Feasibility of a Personalized Stress Appraisal Monitoring Technique in Construction Using Wearable PPG and Physiology-Informed Hierarchical Machine Learning

Francis Xavier Duorinaah[1], Ghanim Saqib[1], Vicente A. Gonzalez[2], Gaang Lee[1]

[1]Empathetics Lab, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada
[2]Infrastructure and Human Tech (IHT) Lab, Strategic Projects Insight Centre in Engineering (SPICE), Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada
duorinaa@ualberta.ca, gsaqib@ualberta.ca, vagonzal@ualberta.ca, gaang@ualberta.ca

Abstract – Worker stress is a primary contributor to accidents on construction sites. While wearable biosensors enable continuous stress monitoring, existing approaches treat stress either as a binary problem or a flat level-classification task, overlooking the distinction between two stress appraisals (i.e., challenge and threat) and the hierarchical nature of the stress appraisal process and its reflection on the physiological mechanisms. To address these gaps, this study proposes a two-stage hierarchical classifier aligned with the stress appraisal theory, where Stage A distinguishes "low stress" class from "stress", and Stage B differentiates challenge (positive stress) from threat (negative stress). To address inter-subject physiological variability, the framework incorporates a learnable subject calibration layer and subject-specific fine-tuning. The model was evaluated using photoplethysmography (PPG) data from eight participants under psychological and physiological stress induction protocols, with a multi-dimensional set of features extracted to capture the cardiovascular responses specific to each appraisal state. The hierarchical model achieved average macro-F1 scores of 0.81 and 0.79, outperforming a standard 3-class model (0.78 and 0.72). Optimal model performance was achieved with only 240-320 seconds of fine-tuning data per class. These findings demonstrate the feasibility of personalized stress monitoring for proactive construction safety interventions.

Keywords –

Photoplethysmography (PPG), Stress appraisal, Hierarchical classification, Construction safety

1 Introduction

Occupational stress is a major concern in the construction industry, where workers routinely face physically demanding tasks, tight deadlines, and hazardous conditions [1]. Excessive stress levels are associated with unsafe behavior [2], which is a primary cause of 80-90% of construction accidents [3]. It also reduces productivity by influencing psychological dimensions such as job satisfaction and worker motivation [4]. Beyond the safety and productivity impacts, chronic stress has extensive health implications, often leading to psychological and systemic disorders such as depression and cardiovascular diseases [5]. Given these consequences, implementing robust stress monitoring and management techniques is vital.

Traditional stress monitoring has relied on standardized self-report instruments such as the Perceived Stress Scales [6] and the Stress and Adversity Inventory [7]. While these surveys offer validated metrics for psychological stress assessment, they are not suitable for continuous stress monitoring in dynamic work environments because they are intrusive.

Recent advancements in wearable biosensors, specifically devices integrating photoplethysmography (PPG), electrocardiogram (ECG), and electrodermal activity (EDA) sensors, have offered a promising alternative for continuous, less-intrusive stress monitoring. These devices are particularly effective because they capture the distinct physiological patterns associated with stress [8]. Several studies have successfully applied these technologies to detect construction workers' stress levels in real-time, establishing the technical feasibility for continuous stress monitoring using biosensors [9–11].

Despite the promising results of simple biosensor-based stress detection, most studies treat stress classification as a binary task, and simply classify workers as either being "stressed" or "not stressed". This simplification is insufficient for effective on-site interventions as it fails to distinguish between the specific stress types, which have different impacts on worker performance and safety. Stress conditions can be classified into two categories: challenge and threat. The

Transactional Model of Stress and Coping [12] provides a framework for understanding challenge and threat, as well as how these conditions arise. The model conceptualizes stress as a result of cognitive appraisal processes which arise from the interaction between individuals and their environment. The model establishes that stress appraisal occurs through two sequential stages: primary and secondary appraisal. During primary appraisal, individuals evaluate a situation's relevance and potential harm. When deemed relevant, they engage in secondary appraisal where they assess their available coping resources such as knowledge, skills, and social support, against the perceived demands of the situation [12]. This two-stage appraisal process determines the nature of the stress response. When individuals perceive their resources as adequate to meet demands, the situation is appraised as a challenge, which is positive and leads to higher task engagement and motivation. Conversely, when demands are judged to exceed available resources, a threat appraisal occurs, which has negative impacts and leads to physiological strain, lethargy, depression, and reduced task focus. As such, distinguishing between these stress types is more beneficial for safety management since it provides an opportunity to put in interventions for threat conditions while leveraging the positive outcomes of challenge.

