2025 IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3315-5899-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBE66822.2025.00079

Personalized Threshold-Based HRV Stress Detection from PPG Data: Preliminary Evaluation with a Biofeedback Serious Game

Stylianos M. Papelis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
steliospapelis@gmail.com

Nikolaos Bothos-Vouterakos
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
nikolasbothos@gmail.com

Konstantinos Mitsis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
kmhtshs@biosim.ntua.gr

Aikaterini Fragkou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
aifragkou@biosim.ntua.gr

Glykeria Theodorou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
gtheodorou@biosim.ntua.gr

Eleftherios Kalafatis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
leftkal@biosim.ntua.gr

Theofanis Ganitidis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
theogani@biosim.ntua.gr

Konstantina S. Nikita
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
knikita@ece.ntua.gr

***Abstract*—In this study, we present preliminary results from the evaluation of an automated real-time stress detection approach during cognitively demanding human–computer interaction. The proposed method relies on heart rate variability (HRV) analysis derived from photoplethysmography (PPG) data collected via a wireless wearable sensor. A personalized HRV threshold is established during a Stroop task of escalating difficulty and subsequently used for binary classification between calm and stressed states. The stress-detection approach is evaluated through an experimental protocol employing a novel dynamic biofeedback serious game (SG) and continuous post-game self-annotations of perceived stress levels. Fourteen individuals participated in the study, with three excluded due to sensor-related issues. The collected data were analyzed in two directions—absolute values and ordinal-centric measures—examining associations between the detected state and user-reported annotations. Statistically significant differences between calm and stressed game segments were observed for the mean (p = 0.002), median (p = 0.003), and trapezoidal interval (p = 0.002) of the annotation values. Ordinal analysis further confirmed this relationship, with positive Spearman rank correlations for the same features (p = 0.002) and consistent directionality in ten of eleven participants. These preliminary findings underscore the potential of the proposed thresholding approach and motivate further refinement of rule-based stress detection systems.**

Keywords—stress detection, heart rate variability (HRV), photoplethysmography (PPG), serious game, procedural content generation, biofeedback

I. Introduction

Physiological signals for automatic and real-time detection of stress levels have been widely studied in recent years as a means for personalization in various interventions [1]. Stress is defined as the physiological and psychological response to situations perceived as demanding or exceeding an individual's coping resources [2]. In research and practice, stress is typically categorized as either acute or chronic. Acute stress refers to short-term, situational responses to specific demands or challenges, such as cognitive tasks, whereas chronic stress describes the cumulative impact of prolonged exposure to stressors [3]. Real-time stress detection primarily targets acute stress, as rapid changes in physiological and cognitive states are more relevant during human computer interaction. By monitoring these changes, interventions adjust their behavior in real-time, supporting goals such as mental health improvement and cognitive enhancement. Additionally, stress detection contributes to improving user experience during challenging or demanding tasks, ultimately optimizing intervention outcomes [4]. However, reliably identifying stress in real time remains challenging, especially during highly dynamic and interactive scenarios [5].

Stress detection commonly relies on detecting changes in the autonomic nervous system. Several types of biosignals are used, such as electroencephalography (EEG) for monitoring brain activity, electrocardiography (ECG) and photoplethysmography (PPG) for capturing cardiac activity and electrodermal activity (EDA) for measuring skin conductance [6]. These signals provide objective markers of physiological arousal, which is closely linked to stress responses. Among them, cardiac-based signals such as ECG and PPG are frequently used because they allow for the extraction of heart rate variability (HRV), a well-established indicator of autonomic nervous system modulation and a widely applied metric in stress-related applications [7].

In HRV analysis, time-domain features are commonly used due to their sensitivity to short-term stress related changes [8]. Given the need to capture these fluctuations efficiently, recent studies highlight the potential of ultra-short-term HRV analysis, where HRV is estimated over windows shorter than one minute to support real-time applications [9]. PPG signals are increasingly favored in this context as they offer a practical and non-obtrusive method of capturing the pulse wave, making them well-suited for wearable devices and interactive systems. As a result, PPG-derived HRV has been the focus of numerous attempts to estimate stress states in real time, offering a balance between accessibility and physiological validity [10].

In real-time stress detection, rule-based systems using individualized thresholds are commonly employed as an alternative to machine learning approaches [11]. While machine learning has the capacity to model complex physiological patterns, these methods often operate as black boxes, offering limited insight into how decisions are made.

2471-7819/25/$31.00 ©2025 IEEE
DOI 10.1109/BIBE66822.2025.00079

In contrast, rule-based systems provide a transparent, easily interpretable framework, allowing precise understanding of how decisions are derived [12]. These systems typically rely on participant-specific calibration sessions to establish physiological baselines under calm and stressed conditions, from which thresholds are computed [13]. Once established, these thresholds enable the identification of changes in physiological states in a straightforward and interpretable manner, supporting applications where clarity, simplicity and responsiveness are prioritized. As such, proper and systematic calibration of these thresholds remains a critical step, as it directly impacts the reliability of individualized detection [14].

