Trust What You Fuse: Reliability-Aware Cross-Attention for Multimodal Physiological Stress Assessment in the Wild

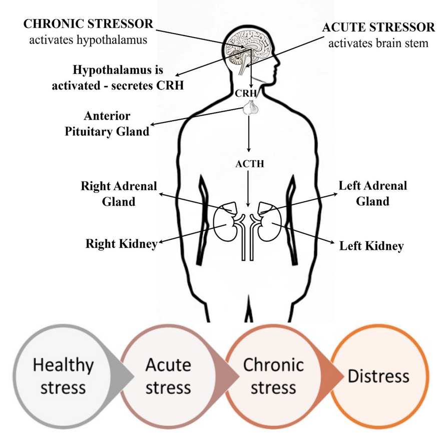
భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Kaveti Pavan Kumar[1]; Nagarajan Ganapathy[1]
[1]Indian Institute of Technology Hyderabad, Telangana, India

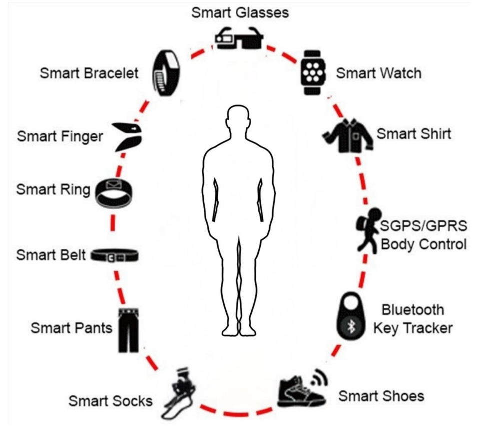
CVPR JUNE 3-7, 2026 DENVER COLORADO

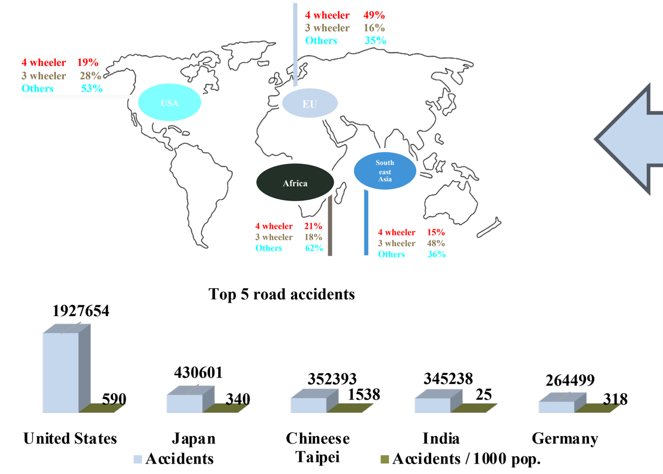
Motivation & Challenges

Acute and Chronic Stressor Effects Brain Differently [1]



Enables non-invasive, real-time assessment of stress and emotional states [2]





