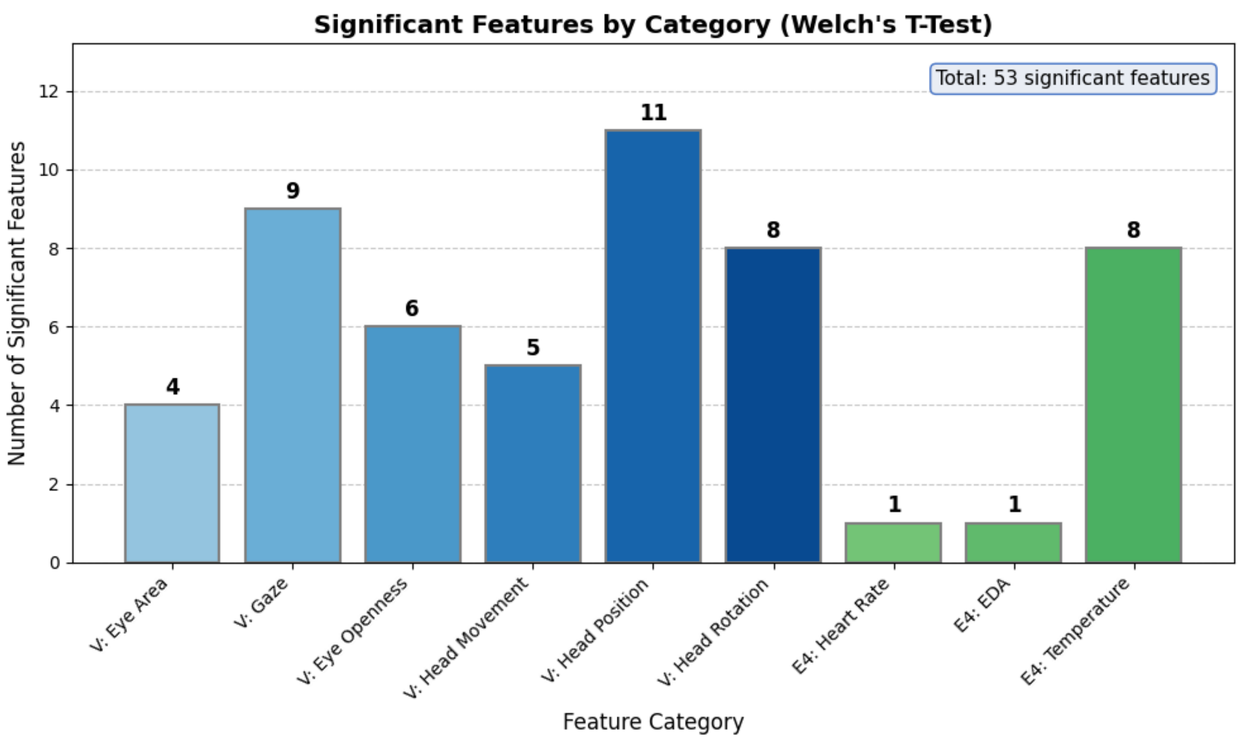
- **Enhanced Robustness:** Mitigates the impact of individual sensor failure
- **Higher Reliability:** Fusing modalities captures a more accurate representation of real-world stress responses

Dataset



Analysis Results

Welch's t-test was performed to assess feature discrepancies between stress groups, dichotomized by a threshold score of 2 (**Non-Stress**: 1–2 vs. **Stress**: 3–5).