Standard calibration practice involves recording physiological responses during periods of relaxation to capture a calm baseline and employing cognitive stressors to establish a stressed baseline [15]. However, obtaining consistent and representative baselines is inherently challenging. Physiological signals are influenced by numerous factors beyond experimental control, including mood, fatigue, and environmental conditions. Moreover, within-individual variability across sessions further complicates the process, as baseline responses can fluctuate over time [16]. These factors make calibration a particularly difficult aspect of real-time stress detection, where the accuracy of threshold-based systems depends heavily on the robustness of this procedure, especially in cognitive demanding interventions [13].

Multiple studies have combined individualized calibration with rule-based logic for stress detection in controlled and non-invasive environments. For example, [17] employs calm and stressed baselines to construct fuzzy logic templates capable of binary classification. Similarly, [18] derives personalized HRV thresholds from categorized conditions, enabling simple threshold-based detection without relying on machine learning models. Other approaches integrate case-based reasoning with fuzzy logic systems, using calibration data to build individualized profiles that inform decision making [19]. Template-based methods are also been used to define rule-based thresholds for stress detection, particularly in systems designed for static and non-adaptive contexts [20], [21]. These studies highlight the reliability and simplicity of threshold-based calibration approaches. However, they remain largely situated within structured and low-complexity settings, offering limited insight into their performance under cognitively rich or adaptive scenarios.

A limited number of rule-based logic approaches integrated in dynamic or real-time contexts are found in the literature. For instance, [11] presents a wearable system for stress detection during everyday activities, combining a rule-based model with online segmentation to infer physiological changes in real time. In biofeedback applications, [22] introduces a fuzzy logic classifier that relies on participant-specific baselines established before gameplay. In a related approach, [21] applies fuzzy rule-based detection within wireless sensor networks, using thresholds derived from calibration sessions to enable real-time classification. Although these studies demonstrate solutions towards actual deployment, they remain limited in task complexity or user interaction. As a result, the ability of such methods to perform reliably under cognitively rich and adaptive conditions remains largely unexplored [11], [20]. Additionally, calibration-driven approaches continue to face several limitations in the context of real-time stress detection. Among the most prominent are the high degree of physiological variability across individuals [23], the sensitivity of threshold values to session-specific conditions [24] and the need for repeated calibration in multi-session or longitudinal use cases [25]. Given these limitations, the robustness of rule-based calibration systems under such conditions remains underexplored, motivating the need for further investigation into their practical applicability in dynamic and interactive contexts.

In this work, we present preliminary evaluation results of a novel threshold-based framework for real-time binary stress detection during cognitively rich interactions. The proposed approach employs ultra-short-term HRV analysis based on PPG signals. Individualized thresholds are calculated through participant-specific calibration sessions involving calm and stressed conditions. Based on the computed threshold, stress state classification is performed in real time, supporting applications where timely adaptation to the user's physiological state is desired. The proposed approach is integrated for evaluation purposes in an adaptive biofeedback serious game (SG) promoting stress self-management, to produce a rich interaction environment where the validity of stress detection directly impacts the desired intervention.

II. Proposed Framework

The proposed threshold-based framework for real-time automated stress detection is illustrated in Fig. 1. It comprises four modules: data collection and processing, calibration, stress classification, and intervention. During the intervention, HRV analysis is conducted in real time using sensor data. The results are compared against a personalized threshold, established during the calibration process, to classify the user as being in either a calm or a stressed state. The classified state is then transferred to the intervention module, which controls the generation of tailored content. This approach aims to establish a robust personalized threshold capable of distinguishing between calm and stressed states during cognitively demanding interactions. To this end, the calibration module incorporates both a relaxation period and a cognitive task of increasing difficulty. The four modules of the proposed framework are described in detail below.

A. Data Collection and Processing Module

PPG data are collected using the EmotiBit sensor [26], a wearable device designed for multi-modal biosignal acquisition. The sensor is placed on the finger to ensure optimal signal quality, as this location provides strong PPG signals due to its rich vascularization and accessibility for stable placement [26]. PPG data are sampled at 100 Hz. The EmotiBit device transmits the PPG data in real time to the stress classification module via the Open Sound Control (OSC) protocol. The collected PPG data are processed during the intervention using a 30-second sliding window, updated every 5 seconds to ensure continuous tracking of short-term changes in HRV. This window length has been validated for ultra-short-term HRV estimation in real-time stress detection scenarios [9], while frequent updates support responsive feedback in interactive contexts [27]. Within each window, the signal is band-pass filtered and its peaks are identified using the scipy.signal Python library [28]. The resulting peaks are used to compute the RR intervals between successive