1 Physiological signals are interdependent

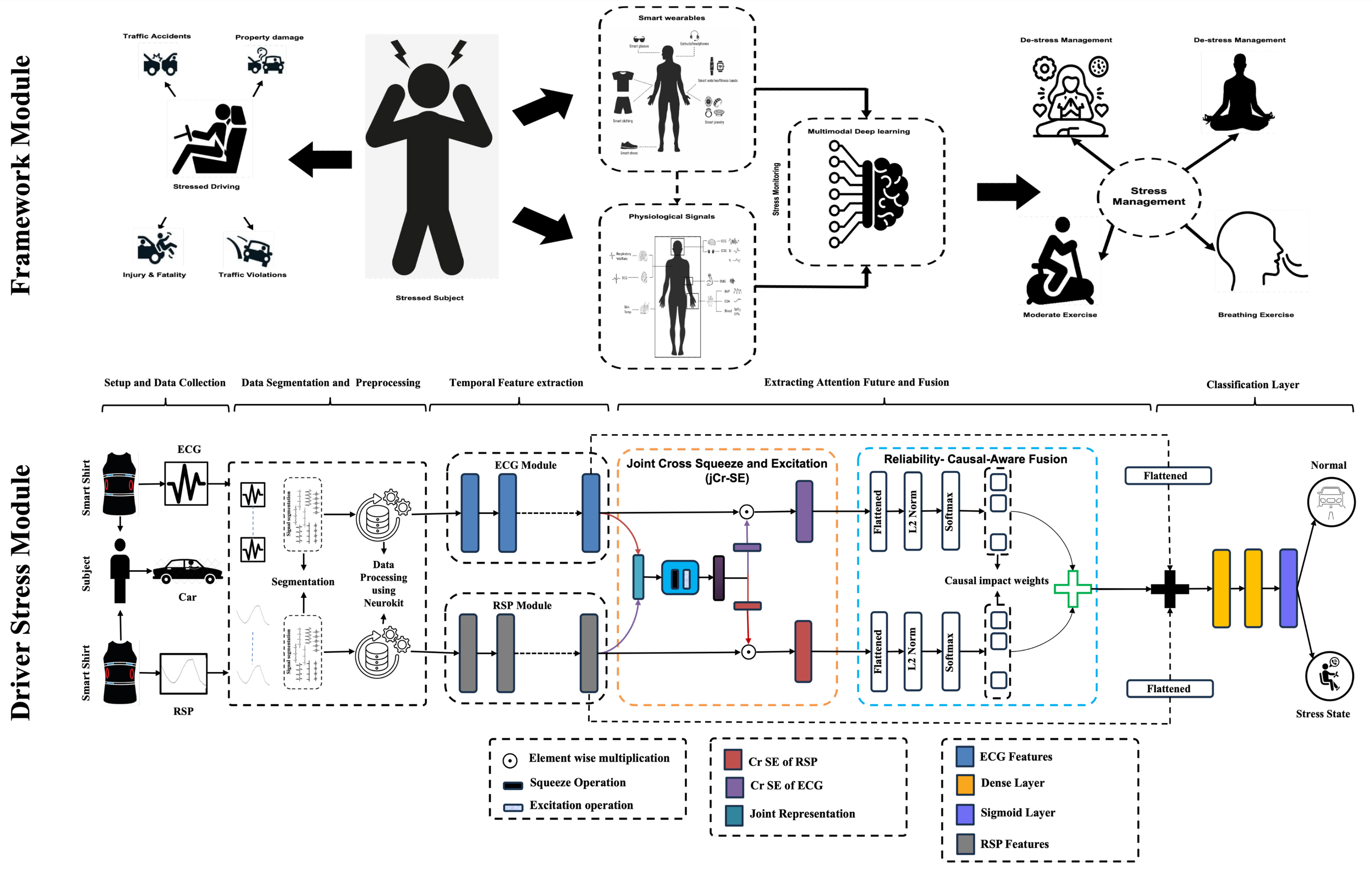
ECG and respiration directly reflect autonomic nervous system responses to driver stress. These modalities are not independent — cardiorespiratory coupling (e.g. respiratory sinus arrhythmia) means one signal actively shapes the other.
→ Simple concatenation or independent fusion ignores this coupling, leaving complementary information unexploited.
→ Cross-modal interaction modeling is essential to capture the full physiological picture.

2 Wearable signal quality varies — reliability is not equal

In real-world driving, wearable sensors are susceptible to motion artifacts, sensor displacement, and noise — causing modality reliability to shift across time and subjects.
→ Fixed-weight fusion amplifies unreliable signals, degrading robustness across subjects.
→ Dynamic, reliability-aware weighting is needed to adapt to real-world conditions.

Challenges of Stress During Driving in real world

Methods



Proposed Multimodal Feature Learning and Reliability-Guided Fusion Pipeline

Feature Extraction

$$F_{m_1} = f_{m_1}(X_{m_1}), \quad F_{m_n} = f_{m_n}(X_{m_n})$$

Cross-SE Interaction Module [7]

$$z_{m_1} = GAP(F_{m_1}), \quad z_{m_n} = GAP(F_{m_n})$$

$$z_{concat} = [z_{m_1}, z_{m_n}]$$

$$SE = \sigma\left(W_2 \cdot \delta(W_1 \cdot z_{concat})\right)$$

$$F^{cross}_{m_1} = F_{m_n} \odot SE, \quad F^{cross}_{m_n} = F_{m_1} \odot SE$$

Causal-Aware Fusion [3]

