Detection of in-the-wild and long-term stress

based on cross-attention transformer and context-aware ensemble model

Authors: Z Xia, CH Chen, MH Hsieh, J Da Lim, HCK Sng, MK Ahmad

Advanced Engineering Informatics (2025) — Elsevier

Presented by Nooshin Taheri

6/24/2026

Motivation: why stress detection matters in VTS

- ~90% of world merchandise trade moves by sea.
- Vessel Traffic Service (VTS) operators keep ships safe in busy waterways — under high workload, time pressure, and safety-critical decisions.
- Rising vessel traffic, size, and density increase the risk of collisions, groundings, and near-misses.
- 89.5% of 2014–2020 maritime accidents were linked to human factors.
- Stress can impair information processing, emergency detection, and decision-making.
- PPG is practical for this setting because it is wearable, non-invasive, privacy-preserving, and suitable for long-term monitoring.
- **The goal is timely and reliable stress detection during routine real-world operations, not only in lab tasks.**

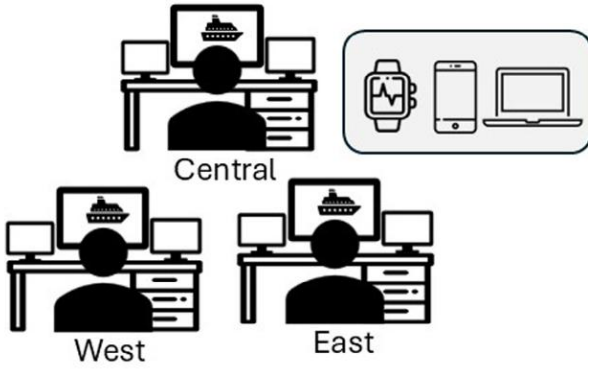


Fig. 1. Experiment setup.

Research gaps and main contributions

Gaps in prior work

- Most datasets are collected in controlled lab settings with artificial stressors.
- Lab datasets often miss motion artifacts, imbalanced stress labels, individual differences, and cross-day drift.
- VTS stress depends on operational context, especially shift and sector.

Contribution of this paper

- First on-site, long-term stress dataset from real VTS operators during routine work.
- Multi-source PPG representation: raw PPG + temporal HRV + global HRV.
- Cross-attention Transformer to fuse physiological inputs.
- Context-aware Mixture-of-Experts (MoE) using shift and sector.
- Cross-subject and cross-day adaptation for new operators and shifts.

Dataset and experiment protocol

- Location: Singapore Port Operations Control Center.
- Participants: 8 licensed front-line VTSOs initially; 7 completed the study.
- Duration: 5 months, 89 shifts in total.
- Shifts: 32 morning, 29 day, 28 night; approximately 6 hours each.
- Sectors: west, central, and east workstations.
- Sensor: Polar Verity Sense PPG at 55 Hz; worn below the elbow on the inner side. (3 signal channels + 1 ambient channel)
- Labeling: every 15 minutes, operators rated stress over the previous 5 minutes.

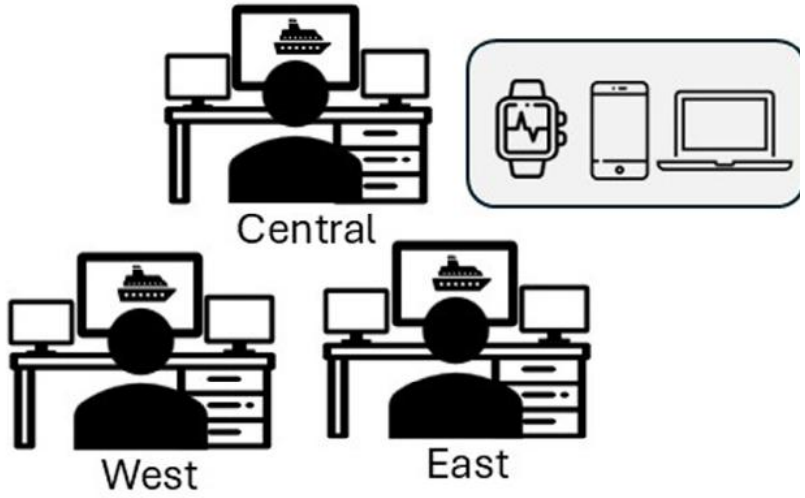


Fig. 1. Experiment setup.