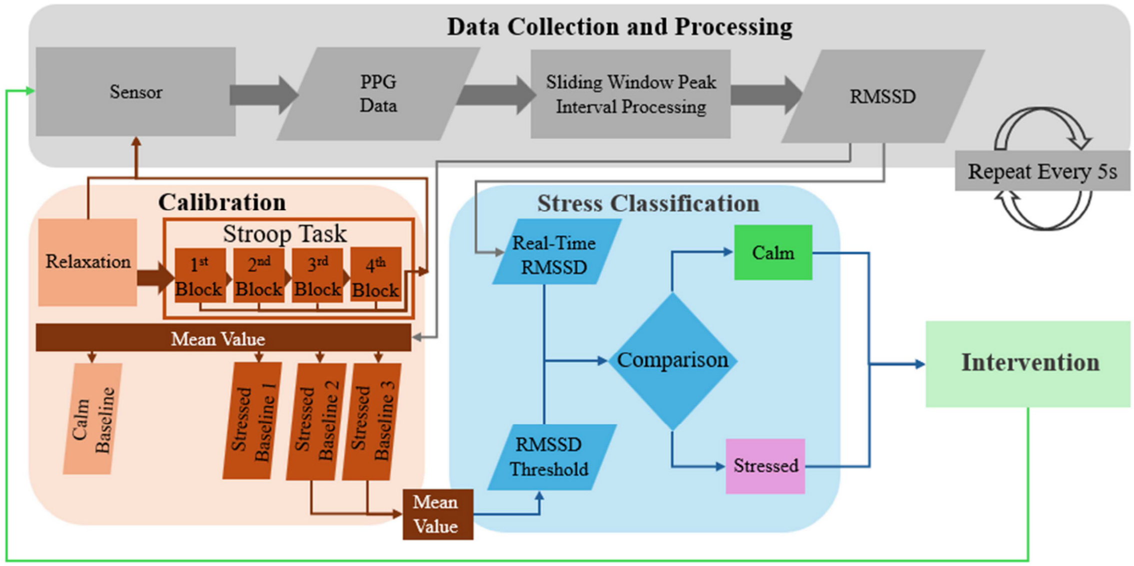


Fig. 1. Flowchart of the calibration process, thresholding and binary classification of stress during the intervention.

heartbeats. For the HRV analysis, the root mean square of successive differences (RMSSD) is calculated for each window, as it is a widely adopted time-domain feature that is particularly sensitive to short-term fluctuations in HRV, making it suitable for real-time stress monitoring in interactive contexts [9].

B. Calibration Module

The calibration module consists of two phases, a five-minute relaxation phase and a cognitive challenge phase. PPG signals are collected and RMSSD values are calculated during both phases in the data collection and processing module. The relaxation period is employed to establish an individualized baseline corresponding to a calm physiological state as recommended in previous HRV studies [29]. The user is instructed to remain still while seated comfortably and listening to soothing music. Only the last 30 segments of relaxation period are used for RMSSD calculation, a practice recommended to minimize the influence of initial signal instability and participant adjustment [30].

In the second phase the user undertakes a digitized Stroop Task [31], a test designed to induce cognitive stress. They are instructed to identify the font color of words displayed on the monitor, by pressing one of the three matching colored keys [32]. Stimuli consist of either congruent trials where the meaning of the word matches its font color (e.g., “Blue” displayed in blue) or incongruent trials where meaning and color are conflicted (e.g., “Green” displayed in red). This conflict produces what is known as the “Stroop Interference” [33], increasing response latency and cognitive stress.

To produce an efficient personalized threshold for cognitively demanding interventions, the Stroop Task is split into four blocks. The first block, containing 15 congruent and 15 incongruent trials. aims to familiarize users with the task and response modality [34]. The remaining three test blocks consist of 80 trials each, progressing from exclusively congruent items on the first block to exclusively incongruent items on the third block. The second block presents an equally divided combination of congruent and incongruent stimuli, thus avoiding prediction mechanisms that reduce the interference [35]. Each block consists of 80 trials to provide adequate interaction with the user [36]. After each block, participants are given a 30-second break to reduce fatigue [34].

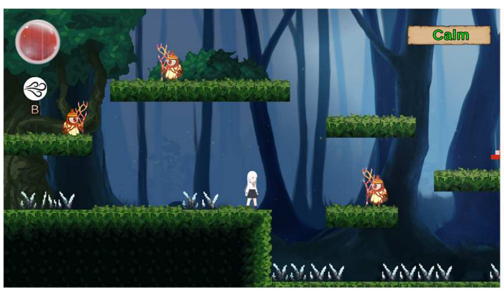
RMSSD is calculated in the last three blocks using the 11 final 5-second segments to minimize the influence of initial signal instability and participant adjustment. The selection of the number of segments included was based on trial and error prior to the start of the experimental process due to the variability in trial response time. RMSSD values are then averaged to generate a stressed baseline for each of the three blocks [37]. To define the personalized threshold, the baselines from the third and fourth blocks are averaged. These two blocks contain incongruent stimuli in proportions that produce high and low Stroop interference respectively. In contrast, the second block, comprising only congruent trials, serves primarily to ensure validity. Baselines obtained during the relaxation period and the first block of the Stroop Task are also used to ensure the validity. This thresholding strategy is proposed to establish a clear and individualized distinction between calm and stressed states. By positioning the threshold between empirically observed low and high-stress conditions, this method aims to exclude mild arousals which are consistently present due to task engagement, while being sensitive enough to detect meaningful fluctuations, indicative of temporal stress, achieving a desired balance between specificity and sensitivity.