$$Score(x_m) = \|x_m\|_2$$

$$\lambda_{m_1} = \frac{e^{Score(x_{m_1})}}{e^{Score(x_{m_1})} + e^{Score(x_{m_n})}}, \quad \lambda_{m_n} = \frac{e^{Score(x_{m_n})}}{e^{Score(x_{m_1})} + e^{Score(x_{m_n})}}$$

$$\tilde{x}_{m_1} = \lambda_{m_1} \cdot x_{m_1}, \quad \tilde{x}_{m_n} = \lambda_{m_n} \cdot x_{m_n}$$

Reliability-Guided Feature Integration

$$x^{final}_{m_1} = \tilde{x}_{m_1} \odot x_{m_1}, \quad x^{final}_{m_n} = \tilde{x}_{m_n} \odot x_{m_n}$$

$$x_{fusion} = [x^{final}_{m_1}, x^{final}_{m_n}]$$

$$x^{final}_{fusion} = [x_{fusion}, F_{m_1}, F_{m_n}]$$

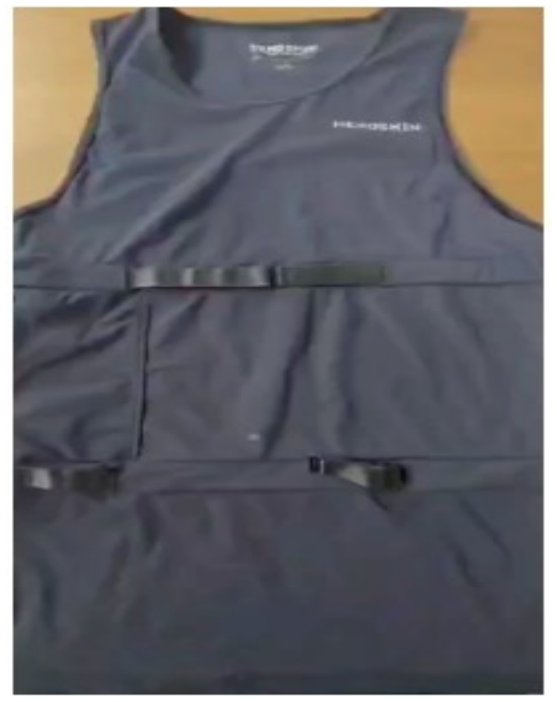
Feature Extraction to Reliability-Guided Fusion

Dataset Details

- **Participants & Setup:** 20 healthy male participants performed controlled driving experiments at Indian Institute of Technology Hyderabad using the Hexoskin Pro Kit smart shirt.
- **Data Acquisition:** ECG (256 Hz) and respiration (128 Hz) were recorded under two conditions: normal driving and phone-call-induced distraction.
- **Preprocessing & Segmentation:** Signals were segmented into non-overlapping 10-second windows with balanced classes and processed using NeuroKit2 [5] with Butterworth filtering (ECG: 0.5 Hz LP, RESP: 0.05–3 Hz).
- **Experimental Design:** Data collected in real driving using Renault Triber [4], with rest intervals between tasks to ensure reliable physiological measurements.

Materials & Experimental Design

Smart Shirt – Hexoskin [6]



Subject wearing the smart shirt and data recording

Subject View point inside experimental vehicle during driving

Contributions

1 Joint cross-modal interaction learning

Cross Squeeze-and-Excitation (JcrSE) enables channel-wise feature recalibration across modalities, explicitly capturing cardiorespiratory coupling between ECG and respiration.

Cross-SE module

2 Causal-aware reliability-guided fusion

Modality weighting via ℓ_2-norm representation scores dynamically adjusts each modality's contribution — suppressing noisy signals and emphasizing reliable ones during integration.

Causal-aware fusion

Experimented Results

Table 1. Architectural Comparison.