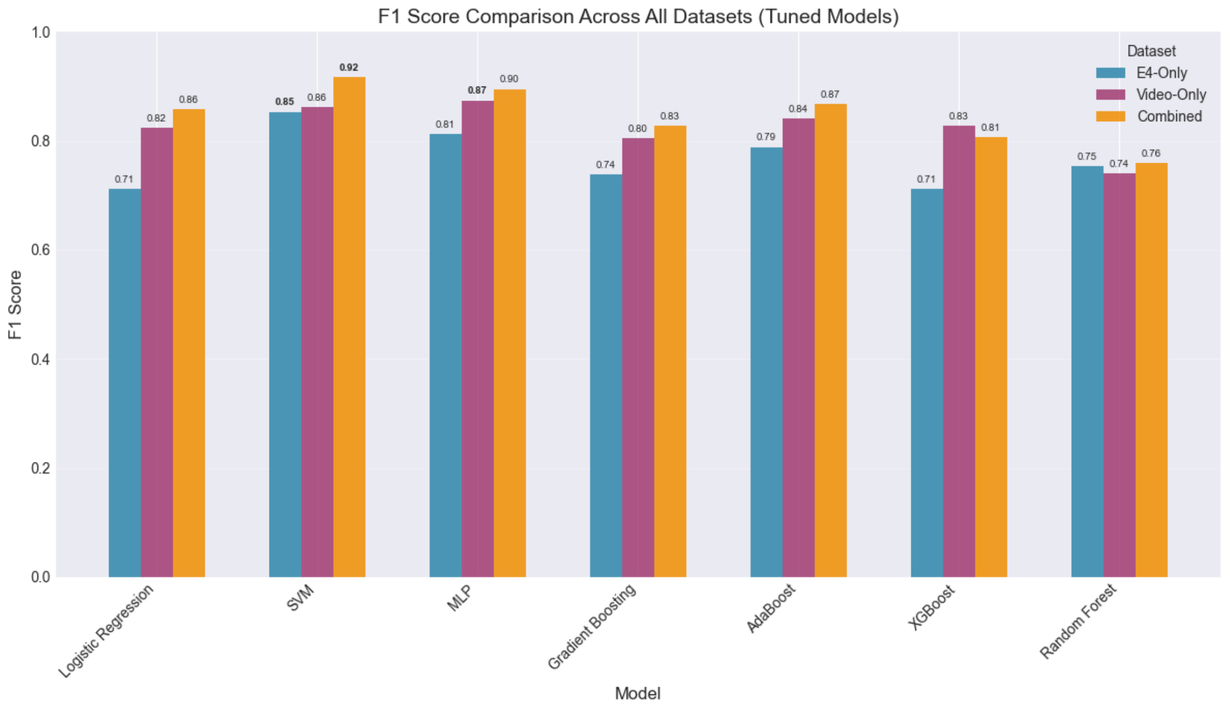


- **Feature Analysis:** Identified **53 significant features** ($p < 0.05$) from a pool of 183 multimodal metrics
- **Key Finding**: Video-based analysis yielded 43 statistically significant features, significantly outnumbering the 10 features identified from E4 physiological data. Video-derived **head position** yielded the highest number of significant features. ($n = 11$)

Model Development

To minimize inter-subject variance, **distinct models** were trained for each driver. The reported results represent the average classification accuracy of **seven algorithms** across **physiological, video, and multimodal datasets**.

For the **binary classification** task, self-reported stress scores (Likert scale 1-5) were dichotomized: scores 1–2 were categorized as 'Non-Stress,' while scores 3–5 were labeled as 'Stress.



Optimal Model Selection per Data Source

Model	Dataset	Accuracy	Precision	Recall	F1	AUROC
SVM	E4-only	0.9420	0.8460	0.8742	0.8532	0.9506
	Video-only	0.9550	0.8548	0.8798	0.8622	0.9759
	Combined	**0.9750**	**0.9157**	**0.9211**	**0.9165**	**0.9860**

- **Model Performance**: SVM achieved peak performance on Multimodal inputs ($\mathrm{F1}: 0.9165$)
- **Modality Trend**: Classification accuracy progressively improved from physiological only, to visual only, to combined fusion
- **Feature Correlation**: The superior performance of video-based models aligns with the higher density of significant features found in video data ($n = 43$) compared to E4 sensors ($n = 10$)
- **Conclusion**: Multimodal integration significantly enhances detection capability by leveraging complementary feature sets

Future Work

❖ Improved Temporal Modeling

- With access to real-time or continuous stress labels, extend the model using **Transformer-based** temporal architectures for finer sequential analysis

❖ Multiclass Stress Prediction

- Expand beyond binary classification to predict **multiple stress levels** for more nuanced driver-state estimation

❖ Stronger Multimodal Fusion

- Develop **deeper fusion** of visual and physiological features to improve robustness against noise and single-sensor failure

❖ Toward Real-Time Alerts

- Advance the system toward **real-time driver notifications** when elevated stress is detected