C. Stress Classification Module

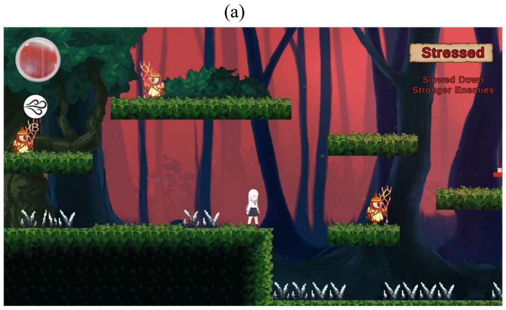
In the stress classification module, RMSSD values obtained in real-time during interaction with the intervention are compared to the personalized threshold generated in the calibration module every five seconds. RMSSD values above threshold are classified as calm state, values below threshold suggest heightened stress levels and are classified as stressed state [38]. The classified state is communicated to the intervention every five seconds through a binary variable, balancing real-time adaptation with stability by avoiding excessive fluctuations. To maintain responsiveness, the communication between classification and intervention modules is handled via an OSC-based protocol, ensuring minimal latency and compatibility with real-time interactive systems.

D. Evaluation Experiment

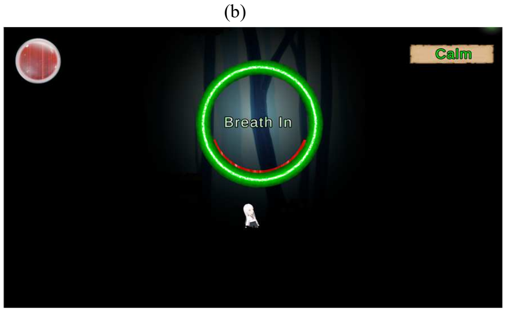
For the evaluation of the suggested stress detection approach a novel 2D platformer SG promoting stress self-management through deep breathing exercises [39] is employed as an intervention. SGs constitute an ideal application domain for the proposed approach, as they incorporate highly interactive environments where adapting to the user's cognitive and emotional state is important for optimizing the delivered intervention [40]. The employed SG is developed using the Unity engine [41] for Windows operating systems and is developed with the capacity to receive sensor data in real time through a processing server. It features procedural content generation and dynamic difficulty adjustment based on automatic real-time recognition of player stress levels. To this end, the SG switches between two distinct game states, denoted as calm (Fig. 2a) and stressed (Fig. 2b). During the calm game state the SG retains normal difficulty levels, while during the stressed game state, difficulty increases by generating additional obstacles and hazards. This real-time adaptation aims to reinforce the connection between physiological state and in-game experience, aligning the gameplay environment with the participant's current stress level. Additionally, the game includes a player-controlled mechanic which allows them to perform guided deep breathing exercises (Fig. 2c). As the player performs the exercise, their in-game progress is facilitated in two ways, through in-game character restoration of health points and their transition towards the calm state.



(a)



(b)



(c)

Fig. 2. Screenshots from the two game states and a breathing session: (a) Calm; (b) Stressed; (c) Breathing.

A total of 14 individuals, 8 male and 6 female, aged 27.1±5.87 participate in the experimental process. All participants confirm no consumption of caffeine and report having slept well for 5 days prior to their experiment session. All participants report no chronic or neurological condition, or taking medication that may affect stress levels, heart activity or concentration. After a detailed description of the procedure and the aim of the study, all participants provide written informed consent and are given the opportunity to ask questions. The study has been approved by the Ethics Committee of the National Technical University of Athens (approval number : 45610/10.07.2025).

Participants are seated comfortably in front of a computer monitor for the duration of the experiment. The EmotiBit sensor is placed on the index finger of the non-dominant hand. Their hand is positioned on the table in a relaxed and stable posture to minimize movement and reduce noise in the PPG signal. Participants then go through the calibration process followed by a gameplay session of one SG stage using their dominant hand. During this session the proposed approach for stress detection controls game states in real time. Afterwards, participants are asked to annotate their perceived level of stress while watching a screen recording video of their gameplay session at double speed [42], to minimize fatigue. For this purpose, a specialized annotation tool, based on RankTrace [43], is employed to approximate ground truth regarding stress levels. This tool allows for unbounded and continuous annotation of affect and is developed using GameMaker Studio [44]. Participants use the mouse scroll wheel to indicate changes in perceived stress, with upward scrolling corresponding to increased stress and downward scrolling indicating reduced stress. Annotation traces are used to evaluate the proposed framework for automated real-time stress detection. Annotations are recorded at a frequency of 1 Hz throughout the duration of the replay [45]. Finally, participants complete the Game Experience Questionnaire (GEQ) [46] to assess their subjective experience during the game.

III. Data Analysis

A. Data Preparation

Data used in the analysis for the evaluation of the proposed stress detection approach are obtained from two primary sources; SG logs and the annotation trace. Logs provide a record of the classified state during the gameplay session, enabling its division into three types of segments, calm, stressed and deep breathing of varying durations. Deep breathing segments are not used in the analysis as they are initiated by the participant and do not include SG content. 3 participants were excluded due to excessive movements detected on the EmotiBit sensor, which compromised data integrity. Consequently, the final analytical sample comprised eleven participants, who contributed on average 15.4 ± 3.4 calm segments with an average duration of 30.8±5.7 seconds and 9.4 ± 3.6 stressed segments with an average duration of 14.9 ±10.4 seconds.

Temporal harmonization of the data is conducted to account for differences in sampling frequencies and the speed of the screen recording video. In particular, linear interpolation is applied to the annotation trace, producing estimated annotation values at 100 Hz. Subsequently, normalization of annotation values to the range [0,1] is performed for each participant [43]. The final dataset consists of calm and stressed segments each aligned and associated with the participants' corresponding annotation data and prepared for feature extraction. An example of a normalized annotation trace from a representative participant, aligned with the gameplay segments is presented in Fig. 3. In this example, peaks in the annotation trace are mostly followed by player inititated deep breathing segments as expected.

