Personalized Multimodal Monitoring for Detecting High-Risk Human Performance States Under Workload

Yuehai Yu[1], Joel Martin[2], Megan Sax van der Weyden[2], Mohammad Fili[1], Lizhi Wang[3], Craig Yu[1], Guiping Hu[1] , Harry Wang

[1] Department of Systems Engineering and Operations Research, George Mason University, Fairfax, VA, USA
[2] Sports Medicine Assessment, Research & Testing (SMART) Laboratory, George Mason University, Fairfax, VA, USA
[3] Department of Computer Science, George Mason University, Fairfax, VA, USA

Introduction

Operational workload can affect thinking and physical performance in different ways. Slower responses or missed signals may appear before a large decline is visible in average performance.

Purpose

Evaluate whether changes from each participant's own baseline can identify workload conditions and periods of elevated performance risk.

Materials & Methods

- 12 participants (9 men, 3 women; age 31.3 ± 10.1 years) completed march-only (STP), thinking-task-only (STC), and combined loaded-march plus thinking-task (DT) conditions.
- STP and DT included a 60-minute loaded march at 4.8 km/h with a 30-kg rucksack; DT added Friend-or-Foe target detection.
- Thinking, physical, body-response, and perceived-workload measures were expressed as changes from each person's normal starting point.
- Leave-one-subject-out (LOSO) cross-validation trained on 11 participants and tested the 12th; risk definitions from 0.5 to 2.0 standard deviations (SD) were examined.

Results

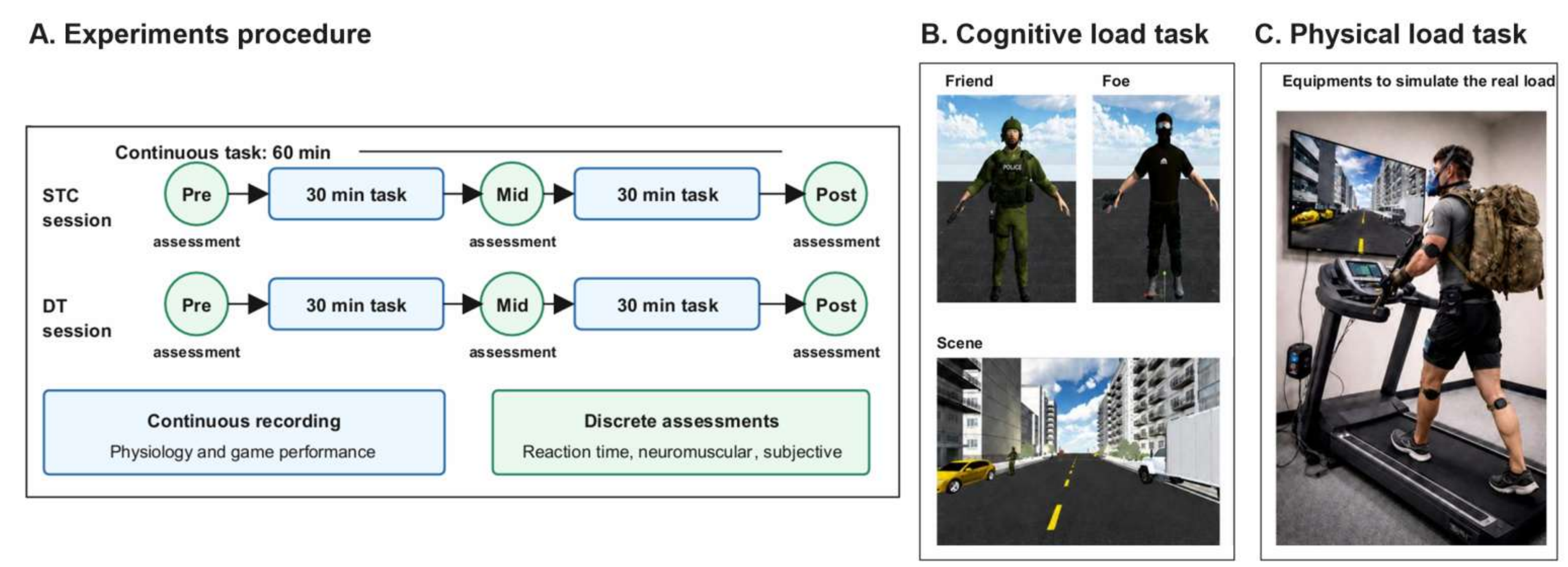
- Area under the curve (AUC) values for the three workload states ranged from 0.927 to 1.000.
- At 1.0 SD: sensitivity 97.4%, specificity 73.9%, and accuracy 82.4%.
- Precision was 67.9% and F1 was 80.0%; the model produced 18 false alerts and missed 1 high-risk observation.
- Sensitivity remained 91.9%–100.0% across four high-risk definitions.