Stress labels were designed to be simple to minimize interruptions during VTS operations.

Dataset challenge: imbalance and variability

- Total stress responses: 1,249.
- Stress ratings were highly imbalanced: Level 0 dominated; Level 4 did not appear.
- The authors converted labels to two classes: Level 0 = no stress; Levels > 0 = stress.
- Even after binarization, the data stayed imbalanced: 849 no-stress vs. 443 stress labels.
- Subjects showed different stress distributions; for example, Subj. 2 & 3 are almost all Level 0, while Subj. 4 sits mostly at Levels 2–3.

Table 1
Summary of stress questionnaire.

Subject ID	Level 0	Level 1	Level 2	Level 3	Level 4
1	85	176	5	0	0
2	200	16	5	0	0
3	163	9	2	0	0
4	1	14	75	12	0
5	48	3	1	0	0
6	267	56	2	0	0
7	85	20	4	0	0
Sum	849	294	94	12	0
Total	1249				

Data processing and input representation

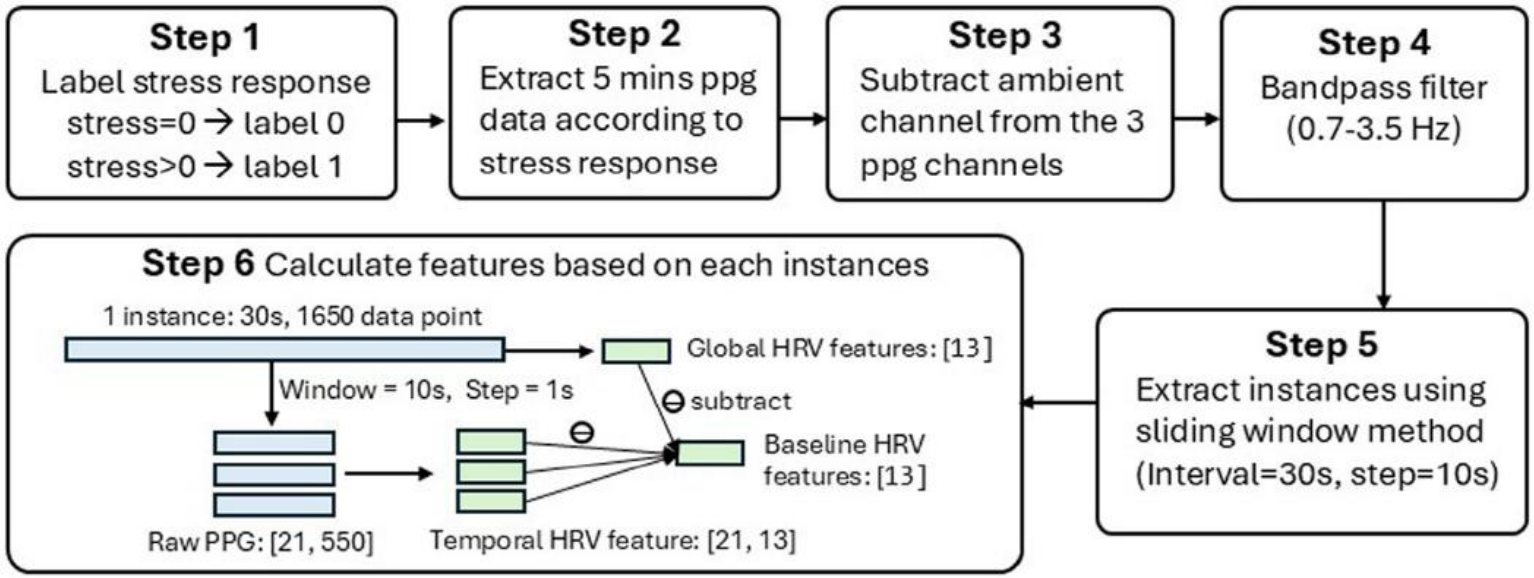
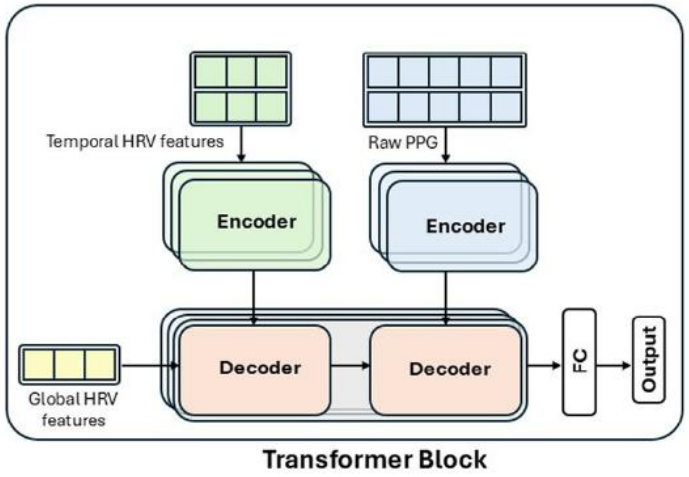