From each segment, five statistical features [43] are calculated: the mean annotation value (μA); the median annotation value ($\tilde{A}$); the area under the annotation trace ($\int A$), computed using the composite trapezoidal rule and normalized by the segment duration; the amplitude ($\hat{A}$), calculated as the difference between the maximum and minimum annotation values within the segment; and the average gradient of the annotation trace (ΔA), representing the average rate of change over time. Finally, for each participant, the mean of each of these five features is computed across all calm segments and, separately, across all stressed segments.

B. Statistical Analysis

Statistical analysis is performed in two different directions. The first focuses on evaluating whether the binary classification of stress derived from the game logs corresponds to measurable differences in participants' subjective perception of stress. For each participant, the aggregated annotation features from calm and stressed segments are compared in pairs to assess whether perceived stress levels differed consistently between the two states. The non-parametric Wilcoxon signed-rank test is applied, with statistical significance set at $p < 0.05$.

The second direction features a complementary ordinal approach [47], motivated by the increasing recognition within affective computing that emotions and subjective experiences are inherently ordinal constructs rather than interval-scaled quantities [48],[49]. An ordinal-centric perspective distinguishes between first-order approaches, where data are directly collected in the form of ranks or preferences and second-order approaches, where interval or amplitude-based annotations are transformed into ranks to capture their underlying ordinal nature. Although annotation traces are initially recorded as amplitude values, the second-order approach is followed by ranking all calm and stressed segments for each participant based on the extracted annotation features. Spearman's rank correlation coefficients between these ranks and the binary state labels (calm = 0, stressed = 1) are then computed to assess whether higher perceived stress, thus higher ranked segments are associated with the stressed binary label.

To test the consistency of this relationship across participants, we apply the Wilcoxon test on the resulting correlation coefficients [43] to examine whether they are significantly different from zero, thus indicating a systematic ordinal association between perceived stress levels and the proposed binary classification. Additionally, a binomial test is conducted to evaluate whether the number of positive correlations significantly exceed what would be expected by chance, providing further evidence that higher feature ranks align with segments labeled as stressed.

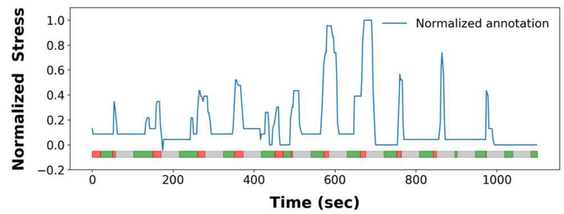


Fig. 3. Example of an annotation trace of perceived stress, divided into calm (green), stressed (red) and breathing (grey) segments

IV. RESULTS

In the absolute value analysis, systematic differences between calm and stressed segments are observed, as the application of the Wilcoxon signed-rank test yields statistically significant higher values during stressed segments for the mean value ($p = 0.002$), median ($p = 0.003$) and area under the annotation trace ($p = 0.002$). No statistically significant differences are identified in amplitude ($p = 0.240$) or average gradient ($p = 0.966$). The observed elevations in mean value, median value and area under the annotation trace during stressed conditions are indicative of an association between system-labeled stressed segments and increased levels of perceived stress as reported by participants. The average value across all participants, for each annotation feature of both calm and stressed segments is presented in Fig. 4.

The ordinal-centric analysis displays positive correlations across all participants, with statistically significant difference from zero for mean rank ($p = 0.002$), median rank ($p = 0.002$) and area rank ($p = 0.002$), suggesting that stressed segments are generally associated with higher perceived stress ranks. By contrast, amplitude rank and gradient rank do not reach statistical significance ($p = 0.465$ and $p = 0.831$, respectively).

Finally, the binomial tests display positive correlations for mean rank, median rank and area rank, in ten out of eleven participants ($p = 0.006$), providing evidence of a systematic relationship across the sample. No significant effects are identified for gradient rank (seven out of eleven, $p = 0.274$) or amplitude rank (three out of eleven, $p = 0.967$). The results of both Wilcoxon tests are summarized in Table 1.

TABLE I. WILCOXON TEST P-VALUES FOR BOTH ANALYSES. SIGNIFICANT VALUES (P<0.05) ARE DEPICTED IN BOLD.

Annotation Features	Wilcoxon p-values	
	Absolute Value Analysis	*Ordinal-Centric Analysis*
Mean (μA)	**0.002**	**0.002**
Median ($\tilde{A}$)	**0.003**	**0.002**
Area ($\int A$)	**0.002**	**0.002**
Amplitude ($\hat{A}$)	0.240	0.465
Gradient (ΔA)	0.966	0.831

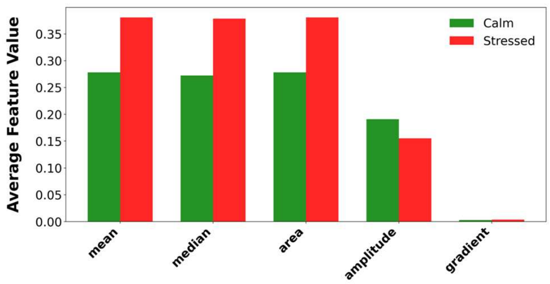


Fig. 4. Average values of perceived stress annotation features across all participants in calm (green) and stressed (red) segments.