Conclusions

- **Individual baselines can support early reassessment and workload decisions.**
- **Use the alert to trigger reassessment or workload adjustment—not to replace leader judgment; validation in new operational settings is still required.**

BLUF: Individual baselines helped distinguish workload conditions and identify nearly all periods of elevated performance risk. Because the system also produced false alerts, it should be used to support reassessment and workload decisions rather than replace leader judgment.

Procedures



Monitoring Workflow

Personal baseline > Changes in four areas > Model trained on other participants > Personal risk score > Alert setting

How the Model Was Built and Tested

1 Personal baseline

Each domain was expressed as change from the participant's own baseline instead of one cutoff for everyone.

2 High-risk definition

High risk was defined when thinking or physical performance exceeded the selected SD threshold.

3 LOSO validation

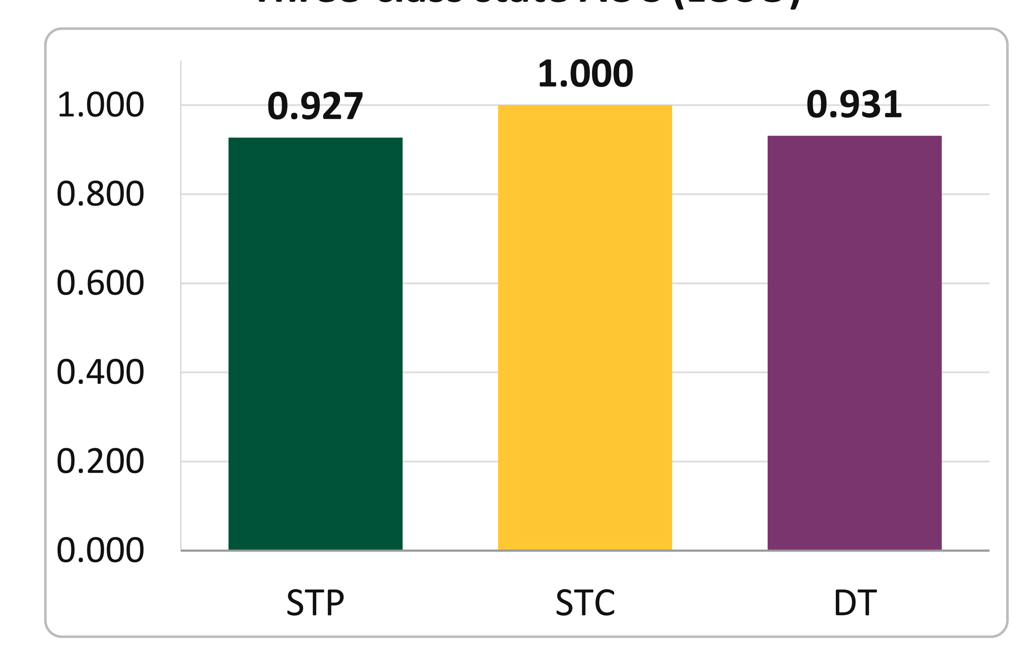
The model trained on 11 participants and then predicted the participant left out of training.