Model	Cross-SECA	Acc	F1	Prec	Rec
1D CNN	No	66.66	60.56	73.48	66.66
	Yes	**68.05**	**63.44**	**70.38**	**67.93**
InceptionTime	No	66.52	63.01	67.64	65.95
	Yes	**69.30**	**66.23**	**71.35**	**68.89**
TCN	No	63.19	60.44	66.11	62.80
	Yes	**64.99**	**62.78**	**67.47**	**64.50**
MobileNet-1D	No	68.33	64.43	69.10	68.01
	Yes	**71.52**	**68.99**	**76.28**	**70.91**
ConvMixer-1D	No	58.05	47.59	49.02	57.97
	Yes	**61.52**	**53.92**	**61.62**	**60.85**
1D ResNet	No	70.41	68.23	71.37	69.78
	Yes	**74.16**	**71.56**	**76.87**	**73.99**
DenseNet	No	69.30	66.35	**70.21**	68.87
	Yes	**69.72**	67.32	**70.87**	**68.97**
Transformer	No	55.41	54.63	56.03	54.51
	Yes	**66.94**	**64.58**	**71.36**	**66.50**
ConvTransformer	No	55.41	54.63	56.03	54.93
	Yes	**66.94**	**64.58**	**71.36**	**66.94**

Table 2. Different Fusions.

Configuration	Acc	F1	Prec	Rec
Baseline (No fusion)	68.33	64.43	69.10	68.01
Cross-SE	73.61	70.29	77.91	73.61
Causal-aware	68.05	65.29	70.97	67.87
Cross-SECA	**74.44**	**72.08**	**76.45**	**73.90**

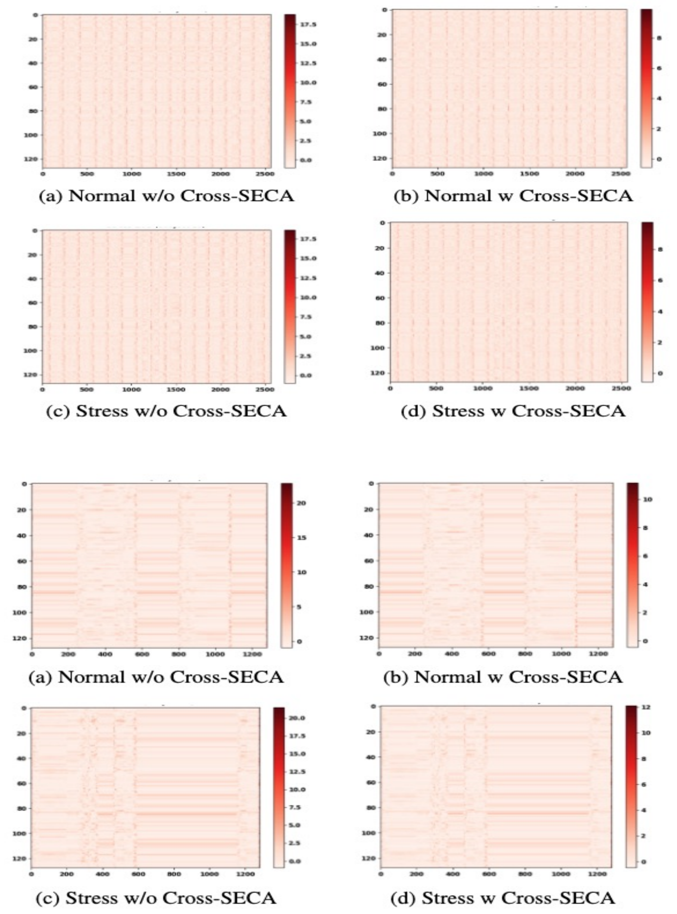
Table 3 Stages Comparison.