V. Discussion

The presented preliminary results are in accordance with those reported in the literature, demonstrating that the RMSSD feature of ultra-short-HRV analysis can be associated with cognitive load and used as a psychological stress indicator. Additionally, comparisons with the self-reported stress annotation trace indicate that the proposed stress detection approach can identify stress states that align with the participant's subjective perception of stress during a cognitive demanding interaction.

Results indicative of this are observed in both statistical analysis directions. In the absolute value approach, the statistically significant higher mean (μA), median ($\tilde{A}$), and area under the annotation trace ($\int A$) observed during segments classified as stressed, indicate that the proposed thresholding and classification system are able to effectively identify heightened cognitive load, as perceived by users. This finding is reinforced by the results of the ordinal approach, where the stress segments as denoted by the detection system are consistently positioned higher in perceived stress ranks across participants, as indicated by the significantly positive correlations. However, both approaches suggest no significant correlation with the average gradient (ΔA) and Amplitude ($\hat{A}$) of the annotation trace, despite their reported robustness [43]. This could be attributed to the inherent variability of physiological signals, the complexity of manually annotating affect and the limited participant sample.

The presented preliminary results advocate towards the reliability of the proposed thresholding approach and build upon previous work regarding real-time stress detection by demonstrating classification capability in a rich interactive environment. The proposed method is low cost, wearable-friendly and light in computational requirements. Unlike machine learning models which often operate as black boxes, the personalized threshold baselines provide transparency, allowing direct observation of how stress states are classified. This makes the approach suitable for real-time applications where low latency and computational efficiency are required, such as adaptive and biofeedback interventions.

However, some limitations should be acknowledged. Generalizability of the findings is limited due to the small sample size, underscoring the need for larger scale studies. The requirement for an explicit calibration procedure using cognitive stressors like the Stroop task also poses practical challenges, as it might prove burdensome in real world applications. Efforts towards streamlining the process for outside research setting application need to be undertaken. Additionally, relying exclusively on HRV as a physiological marker, while well established in literature, may not fully capture the multimodal nature of stress responses, which are also reflected in other autonomic signals such as EDA and EEG [50]. Incorporating additional modalities in future work could help address these constraints and enhance the system's robustness and applicability. Finally, insights from this preliminary evaluation will guide further analysis and fine-tuning of the proposed approach.

VI. Conclusions

The present study investigates the potential of real-time stress detection in cognitively demanding interventions using a rule-based system based on ultra-short-term HRV derived from PPG signals. Personalized thresholds, established through calibration, enable binary classification of calm and stressed states, which dynamically guide the adaptation of the intervention's content. The classification capacity of the proposed approach is validated against continuous annotations of perceived stress, demonstrating significant alignment and supporting the effectiveness of the threshold computation method. While the method shows promising accuracy and practical simplicity, these preliminary results should be further validated and extended through the integration of additional biosignals, larger participant samples, and fine-tuning. Nevertheless, the findings highlight the potential of rule-based stress detection systems and their future integration into adaptive intervention environments involving cognitively demanding human–computer interactions.

Acknowledgment

S. M. Papelis and N. Bothos-Vouterakos contributed equally to this paper.

References

[1] E. Lazarou and T. P. Exarchos, "Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices," *AIMS Neurosci.*, vol. 11, no. 2, pp. 76–102, 2024, doi: 10.3934/Neuroscience.2024006.

[2] B. Chu, K. Marwaha, T. Sanvictores, A. O. Awosika, and D. Ayers, "Physiology, Stress Reaction," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: July 31, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK541120/

[3] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and Health: Psychological, Behavioral, and Biological Determinants," *Annu. Rev. Clin. Psychol.*, vol. 1, no. 1, pp. 607–628, Apr. 2005, doi: 10.1146/annurev.clinpsy.1.102803.144141.

[4] S. Vaca Narvaja, E. Dalal, M. Dong, Y. Ni, and I. Gonsher, "A Biofeedback-Driven Interaction System for Real-Time Stress Detection and Intervention," presented at the Human Interaction and Emerging Technologies (IHIET 2025), 2025. doi: 10.54941/ahfe1006725.

[5] M. Paniagua-Gómez and M. Fernandez-Carmona, "Trends and Challenges in Real-Time Stress Detection and Modulation: The Role of the IoT and Artificial Intelligence," *Electronics*, vol. 14, no. 13, p. 2581, June 2025, doi: 10.3390/electronics14132581.

[6] J. Chen, M. Abbod, and J.-S. Shieh, "Pain and Stress Detection Using Wearable Sensors and Devices—A Review," *Sensors*, vol. 21, no. 4, p. 1030, Feb. 2021, doi: 10.3390/s21041030.

[7] E. Zhou, M. Soleymani, and M. J. Matarić, "Investigating the Generalizability of Physiological Characteristics of Anxiety," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2023, pp. 4848–4855. doi: 10.1109/BIBM58861.2023.10385292.