Beyond the theoretical distinction between challenge and threat, these conditions produce subtle but distinct autonomic and neuroendocrine responses which can be captured by wearable biosensors. The Biopsychosocial Model of Challenge and Threat (BPSM) [13] outlines that while both conditions activate the Sympathetic-Adreno-Medullary (SAM), resulting in increased heart rate and arousal [14], they diverge in their underlying hemodynamics. Specifically, threat appraisals uniquely activate the hypothalamic-pituitary-adrenal (HPA) axis, which triggers cortisol release and increases total peripheral resistance (TPR) through vasoconstriction [15]. In contrast challenge states produce a more efficient cardiovascular profile characterized by increased cardiac output and reduced TPR, reflecting physiological mobilization through vasodilation [16]. These distinct patterns allow for real-time differentiation of stress appraisal types using physiological signals such as PPG. By optically detecting blood volume fluctuations, PPG enables extraction of specific metrics such as pulse morphology and acceleration photoplethysmography (APG) [17], which are indicators of TPR. Beyond its physiological potential to distinguish threat and challenge, PPG is also highly suitable for field deployment because it is less-intrusive and widely available in commercial wearable devices.

Although biosensors offer the potential to identify specific stress types, very few studies have attempted to make this distinction [18,19], and these existing studies possess some limitations. Current approaches have either framed the low stress, challenge, and threat classification task as a standard three-class problem, treating each class as an independent and flat category [18], or have focused exclusively on the binary distinction between threat and challenge while ignoring low stress [19]. This formulation is not aligned with the theoretical foundations and physiological mechanisms established by the Transactional Model [12]. The model establishes that the stress appraisal process is hierarchical, which is reflected in the sequential activation of the SAM and HPA axis. Low stress involves no SAM activation, representing a clear baseline. However, when primary appraisal identifies a situation as relevant, SAM is triggered, resulting in broad arousal responses common to both challenge and threat. The distinction between threat and challenge occurs in the secondary appraisal; If resources are sufficient, the response is predominantly SAM-driven (challenge), but if demands exceed resources, the HPA is engaged in addition to the SAM, triggering threat. Consequently, separating low stress from challenge or threat is comparatively straightforward, whereas separating challenge from threat is more subtle because it depends on HPA-related effects conditional on existing sympathetic activation. Flat classification architectures do not reflect this conditional structure, which can limit performance by making the model biased toward the easier low stress versus stress separation while reducing sensitivity to the subtle physiological patterns that differentiate challenge from threat. Furthermore, while physiological signal-based stress detection systems are known to suffer from inter-subject variability due to differences in physiological profiles [20], the prior stress appraisal studies have not addressed this issue. This gap is particularly critical because the physiological distinctions between challenge and threat, including vascular resistance patterns, cardiac efficiency, and HPA axis activation, may exhibit greater inter-individual variability than general stress detection, which is often driven primarily by broad arousal patterns [21]. Without strategies to manage this variability, the practical deployment of stress appraisal detection frameworks in real construction environments remains difficult.

To address these limitations, this study presents a hierarchical deep learning framework that operationalizes the sequential nature of the stress appraisal theory. The proposed architecture decomposes the classification into two stages during both training and inference. Stage A distinguishes "low stress" from "stress" (challenge and threat combined) and is trained on the full dataset. Conversely, Stage B is trained and evaluated exclusively on stress samples to differentiate between challenge and threat. This design enables stage-specific optimization to address the inherent challenge of discriminating the subtle cardiovascular distinctions between challenge and threat states. To address the issue of subject physiological variability, a learnable calibration layer is incorporated into each classification stage, and subject-specific fine-tuning is performed. This study advances toward field-deployable stress appraisal monitoring systems for construction safety applications where early detection of threat responses could inform targeted interventions to reduce negative impacts.

2 Methods

To develop a hierarchical classifier capable of distinguishing stress appraisal states, four methodological steps were followed, as shown in Figure 1. PPG data were collected in controlled laboratory experiments to obtain reliable recordings under low stress, challenge, and threat conditions with validated ground-truth labels required for training the proposed model. This was followed by standard data preprocessing, data labeling, feature extraction, and finally training of the proposed hierarchical classifier

2.1 Data Collection

2.1.1 Participants

Data were collected from 8 participants (5 males and 3 females; age range: 23-34) who reported good physical and mental health, with no history of cardiovascular or psychiatric conditions. As an initial feasibility study, this limited sample size was selected to examine whether low stress, challenge, and threat-related physiological patterns could be identified under laboratory conditions before broader field validation. Participants without prior construction experience were recruited, and while this limits direct generalizability of findings to construction sites, it ensured that appraisal-related physiological responses were not influenced by confounding factors such as occupational familiarity or site-specific stress adaptation. All experimental protocols were approved by the University of Alberta's Ethics Board (Pro00137747).