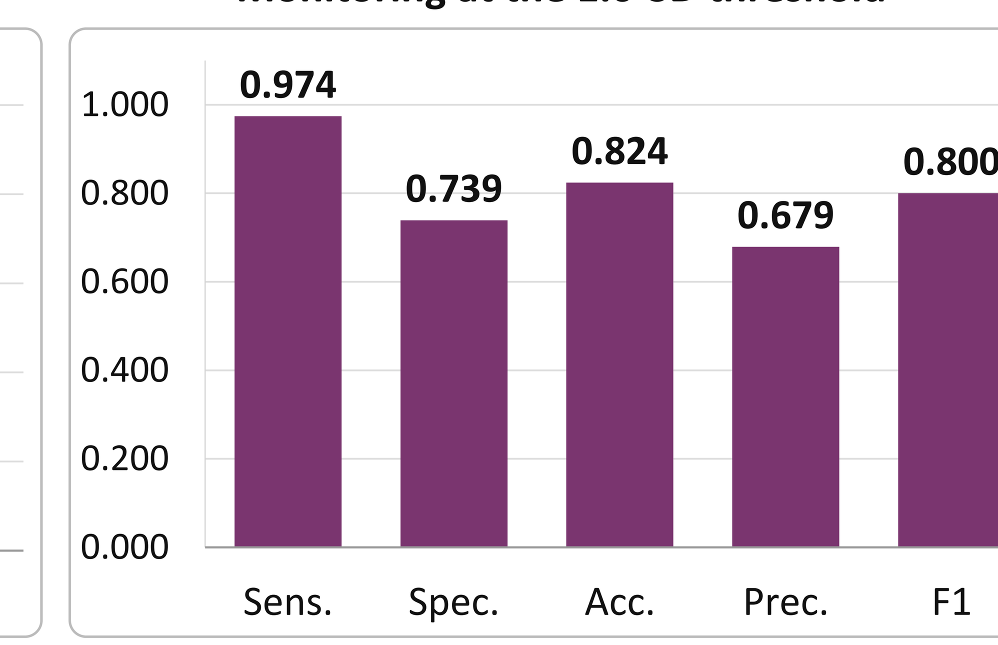
4 Alert-threshold tradeoff

Sensitivity, specificity, precision, and F1 were reported together to show detections and false alerts.

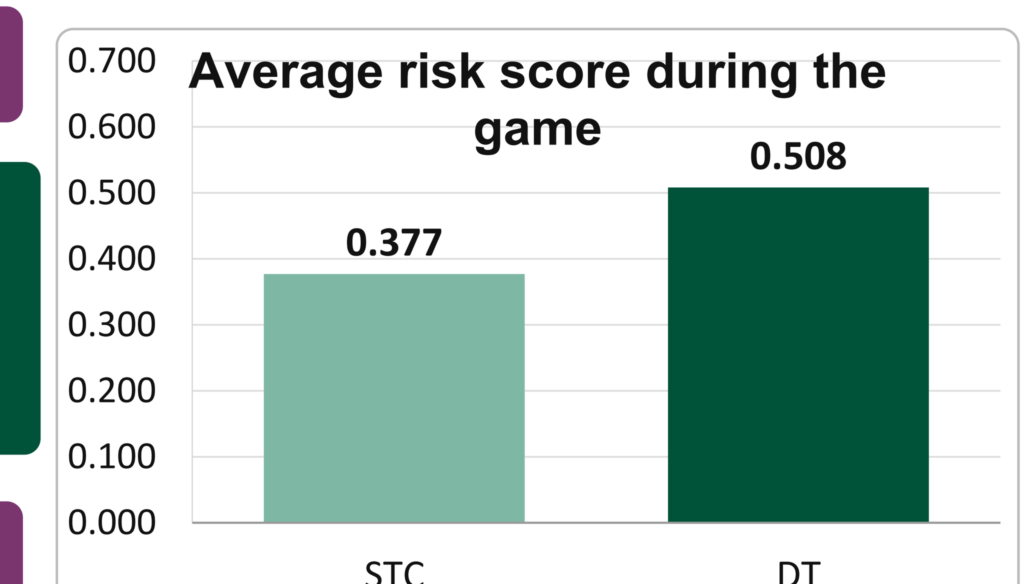
Knowing the workload condition is not the same as knowing whether performance is at risk.