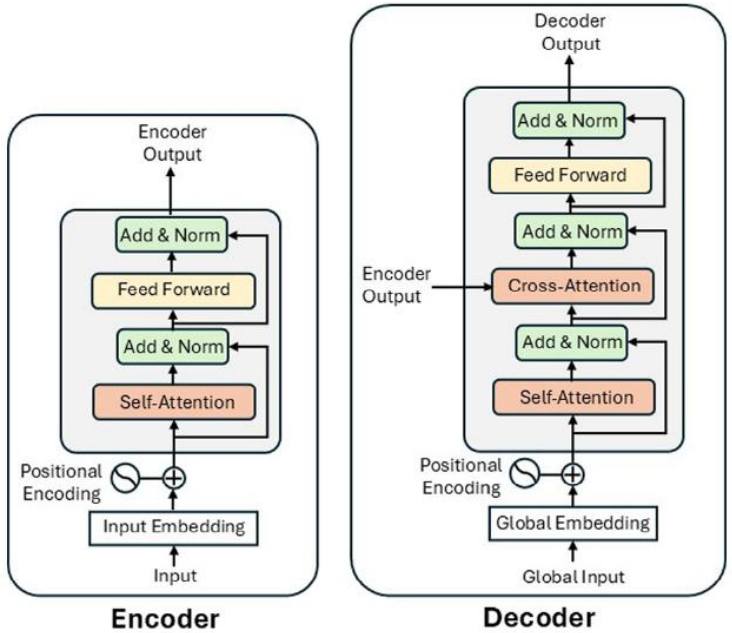
Fig. 2. Data processing.

- Each questionnaire label was paired with the previous 5 minutes of PPG.
- Ambient signal was subtracted from the three PPG channels; then a 0.7-3.5 Hz bandpass filter was applied.
- 30-second windows with 10-second step produced 75,312 labeled instances.
- Final model inputs: raw PPG [21,550], temporal HRV [21,13], and global HRV [13].
- The **baseline** is the resting physiological recording collected at the beginning of each operator's shift, when participants stayed at rest for 30 seconds.