2.1.2 Experimental Procedures

A controlled laboratory experiment was conducted with protocols designed to elicit low stress, challenge, and threat across multiple trials while PPG signals were recorded. Each participant completed the protocol twice on separate days, with different threat-induction methods used on each day. This multi-session design was implemented to ensure the recorded stress patterns were a result of experimental protocols, while mitigating potential biases arising from other physiological stresses. PPG data were recorded using the Polar Verity Sense armband sensor, which features three PPG channels and a 55 Hz sampling rate.

Each experiment session commenced with a 20-minute rest period, during which participants relaxed while listening to neutral nature sounds. This phase served as a physiological baseline (low stress) and helped minimize external confounding stress responses unrelated to the experiment. Following relaxation, participants performed ten task trials per session: five designed to elicit challenge appraisals and five targeting threat appraisals, with each trial lasting 2 minutes. In the challenge condition, participants transported a 10 kg ergonomic weight between two marked points 12 meters apart. During challenge trials, researchers maintained neutral body language and provided no performance feedback. In contrast, threat sessions involved the same material handling task but integrated additional stressors to induce threat. On Day 1, a psychological threat condition was implemented by requiring strict adherence to specified ergonomic rules (such as maintaining elbow proximity and avoiding trunk rotation) while providing timed, standardized critical feedback every 30 seconds emphasizing deviations from the required technique. Participants were also informed that an 80% success rate was required across trials, consistent with procedures intended to elicit social-evaluative stress and psychological strain [22]. On Day 2, a physiological threat condition was induced using a cold-pressor arm-wrap maintained at 0–5°C. Cold-pressor exposure is widely used to elicit threat-related responses because it disrupts physiological homeostasis and reliably triggers stress reactivity [15,22].

To validate the appraisal states elicited by the experimental conditions, two self-report instruments were administered: the Self-Assessment Manikin (SAM) and the Cognitive Appraisal Ratio (CAR) [23,24]. SAM was used to quantify valence and arousal, which are

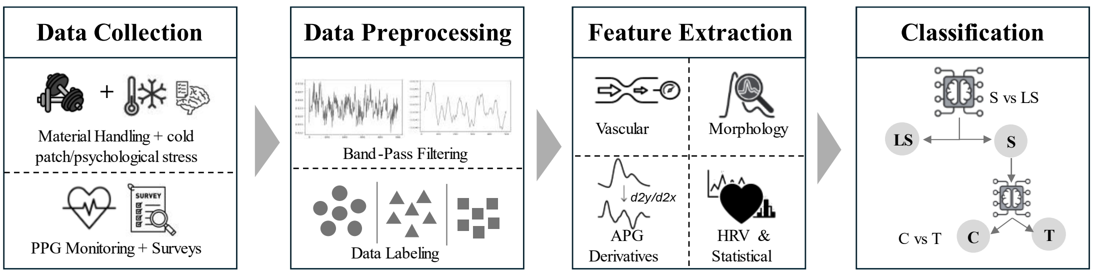

Figure 1. Proposed stress appraisal monitoring framework

relevant to differentiating challenge and threat appraisals. CAR was used to compute perceived demands and perceived coping resources, such that challenge is expected when perceived resources exceed demands, whereas threat is expected when perceived demands exceed resources. CAR was administered immediately before each trial to capture anticipatory appraisals, and SAM was administered during recovery periods to capture post-trial valence and arousal [23,24].

2.2 Data Preprocessing

A multi-stage data preprocessing procedure was applied to attenuate motion artifacts and sensor noise while preserving physiologically relevant PPG components. Signals were filtered using a fourth-order Butterworth bandpass filter (0.5–5.0 Hz) to suppress both low-frequency baseline drift and high-frequency noise. This was followed by a Gaussian smoothing step to further reduce residual high-frequency fluctuations which can destabilize derivative-based feature extraction [25].