[8] T. Pereira, P. R. Almeida, J. P. S. Cunha, and A. Aguiar, "Heart rate variability metrics for fine-grained stress level assessment," *Comput. Methods Programs Biomed.*, vol. 148, pp. 71–80, Sept. 2017, doi: 10.1016/j.cmpb.2017.06.018.

[9] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term HRV features as surrogates of short term HRV: a

case study on mental stress detection in real life," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 12, Dec. 2019, doi: 10.1186/s12911-019-0742-y.

[10] H. Barki and W.-Y. Chung, "Mental Stress Detection Using a Wearable In-Ear Plethysmography," *Biosensors*, vol. 13, no. 3, p. 397, Mar. 2023, doi: 10.3390/bios13030397.

[11] K. Kyriakou *et al.*, "Detecting Moments of Stress from Measurements of Wearable Physiological Sensors," *Sensors*, vol. 19, no. 17, p. 3805, Sept. 2019, doi: 10.3390/s19173805.

[12] D. Jaber, H. Hajj, F. Maalouf, and W. El-Hajj, "Medically-oriented design for explainable AI for stress prediction from physiological measurements," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 38, Dec. 2022, doi: 10.1186/s12911-022-01772-2.

[13] U. Zalabarria, E. Irigoyen, R. Martinez, M. Larrea, and A. Salazar-Ramirez, "A Low-Cost, Portable Solution for Stress and Relaxation Estimation Based on a Real-Time Fuzzy Algorithm," *IEEE Access*, vol. 8, pp. 74118–74128, 2020, doi: 10.1109/ACCESS.2020.2988348.

[14] S. Begum, M. U. Ahmed, P. Funk, and N. Xiong, "Individualized Stress Diagnosis Using Calibration and Case-Based Reasoning".

[15] R. E. Diaz-Ramos, D. A. Gomez-Cravioto, L. A. Trejo, C. F. López, and M. A. Medina-Pérez, "Towards a Resilience to Stress Index Based on Physiological Response: A Machine Learning Approach," *Sensors*, vol. 21, no. 24, p. 8293, Dec. 2021, doi: 10.3390/s21248293.

[16] J. Hannon, A. O'Hagan, R. Lambe, B. O'Grady, and C. Doherty, "Associations Between Daily Heart Rate Variability and Self-Reported Wellness: A 14-Day Observational Study in Healthy Adults," *Sensors*, vol. 25, no. 14, p. 4415, July 2025, doi: 10.3390/s25144415.

[17] A. De Santos Sierra, C. Sanchez Avila, J. Guerra Casanova, and G. Bailador Del Pozo, "A Stress-Detection System Based on Physiological Signals and Fuzzy Logic," *IEEE Trans. Ind. Electron.*, vol. 58, no. 10, pp. 4857–4865, Oct. 2011, doi: 10.1109/TIE.2010.2103538.

[18] D. Muhajir, F. Mahananto, and N. A. Sani, "Stress level measurements using heart rate variability analysis on android based application," *Procedia Comput. Sci.*, vol. 197, pp. 189–197, 2022, doi: 10.1016/j.procs.2021.12.200.

[19] S. Begum, M. U. Ahmed, P. Funk, N. Xiong, and B. Von Schéele, "Classify and Diagnose Individual Stress Using Calibration and Fuzzy Case-Based Reasoning," in *Case-Based Reasoning Research and Development*, vol. 4626, R. O. Weber and M. M. Richter, Eds., in Lecture Notes in Computer Science, vol. 4626. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 478–491. doi: 10.1007/978-3-540-74141-1_33.

[20] A. De, C. Snchez-Avila, J. Guerra-Casanova, and G. Bailador-Del, "Real-Time Stress Detection by Means of Physiological Signals," in *Recent Application in Biometrics*, J. Yang, Ed., InTech, 2011. doi: 10.5772/18246.

[21] M. Koussaifi, C. Habib, and A. Makhoul, "Real-time stress evaluation using wireless body sensor networks," in *2018 Wireless Days (WD)*, Dubai: IEEE, Apr. 2018, pp. 37–39. doi: 10.1109/WD.2018.8361691.

[22] L. Chittaro and R. Sioni, "Affective computing vs. affective placebo: Study of a biofeedback-controlled game for relaxation training," *Int. J. Hum.-Comput. Stud.*, vol. 72, no. 8–9, pp. 663–673, Aug. 2014, doi: 10.1016/j.ijhcs.2014.01.007.

[23] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on Psychological Stress Detection Using Biosignals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 440–460, Jan. 2022, doi: 10.1109/TAFFC.2019.2927337.

[24] M. Bolpagni, S. Pardini, M. Dianti, and S. Gabrielli, "Personalized Stress Detection Using Biosignals from Wearables: A Scoping Review," *Sensors*, vol. 24, no. 10, p. 3221, May 2024, doi: 10.3390/s24103221.

[25] S. N. Mohammad Affendi *et al.*, "Development and Evaluation of the Inner Peace Meter: A Galvanic Skin Response-Based Stress Monitoring System," *Semarak Int. J. Public Health Prim. Care*, vol. 4, no. 1, pp. 14–26, June 2025, doi: 10.37934/sijphpc.4.1.1426.

[26] S. M. Montgomery, N. Nair, P. Chen, and S. Dikker, "Introducing EmotiBit, an open-source multi-modal sensor for measuring research-grade physiological signals," *Sci. Talks*, vol. 6, p. 100181, May 2023, doi: 10.1016/j.sctalk.2023.100181.