Cross-attention Transformer: how the fusion works



Transformer Block



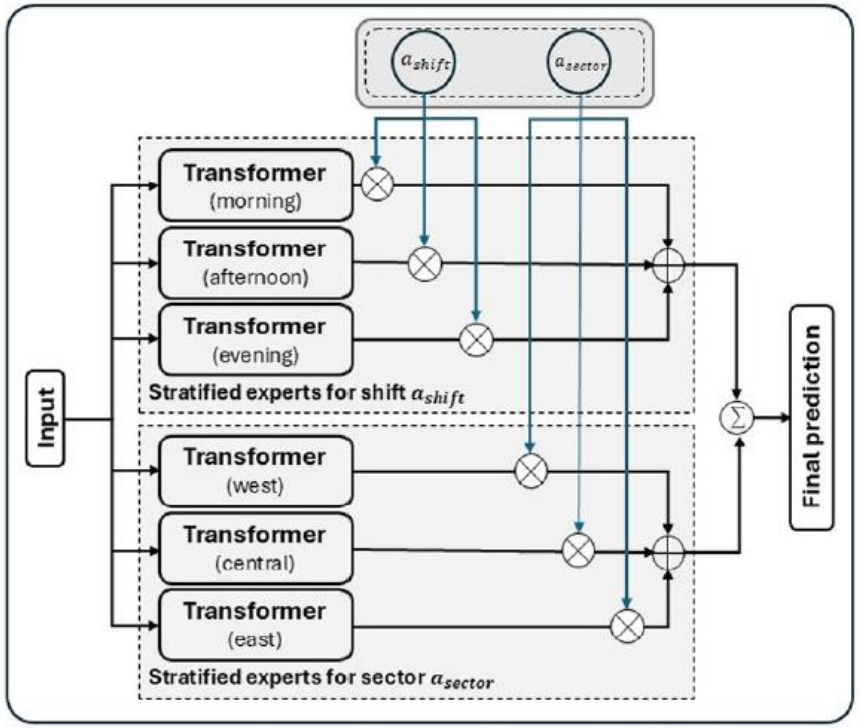
$$x_{\text{global}}^{(k)} = \begin{cases} \text{CrossAttention}(x_{\text{g\_hrv}}^{(k-1)}, E_{\text{ppg}}) & \text{if } k \text{ is odd} \\ \text{CrossAttention}(x_{\text{g\_hrv}}^{(k-1)}, E_{\text{t\_hrv}}) & \text{if } k \text{ is even} \end{cases}$$

For each decoder layer $k = 1, 2, \ldots, 2M$

- Raw PPG and temporal HRV are processed by separate encoders.
- Encoders first create two encoded representations: E_ppg from raw PPG and E_t_hrv from temporal HRV.
- The decoder starts with global HRV features (x_g_hrv); each layer uses the previous decoder output as the query.
- Odd decoder layers (k = 1, 3, 5, ...): query attends to encoded raw PPG, E_ppg, as the key/value input.
- Even decoder layers (k = 2, 4, 6, ...): query attends to encoded temporal HRV, E_t_hrv, as the key/value input.
- After all 2M decoder layers, the fused representation goes to the FC layer for stress / no-stress prediction.

Context-aware MoE: adapting to shift and sector

- Motivation: VTS stress patterns vary by operational context.
- The paper uses a Mixture-of-Experts framework with six Transformer-based experts.
- Shift experts: morning, day, night.
- Sector experts: west, central, east.
- The one-hot gate activates the relevant shift expert and sector expert for each sample.
- Example: morning + central activates the morning expert and central-sector expert.



Training and evaluation setup

Model configuration

- Hidden dimension: 128.
- Feedforward dimension: 1024.
- Attention heads: 2.
- Encoder layers: 1; decoder layers: 2.
- Two-class output: stress vs. no stress.

Evaluation design

- 5-fold cross-validation.
- Segments from the same 5-minute PPG signal were kept only in train or test, avoiding data leakage.
- Optimizer: AdamW, initial LR 0.001.
- Batch size 256, 250 epochs, cross-entropy loss.
- Metrics: accuracy, precision, recall, and F1; recall is especially important for safety.

Key evaluation point: because missing a stressed operator is risky, recall is more meaningful than accuracy alone.

Main results: proposed model vs. baselines

Table 2
Model comparison.