To ensure that the final dataset represented the intended appraisal states, a consistent-label filtering step was implemented. This step cross-referenced the experimental condition labels with participants' self-reports and retained only windows where the intended trial condition (low stress, challenge, or threat) was fully consistent with both the CAR and SAM responses. This ensured that the samples used for training and evaluation cleanly represented the target stress conditions.

2.3 Feature Extraction

Following preprocessing and labeling, a multi-dimensional feature set was extracted from the filtered PPG signals. The feature pipeline was designed to move beyond conventional heart rate variability (HRV) indices, which primarily reflect autonomic arousal but are insufficient for separating challenge from threat appraisals. According to the BPSM [13], the key cardiovascular distinction between challenge and threat is TPR, defined as the resistance to blood flow within the blood vessels [26]. Although gold-standard direct TPR estimation requires continuous blood pressure monitoring (often combined with ECG) [27], studies suggest that specific PPG-derived features, particularly waveform morphology and acceleration photoplethysmography (APG), the second derivative of the PPG waveform, have some correlation with TPR and can serve as indirect proxies [17,28,29]. Accordingly, in addition to HRV features, the extracted feature set in this study included APG-derived descriptors, pulse morphology measures, and vascular resistance proxies such as the reflection index and dicrotic notch (Table 1).

Features were extracted from filtered PPG signals using 17-second windows with 75% overlap to preserve temporal continuity. Within each window, the smoothed PPG waveform was differentiated to obtain the velocity (VPG) and acceleration (APG) signals. Systolic peaks were identified using rule-based criteria on minimum distance, prominence, and height, and troughs were detected from the inverted waveform. Windows were discarded when signal variation was insufficient (std < 0.01), when more than 10% of samples were missing, or when the estimated heart rate fell outside 40–180 bpm. Beat-level features were then computed for each valid pulse cycle and averaged within the window.

Table 1. Summary of Extracted Features

Feature Category	Specific Features
Morphological Features	Pulse amplitude, systolic/diastolic ratio, upstroke slope, pulse area
APG Derivatives	a-peak, b-peak, b/a ratio, e peak, aging index
Vascular and hemodynamic	Perfusion Index, PTT (proxies), dicrotic notch (time/prominence), stiffness index, augmentation index, compliance
HRV time domain	Heart rate, RMSSD, SDNN, NN intervals, pNN50
Statistical/Poincare	SD1, SD2, Skewness, Kurtosis

2.4 Classification framework

2.4.1 Architecture details

To operationalize the sequential nature of the stress appraisal process, a hierarchical hard-cascade classifier with stage-specific optimizations was developed (Figure 2). The architecture decomposes the three-class appraisal problem into two sequential binary decisions (Stage A and Stage B). Stage A performs the primary separation by classifying each window as either Low Stress or Stress (Challenge and Threat combined). Stage B is executed only when Stage A predicts Stress, and is trained exclusively on stress-labeled windows. Restricting Stage B to Challenge and Threat samples reduced contamination from high-variance baseline physiology and enabled specialization in the subtle hemodynamic differences associated with appraisal, rather than general arousal. Stage A and Stage B both used the same dense neural network design to model non-linear relationships in the physiological feature space, but were trained as independent models with separate Stage A and Stage B preprocessing pipelines. Each model consisted of a two-layer core feature extractor (256 and 128 units) with ReLU activation and batch normalization. A two-layer classification head (128 and 64 units) then mapped the learned representation to class probabilities via a SoftMax output. To support lightweight personalization, a learnable subject-calibration layer was inserted immediately after the input of each model. For an input vector x, the layer produced a calibrated representation x' $= \gamma \odot x + \beta$, where γ and β are trainable feature-wise scaling and shifting parameters respectively, and $\odot$ denotes element-wise multiplication. The layer was initialized as an identity transform ($\gamma = 1$, $\beta = 0$), so that it initially left the input unchanged during training and was later used to adapt the population-trained feature representations to each individual subject. Each model was trained using the Adam optimizer with an initial learning rate of 0.001. Training used a batch size of 32 for up to 50 epochs. To reduce the risk of overfitting, dropout (p=0.3) and L2 weight decay (λ=0.005) were applied throughout training. Early stopping was applied based on the validation loss with a patience of 5 epochs, and model weights from the best-performing epoch were restored. Stage A used a sparse categorical cross-entropy with cost sensitive class weighting, whereas Stage B used focal loss (γ=2.0, α=0.25) to better address the more difficult challenge-versus-threat differentiation.