Insertion Stage	Acc	F1	Prec	Rec
After Layer 1	**74.58**	**72.01**	**77.51**	**73.80**
After Layer 2	74.16	72.52	77.30	73.98
After Layer 3	71.11	68.29	77.27	70.77
After Layer 4	74.44	72.08	76.45	73.86

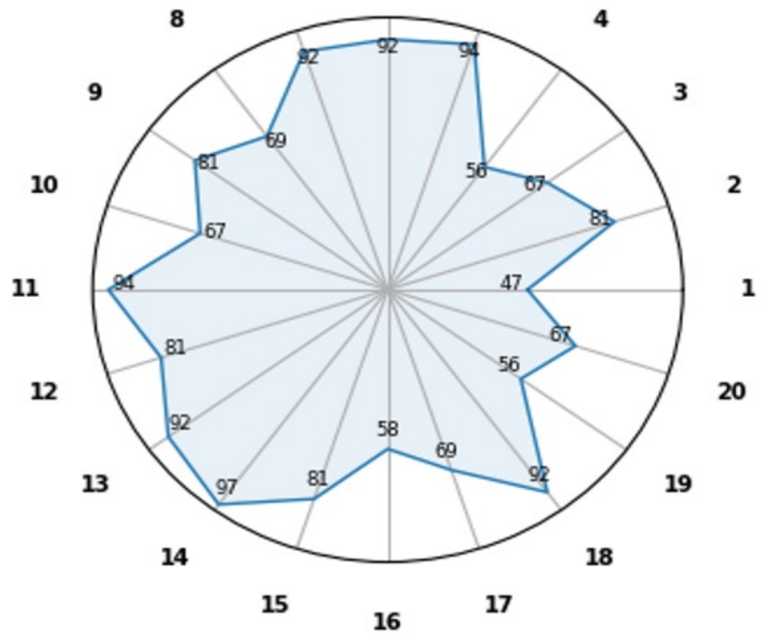
Table 4. Modality Comparison.

Modality	Acc	F1	Prec	Rec
ECG	63.88	61.18	63.44	55.40
RSP_{tho}	50.97	27.84	50.97	35.20
RSP_{abd}	65.05	63.42	66.05	64.48
ECG + RSP_{tho}	63.33	55.98	63.33	54.03
ECG + RSP_{abd}	**66.66**	**60.56**	**73.48**	**66.66**
RSP_{tho} + RSP_{abd}	66.03	61.87	66.84	66.25
ECG + RSP_{tho} + RSP_{abd}	65.50	65.93	63.50	61.33

Overall Tabular Results



ECG and RSP feature activation heatmaps (with/without Cross-SECA) and subject-wise classification accuracy of the best-performing model.



Discussion & Conclusions

- Introduced Cross-SECA, a multimodal physiological learning framework that integrates cross-modal interaction (Cross-SE) with causal-aware reliability fusion for driver stress detection.
- Effectively captures interdependencies between physiological signals while dynamically adapting to modality reliability, improving robustness against variability in real-world conditions.
- Demonstrates consistent performance improvements across temporal architectures, highlighting strong generalizability for multimodal physiological analysis.

Contact

Kaveti Pavan Kumar
Indian Institute of Technology
Email: bm23resch01001@iith.ac.in
Website: https://www.medinfolab.com/
Lab Name: Medical Informatics and Digital Health
Phone: +919848946301

MINDH Medical Informatics & Digital Health Lab

References

1. Healey JA, Picard RW. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on intelligent transportation systems. 2005 Jun 30;6(2):156-66
2. Yang S, Gao Y, Zhu Y, Zhang L, Xie Q, Lu X, Wang F, Zhang Z. A deep learning approach to stress recognition through multimodal physiological signal image transformation. Scientific Reports. 2025 Jul 1;15(1):22258
3. Xu X, Muhammad G. Multi-modal driving fatigue detection with causal-aware fusion and Kolmogorov-Arnold network. Information Fusion. 2026 Feb 26:104255.
4. Pavan K, Singh A, Pawar DS, Ganapathy N. Multimodal Wearable Based Automated Driver Inattention State Assessment Using Multi Devices and Novel Cross-Modal Attention Framework. IEEE Sensors Letters. 2025 Aug 7.
5. Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SA. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior research methods. 2021 Aug;53(4):1689-96.
6. Pavan K, Singh A, Igasaki T, Pawar DS, Ganapathy N. Assessment of Driver's Stress State Using Smart T-Shirt Textile Electrodes and Multimodal Cross-Attention Networks. IEEE Sensors Letters. 2024 Sep 12;8(10):1-4.
7. Joze HR, Shaban A, Iuzzolino ML, Koishida K. MMTM: Multimodal transfer module for CNN fusion. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020 (pp. 13289-13299).