Results





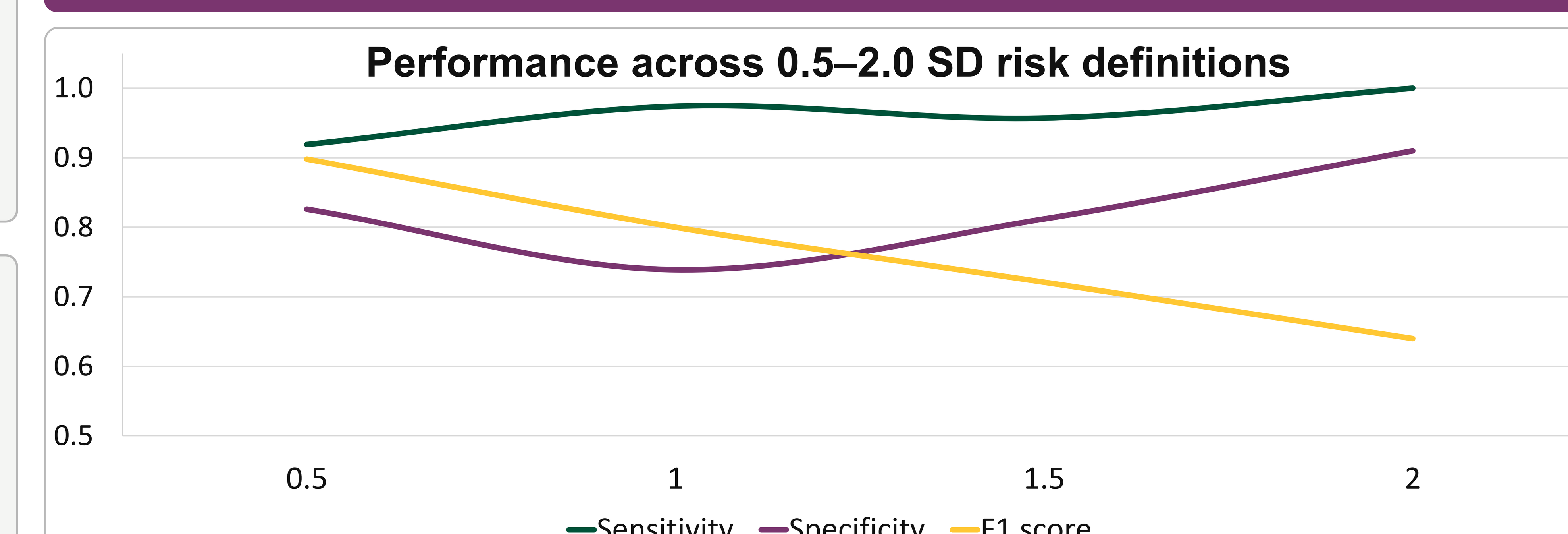
AUC: 1.0 = perfect; 0.5 = chance. DT vs. STP AUC = 0.861 (95% confidence interval 0.689–0.990).
Sensitivity = risk detected; specificity = non-risk dismissed; precision = true alerts; F1 balances sensitivity and precision.



Primary operating point

108 observations from 12 participants
39/108 were labeled high risk (36.1%)
Alert threshold = 0.322
38/39 high-risk observations detected
18 false alerts | 1 missed event

Sensitivity Analysis Across Risk Definitions



Sensitivity remained 91.9%–100.0%. F1 decreased from 0.898 to 0.640 as high-risk events became rarer, showing how threshold choice changes the false-alert burden.