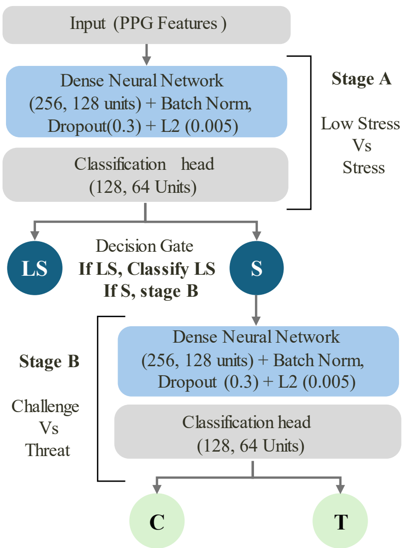

Figure 2. Hierarchical Classifier Logic (LS: Low Stress, S: Stress, C: Challenge; T: Threat)

2.4.2 Generalizability test and subject specific fine-tuning

To evaluate the baseline model generalizability and address the challenge of inter-subject variability, a rigorous two-phase evaluation framework that combines Leave-One-Subject-Out Cross-Validation (LOSO-CV) with adaptive subject-specific fine-tuning for personalization was implemented. Under LOSO-CV, Stage A and Stage B were trained on data from all subjects except one, and the held-out subject was used exclusively for testing. This process was repeated for each subject to simulate deployment on an unseen worker.

Following the baseline LOSO-CV evaluation, subject-specific fine-tuning was performed using limited calibration data from the held-out subject. To prevent temporal leakage and to evaluate personalization on future data, each subject's windows were first sorted chronologically by window start time and then divided using a strict temporal split. The earliest 75% of windows formed a calibration pool, while the final 25% served as a fixed temporal holdout test set. Calibration requirements were systematically varied from 200 to 680 seconds per class by sampling from the earliest portion of the calibration pool. For each subject, the same temporal holdout test set was used across all calibration durations, enabling a controlled comparison of how calibration quantity affects performance on identical future samples. Because windows overlap by 75% (13-second overlap for a 17-second window), a temporal buffer equal to one window length (17 seconds) was inserted between the calibration pool and test-set boundary to reduce adjacency effects and eliminate overlap-driven leakage. This forward-in-time evaluation maximizes ecological validity by clearly representing typical field-deployment conditions where calibration must rely only on past data, while evaluation reflects future use. The fine-tuning process followed a two-stage technique with adaptive hyperparameters to balance personalization against overfitting. In the first stage of fine-tuning, the core feature extractor layers (Dense 1 and Dense 2) were frozen and only the classification head (128- and 64-unit layers) were updated, allowing the decision boundary to shift toward subject-specific feature distributions while preserving population-level representations. In the second stage, the core feature extractor layers were optionally unfrozen, and the model was further updated at a reduced learning rate of 5×10^{-5} selected empirically, to support subject-specific refinement while limiting catastrophic forgetting. This deeper adaptation was enabled only when calibration data reached at least 480 seconds per class, as smaller calibration sets often led to unstable updates and performance collapse. Stage 1 fine-tuning ran for up to 15 epochs, whereas Stage 2 ran for up to 40 epochs with early stopping based on validation loss. This progressive unfreezing reduced instability with small calibration sets while allowing fuller adaptation with sufficient data.

2.4.3 Standard 3-class benchmark model

A standard flat three-class classifier was trained to serve as a benchmark for the hierarchical model. This reference model used the same dense architecture and training configuration as the hierarchical model described previously. The only difference was the decision structure: the flat model performed a single-stage SoftMax classification across the three labels (Low

Stress, Challenge, Threat), treating them as independent categories without hierarchical decomposition. Subject-specific fine-tuning was also conducted using the same technique and same data durations. This controlled configuration isolated the effects of the hierarchical cascade, allowing any observed performance differences to be attributed to the sequential binary technique rather than changes in network capacity or optimization.

3 Results

Table 2 and Table 3 summarize the baseline (LOSO-CV) and fine-tuned performance for the hierarchical and flat models across the two experimental protocols. For Day 1 (psychological threat protocol), both models achieved the same baseline macro-F1 of 0.62, with accuracies of 0.65 (hierarchical) and 0.68 (flat). For Day 2 (physiological threat), baseline performance was substantially different: the hierarchical model achieved F1 = 0.59 with accuracy = 0.64, whereas the flat model obtained F1 = 0.54 with accuracy = 0.66.