Methods	Accuracy	Precision	Recall	F1
Ours	85.43	84.33	84.74	84.50
Bispectrum [17]	84.41	81.92	82.81	82.32
Random Forest (RF) [27,28]	85.72	86.04	75.59	80.47
AdaBoost (AB) [27]	83.31	81.26	74.28	77.61
CNN-LSTM [14]	72.40	70.01	70.97	70.41
CNN-MLP [50]	71.24	67.08	70.40	67.58
Decision Tree [27]	78.54	73.98	69.24	71.53
1D CNN [13]	69.46	65.79	67.91	66.17
GRU [13]	63.93	53.66	59.05	49.99
LSTM [13]	63.54	53.50	58.06	50.06
Naive Bayes [25]	43.50	36.73	62.43	46.25
KNN [25,27,51]	68.84	60.91	55.79	58.23
SVM [25,27,28,51],	71.96	75.25	41,74	53.70

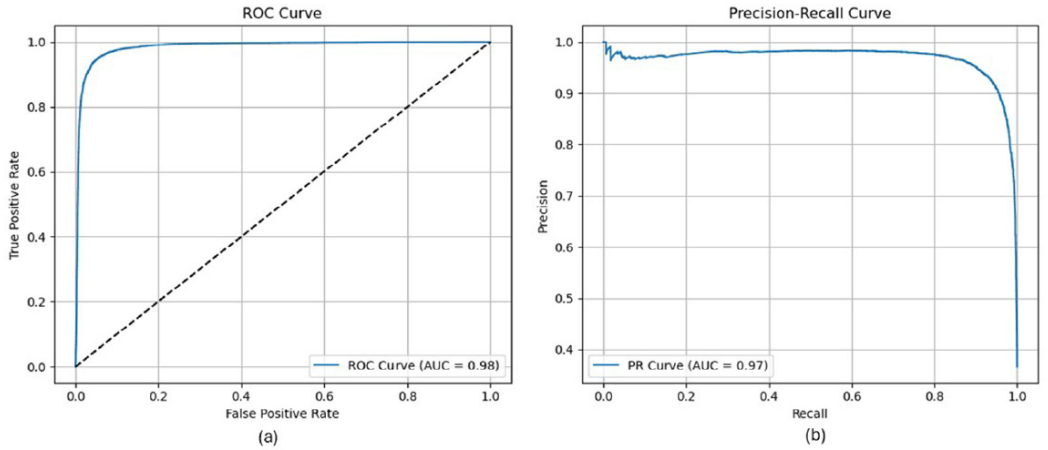


Fig. 5. (a) ROC Curve and (b) Precision–Recall Curve of the proposed model.

- Proposed model: 85.43% accuracy, 84.50% F1.
- Recall was 84.74% in Table 2, meaning the model detected most stress cases.
- ROC AUC = 0.98 and PR AUC = 0.97, showing strong performance across thresholds.
- Random Forest had similar accuracy but lower recall (75.59%), which is less desirable for safety-critical stress detection.
- CNN-LSTM and other deep models performed lower, suggesting raw PPG alone is not enough here.

Ablation study: what components matter?

Table 3
Abalation study.

	Accuracy	Precision	Recall	F1
transformer (g_hrv+t_hrv+ppg)+MoE	85.43 ± 1.42	84.34 ± 1.47	84.74 ± 1.23	84.50 ± 1.36
transformer (g_hrv+t_hrv+ppg)	82.78 ± 80.87	81.45 ± 0.77	81.81 ± 0.64	81.67 ± 0.71
transformer (g_hrv+ppg)	77.25 ± 1.04	74.64 ± 1.02	76.32 ± 0.55	75.17 ± 0.92
transformer (g_hrv+t_hrv)	79.33 ± 0.93	77.13 ± 1.08	78.72 ± 1.24	77.56 ± 0.86
transformer (t_hrv)	75.05 ± 1.51	71.82 ± 0.86	74.45 ± 2.66	72.40 ± 1.02
transformer (ppg)	69.77 ± 1.71	66.08 ± 1.39	67.97 ± 1.18	66.43 ± 1.47
transformer (g_hrv)	74.99 ± 1.15	71.88 ± 1.58	73.91 ± 1.34	72.43 ± 1.49

Table 3: removing MoE or input modalities reduces performance

Interpretation

- Full model (global HRV + temporal HRV + raw PPG + MoE) achieved the best F1: 84.50.
- Removing MoE reduced F1 to 81.67, showing the benefit of shift/sector adaptation.
- Single-source inputs were weaker: raw PPG only reached 66.43 F1; global HRV only reached 72.43 F1.
- The combination of raw signal + temporal HRV + global HRV captures complementary stress information.