After subject-specific fine-tuning, both architectures improved substantially. Under the Day 1 protocol, the hierarchical model reached F1 = 0.81 and accuracy = 0.85, representing an absolute gain of ΔF1 = 0.19 relative to baseline. The flat model improved to F1 = 0.78 and accuracy = 0.82 (ΔF1 = 0.16). Under the Day 2 protocol, the hierarchical model achieved F1 = 0.79 and accuracy = 0.82 (ΔF1 = 0.20), while the flat model reached F1 = 0.72 and accuracy = 0.77 (ΔF1 = 0.18). Overall, the hierarchical architecture consistently produced higher fine-tuned performance, exceeding the flat model by 3 percentage points on Day 1 (0.81 vs. 0.78) and by 7 percentage points on Day 2 (0.79 vs. 0.72). Across both models and both protocols, the mean calibration duration required to reach optimal performance ranged from 240 to 320 seconds per class, supporting the feasibility of achieving acceptable on-site personalization with limited subject-specific data.

Table 2. Results Using Day 1 Protocols

Model	Baseline (LOSO)		Best Fine-tuned		
	F1	Acc	F1	Acc	Duration
Hierarchical	0.62	0.65	0.81	0.85	240 s
Flat Model	0.62	0.68	0.78	0.82	300 s

Table 3. Results Using Day 2 Protocols

Model	Baseline (LOSO)		Best Fine-tuned		
	F1	Acc	F1	Acc	Duration
Hierarchical	0.59	0.64	0.79	0.82	320 s
Flat Model	0.54	0.66	0.72	0.77	285 s

ACC: Accuracy; F1: Macro F1 score; Duration: Average duration to best F1-score during fine-tuning

4 Discussions

Wearable biosensors, specifically those embedded with PPG sensors have great potential to improve construction safety by enabling continuous monitoring of worker stress states. For such stress monitoring systems to be operationally useful in construction environments, they must move beyond identifying whether a worker is stressed to specifically identifying stress types (low stress, challenge, threat), since these states have different implications for performance and safety. Existing studies have overlooked this requirement by framing stress detection as a binary or flat three-class problem and not addressing the issue of inter-subject physiological variability [11,18,19].

This study addresses these limitations by introducing a hierarchical classification framework grounded in the stress appraisal theory established by the BPSM [13]. By decomposing the task into two sequential stages, Stage A separating low stress from stress and Stage B differentiating challenge from threat, the model aligns with the theory-driven transition from broad arousal detection to fine-grained stress appraisal discrimination. Compared to a flat three-class classifier, the hierarchical model demonstrated more stable baseline performance across experimental protocols and had higher performance after subject-specific calibration. These results support the interpretation that hierarchical decomposition improves learning by reducing the effective complexity of each decision: Stage A can focus on robust arousal-related distinctions, while Stage B specializes on the more difficult and physiologically subtle challenge versus threat boundary. In addition, the design enabled stage-specific optimization, which provided an opportunity to further optimize the Stage B model to effectively capture the difficult distinctions between threat and challenge without destabilizing the Stage A training which is straightforward. Collectively, these design properties contributed to improved robustness across protocols and more consistent fine-tuning performance.

The study also evaluated subject-specific fine-tuning as a practical strategy to improve field applicability in the presence of inter-subject physiological variability. Baseline LOSO-CV performance indicated inter-subject variability and moderate model generalization, which is consistent with known challenges in physiological classification [20]. However, fine-tuning with limited calibration data (typically 240–320 seconds per class) led to substantial performance gains for both models, with the hierarchical framework showing the greatest improvements and the highest performance across both protocols. This suggests that although fully generalizable population-level models remain challenging to achieve for stress appraisal classification, deployable performance can be reached through modest fine-tuning.

The temporal splitting strategy further strengthens ecological validity by evaluating models on future, unseen data, consistent with realistic deployment where calibration occurs early and monitoring occurs later during work.

Beyond model performance, the ability to distinguish low stress from challenge and threat has direct implications for construction safety management. A system that detects specific stress types enables targeted interventions that mitigate threat risk without disrupting adaptive challenge responses. The proposed fine-tuning strategy further supports on-site deployment of stress monitoring systems by allowing rapid adaptation to new workers using limited calibration, which is more feasible for construction settings where comprehensive subject data is unavailable and model retraining is almost impossible.

Despite the results achieved, some limitations should be acknowledged. The dataset included only eight participants. While this may be adequate for feasibility validation, it limits generalizability and may not fully represent the demographic and physiological diversity of the construction workforce. Future studies should validate the framework with larger samples spanning broader age ranges, fitness levels, and cardiovascular baselines to better characterize inter-subject variability. Second, although the controlled laboratory protocol enabled reliable labeling of low stress, challenge, and threat conditions, real construction environments introduce additional sources of variability, including motion artifacts, temperature effects, vibration, and task heterogeneity. Future work should therefore evaluate model performance under varying in-the-wild conditions.

5 Conclusions

This study presents a field-applicable hierarchical deep learning framework, grounded in the interaction between the HPA and SAM axes, for discriminating low stress, challenge, and threat appraisals using PPG-derived features. By decomposing the three-class task into a two-stage hard-cascade aligned with the theoretical structure of stress appraisal, and combining this approach with subject-specific fine-tuning, the framework achieved F1-scores of 0.81 (Day 1) and 0.79 (Day 2), improving substantially over baseline LOSO-CV performance (0.62 and 0.59).

The practical implications for construction safety and occupational health are significant. By distinguishing threat from challenge, the system enables targeted stress-management interventions while avoiding unnecessary responses during adaptive challenge states. With limited calibration data (approximately 4–6 minutes per class), the framework can be personalized to individual workers, supporting performance levels suitable for continuous monitoring in construction settings. Validation across two distinct threat-induction protocols further suggests robustness to different stressor mechanisms. Future work should validate the framework with larger and more diverse samples in real construction environments, examine temporal dynamics of appraisal transitions, and evaluate whether additional sensing modalities can further improve discrimination of challenge and threat.

Overall, this work establishes a foundation for personalized stress appraisal monitoring that can support early identification of elevated-risk threat responses and inform timely, targeted interventions in construction and other safety-critical work contexts.

Acknowledgement

This study was supported by the Government of Alberta's Supporting Psychological Health in First Responders Program (Award No. 23SPHIFR27-2), the NSERC Discovery Grants Program (Award No. RGPIN-2025-05214), AGE-WELL (Award No. AW-CAT-2023-10), and the NSERC Alliance Advantage Program (Award No. ALLRP 578484-22). The authors are also grateful to the anonymous participants for their contribution to this research.

References

[1] R. Blair Winkler, C. Middleton, O. Remes, Work-Related Stress and Coping: A Comparative Analysis of On-Site and Office-Based Workers in UK Building Construction, Healthcare 12 (2024) 2117. https://doi.org/10.3390/healthcare12212117.

[2] H.-C. Seo, Y.-S. Lee, J.-J. Kim, N.-Y. Jee, Analyzing safety behaviors of temporary construction workers using structural equation modeling, Saf. Sci. 77 (2015) 160–168. https://doi.org/10.1016/j.ssci.2015.03.010.

[3] H. Lingard, S. Rowlinson, Occupational Health and Safety in Construction Project Management, Routledge, 2004. https://doi.org/10.4324/9780203507919.

[4] M. Leung, I.Y.S. Chan, C.L. Cooper, Stress Management in the Construction Industry, Wiley, 2014. https://doi.org/10.1002/9781118456361.

[5] J.S. Petersen, C. Zwerling, Comparison of health outcomes among older construction and blue-collar employees in the United States, Am. J. Ind. Med. 34 (1998) 280–287.

[6] S. Cohen, T. Kamarck, R. Mermelstein, Perceived stress scale. Measuring stress:A guide for health and social scientists, 10 (1994).

[7] G.M. Slavich, G.S. Shields, Assessing Lifetime Stress Exposure Using the Stress and Adversity

Inventory for Adults (Adult STRAIN): An Overview and Initial Validation, Psychosom. Med. 80 (2018) 17–27. doi.org/10.1097/PSY.0000000000000534.

[8] L.K. McCorry, Physiology of the Autonomic Nervous System, Am. J. Pharm. Educ. 71 (2007) 78. https://doi.org/10.5688/aj710478.

[9] H. Jebelli, M.M. Khalili, S. Hwang, S. Lee, A Supervised Learning-Based Construction Workers' Stress Recognition Using a Wearable Electroencephalography (EEG) Device, in: Constr. Res. Congr. 2018, American Society of Civil Engineers, Reston, VA, 2018: pp. 40–50. https://doi.org/10.1061/9780784481288.005.

[10] H. Jebelli, S. Hwang, S. Lee, EEG-based workers' stress recognition at construction sites, Autom. Constr. 93 (2018) 315–324. https://doi.org/10.1016/j.autcon.2018.05.027.