Cross-subject and cross-day adaptation

- Baseline leave-one-subject-out generalization was poor: 60.02% accuracy and 44.93% F1.
- Adaptation strategy: pretrain on other subjects, fine-tune with **20% of the target** subject data, then test on the remaining 80%.
- Fine-tuning used only **5 epochs**.
- After adaptation: **92.12%** accuracy, 82.45% precision, 75.68% recall, **76.69% F1**.
- Subjects with very few stress labels still had lower recall, showing the need for balanced personalization data.

Table 4
Cross-subject adaptation.

Subject ID	Accuracy	Precision	Recall	F1
1	84.96	82.03	81.39	81.67
2	99.30	98.33	98.53	98.42
3	94.00	76.80	58.71	62.21
4	99.16	90.02	77.48	78.29
5	90.43	63.41	52.70	52.86
6	86.08	78.64	74.68	76.29
7	90.94	88.05	86.29	87.05
mean	92.12 ± 5.30	82.45 ± 10.36	75.68 ± 14.58	76.69 ± 14.06

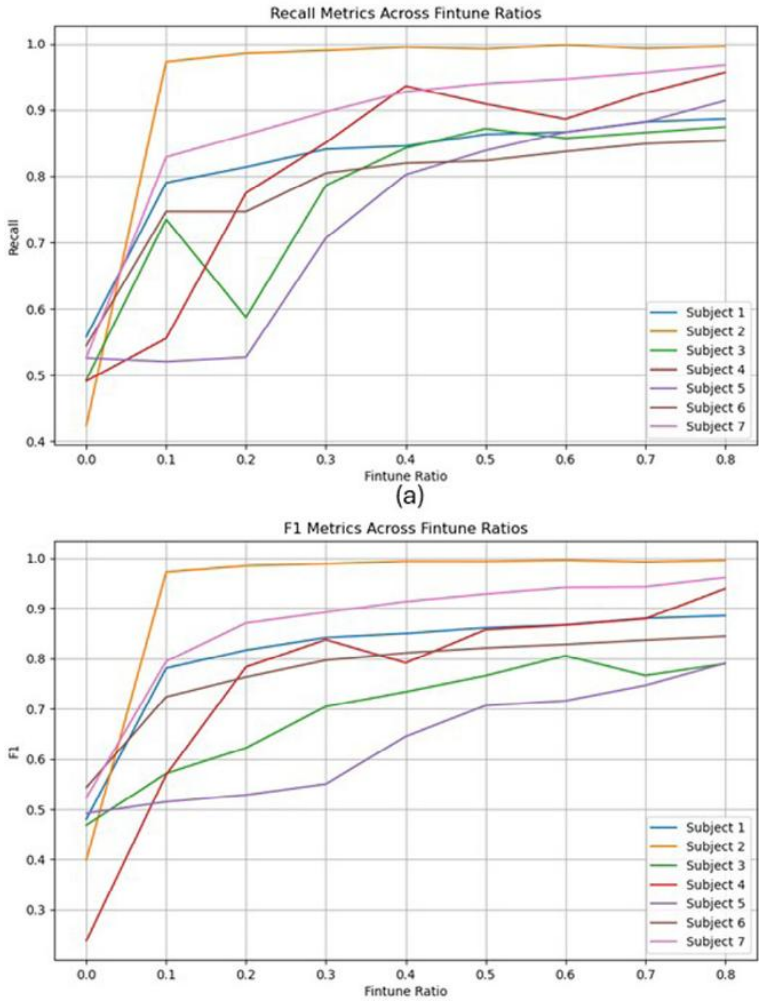


Table 4 and Fig. 10: adaptation results and fine-tuning ratios

Interpretability: cross-attention visualization

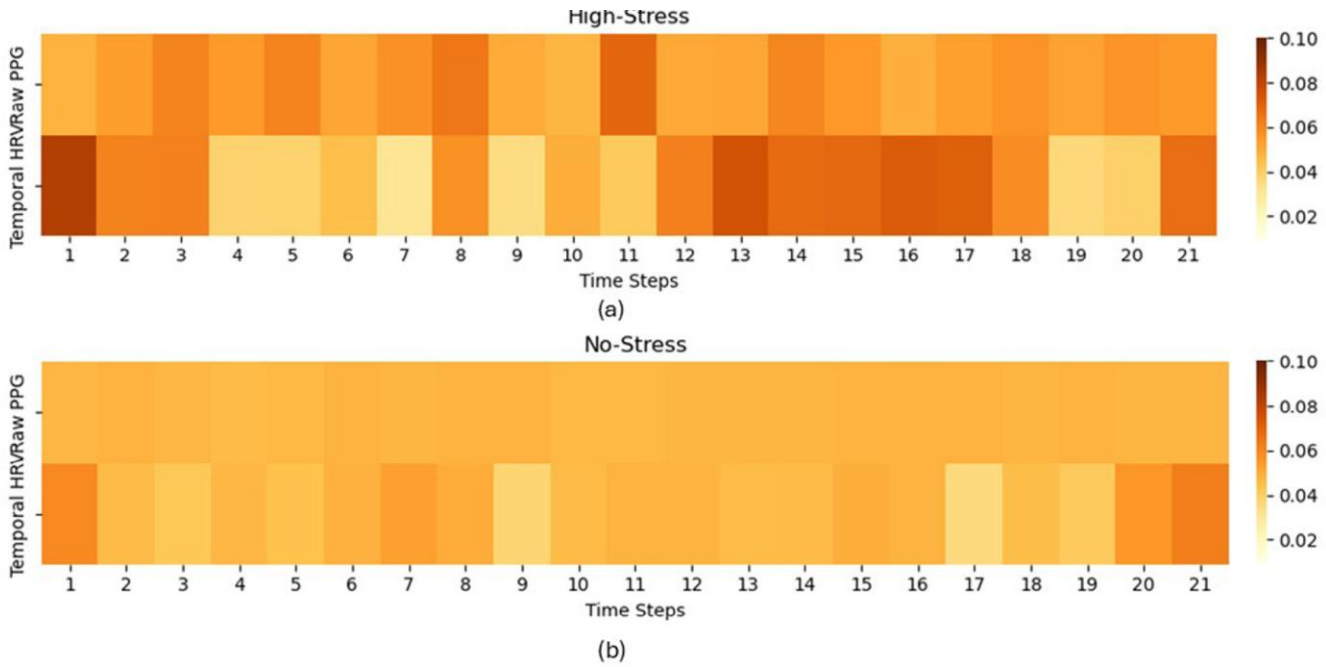


Fig. 9. Cross-attention weight visualization for correctly predicted (a) high-stress and (b) no-stress samples.

- X-axis: 21 time steps inside each 30-second sample.
- Rows: temporal HRV and raw PPG streams.
- Darker color means stronger model attention.
- High-stress sample: attention is sharper and concentrated at specific time steps in both PPG and HRV.
- No-stress sample: attention is more evenly distributed and mostly focused on HRV.

Takeaway: the model provides some temporal and modality-level explanation for its predictions.

Limitations future work

- Few extreme stress cases; no Level 4 labels appeared in the dataset.
- Self-reported stress may be underreported because operators may worry about being evaluated.
- The general stress questionnaire may not fully capture VTS-specific stress expressions.
- Small participant pool limits demographic modeling.
- Supervised fine-tuning requires labels; future work could explore emergency drills, individual-aware MoE, continual learning, meta-learning, KFT, and RLHF.

Summary / conclusion

Key findings

- Real-world VTS stress detection is feasible using wearable PPG data.
- The model combines raw PPG, temporal HRV, and global HRV instead of relying on one signal view.
- Cross-attention helps fuse physiological sources and capture both raw patterns and HRV dynamics.
- Context-aware MoE improves detection by adapting to operational shift and sector.

Takeaways for discussion

- Best overall performance: 85.43% accuracy and 84.50% F1-score.
- Personalization is important: fine-tuning with 20% target-subject data raised adaptation accuracy to 92.12%.
- The main limitation is limited high-stress data and reliance on self-reported stress labels.