[11] H. Jebelli, B. Choi, S. Lee, Application of Wearable Biosensors to Construction Sites. I: Assessing Workers' Stress, J. Constr. Eng. Manag. 145 (2019).

[12] R. Lazarus, R. Folkman, Stress, Appraisal, and Coping, New York Springer 1984 (n.d.).

[13] M.D. Seery, The Biopsychosocial Model of Challenge and Threat: Using the Heart to Measure the Mind, Soc. Personal. Psychol. Compass 7 (2013) 637–653. https://doi.org/10.1111/spc3.12052.

[14] G.G. Berntson, K.S. Quigley, D. Lozano, Cardiovascular Psychophysiology, in: J.T. Cacioppo, L.G. Tassinary, G. Berntson (Eds.), Handb. Psychophysiol., Cambridge University Press, Cambridge, n.d.: pp. 182–210.

[15] S.S. Dickerson, M.E. Kemeny, Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research., Psychol. Bull. 130 (2004) 355–391. https://doi.org/10.1037/0033-2909.130.3.355.

[16] M.D. Seery, Challenge or threat? Cardiovascular indexes of resilience and vulnerability to potential stress in humans, Neurosci. Biobehav. Rev. 35 (2011) 1603–1610. https://doi.org/10.1016/j.neubiorev.2011.03.003.

[17] A.A. Awad, A.S. Haddadin, H. Tantawy, T.M. Badr, R.G. Stout, D.G. Silverman, K.H. Shelley, The relationship between the photoplethysmographic waveform and systemic vascular resistance, J. Clin. Monit. Comput. 21 (2007) 365–372.

[18] G. Lee, S. Lee, Feasibility of a Mobile Electroencephalogram (EEG) Sensor-Based Stress Type Classification for Construction Workers, in: Constr. Res. Congr. 2022,. https://doi.org/10.1061/9780784483985.033.

[19] G. Saqib, S.-P. Chen, G. Lee, Distinguishing Stress Appraisals in Occupational Settings Using Wearable PPG and Deep Learning, in: IEEE Int. Humanit. Technol. Conf., IEEE, Edmonton, 2025.

[20] G. Lee, S. Lee, Importance of Testing with Independent Subjects and Contexts for Machine-Learning Models to Monitor Construction Workers' Psychophysiological Responses, J. Constr. Eng. Manag. 148 (2022).

[21] J.R. Turner, Individual Differences in Cardiovascular Reactivity, in: Cardiovasc. React. Stress, Springer, 1994: pp. 71–89. https://doi.org/10.1007/978-1-4757-9579-0_5.

[22] P.E. Sawchenko, H.-Y. Li, A. Ericsson, Circuits and mechanisms governing hypothalamic responses to stress: a tale of two paradigms, in: 2000: pp. 61–78.

[23] J. Tomaka, J. Blascovich, R.M. Kelsey, C.L. Leitten, Subjective, physiological, and behavioral effects of threat and challenge appraisal., J. Pers. Soc. Psychol. 65 (1993) 248–260. https://doi.org/10.1037/0022-3514.65.2.248.

[24] M.M. Bradley, P.J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, J. Behav. Ther. Exp. Psychiatry 25 (1994) 49–59.

[25] Y.-N. Jeng, P.. Huang, Y.-C. Cheng, Decomposition of one-dimensional waveform using iterative Gaussian diffusive filtering methods, Proc. R. Soc. A Math. Phys. Eng. Sci. (2008). https://doi.org/10.1098/rspa.2007.0031.

[26] J.E. Trammel, A. Sapra, Physiology, Systemic Vascular Resistance, 2025. https://doi.org/32310535.

[27] T. Le, F. Ellington, T.-Y. Lee, K. Vo, M. Khine, S.K. Krishnan, N. Dutt, H. Cao, Continuous Non-Invasive Blood Pressure Monitoring: A Methodological Review on Measurement Techniques, IEEE Access 8 (2020) 212478–212498. https://doi.org/10.1109/ACCESS.2020.3040257.

[28] M. Elgendi, Y. Liang, R. Ward, Toward Generating More Diagnostic Features from Photoplethysmogram Waveforms, Diseases 6 (2018). https://doi.org/10.3390/diseases6010020.

[29] C. Yang, B. Gao, L. Peng, A novel non-invasive approach for estimating cardiac output using photoplethysmography, in: L. Trajkovic, Z. Chen (Eds.), Second Int. Conf. Electron. Electr. Control Syst. (EECS 2025), SPIE, 2025: p. 26.