[27] T. Tagnithammou, L. Trénoras, and A. Ferszterowski, "Stress Detection of Human Using Heart Rate Variability Analysis Based on Low Cost Camera," 2019.

[28] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[29] P. Satish, K. Muralikrishnan, and K. Balasubramanian, "Heart rate variability changes during stroop color and word test among genders," *Indian J Physiol Pharmacol*.

[30] M. Malik *et al.*, "Heart rate variability. Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, 1996, doi: 10.1093/oxfordjournals.eurheartj.a014868.

[31] F. Scarpina and S. Tagini, "The Stroop Color and Word Test," *Front. Psychol.*, vol. 8, p. 557, Apr. 2017, doi: 10.3389/fpsyg.2017.00557.

[32] M. Augustinova, B. A. Parris, and L. Ferrand, "The Loci of Stroop Interference and Facilitation Effects With Manual and Vocal Responses," *Front. Psychol.*, vol. 10, p. 1786, Aug. 2019, doi: 10.3389/fpsyg.2019.01786.

[33] R. Brunetti *et al.*, "eStroop: Implementation, Standardization, and Systematic Comparison of a New Voice-Key Version of the Traditional Stroop Task," *Front. Psychol.*, vol. 12, p. 663786, May 2021, doi: 10.3389/fpsyg.2021.663786.

[34] M. S. Ben-Haim, P. Williams, Z. Howard, Y. Mama, A. Eidels, and D. Algom, "The Emotional Stroop Task: Assessing Cognitive Performance under Exposure to Emotional Content," *J. Vis. Exp.*, no. 112, Art. no. 112, June 2016, doi: 10.3791/53720.

[35] D. H. Weissman and C. E. Saba, "The list-wide proportion congruency effect is larger when the distractor precedes the target: Evidence for conflict-independent control in the prime-probe task," *Mem. Cognit.*, vol. 53, no. 5, pp. 1551–1570, July 2025, doi: 10.3758/s13421-024-01669-7.

[36] J. M. Bugg, L. L. Jacoby, and J. P. Toth, "Multiple levels of control in the Stroop task," *Mem. Cognit.*, vol. 36, no. 8, Art. no. 8, Dec. 2008, doi: 10.3758/MC.36.8.1484.

[37] "development-of-four-stress-levels-in-group-stroop-colour-word-test-using-hrv-analysis," *Biomed Res*, vol. 28, no. 1, Art. no. 1, 2017.

[38] M. Salai, I. Vassányi, and I. Kósa, "Stress Detection Using Low Cost Heart Rate Sensors," *J. Healthc. Eng.*, vol. 2016, pp. 1–13, 2016, doi: 10.1155/2016/5136705.

[39] V. Perciavalle *et al.*, "The role of deep breathing on stress," *Neurol. Sci. Off. J. Ital. Neurol. Soc. Ital. Soc. Clin. Neurophysiol.*, vol. 38, no. 3, Art. no. 3, Mar. 2017, doi: 10.1007/s10072-016-2790-8.

[40] K. Mitsis, K. Zarkogianni, K. Dalakleidi, G. Mourkousis, and K. S. Nikita, "Evaluation of a Serious Game Promoting Nutrition and Food Literacy: Experiment Design and Preliminary Results," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Athens, Greece: IEEE, Oct. 2019, pp. 497–502. doi: 10.1109/BIBE.2019.00096.

[41] S. Singh and A. Kaur, "Game Development using Unity Game Engine," in *2022 3rd International Conference on Computing, Analytics and Networks (ICAN)*, Aug. 2022, pp. 1–6. doi: 10.1109/ICAN56228.2022.10007155.

[42] K. Pinitas *et al.*, "Predicting Player Engagement in Tom Clancy's The Division 2: A Multimodal Approach via Pixels and Gamepad Actions," in *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, Paris France: ACM, Oct. 2023, pp. 488–497. doi: 10.1145/3577190.3614203.

[43] P. Lopes, G. N. Yannakakis, and A. Liapis, "RankTrace: Relative and unbounded affect annotation," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, TX: IEEE, Oct. 2017, pp. 158–163. doi: 10.1109/ACII.2017.8273594.

[44] "GameMaker Studio. Available online: https://www.yoyogames.com/en/gamemaker." Accessed: July 30, 2025. [Online]. Available: https://gamemaker.io/en

[45] K. Mitsis *et al.*, "A Multimodal Approach for Real Time Recognition of Engagement towards Adaptive Serious Games for Health," *Sensors*, vol. 22, no. 7, Art. no. 7, Mar. 2022, doi: 10.3390/s22072472.

[46] W. A. IJsselsteijn, Y. A. W. de Kort, and K. Poels, *The Game Experience Questionnaire*. Eindhoven: Technische Universiteit Eindhoven, 2013.

[47] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions: An Emerging Approach," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, Art. no. 1, Jan. 2021, doi: 10.1109/TAFFC.2018.2879512.

[48] G. N. Yannakakis and H. P. Martínez, "Ratings are Overrated!," *Front. ICT*, vol. 2, July 2015, doi: 10.3389/fict.2015.00013.

[49] G. N. Yannakakis and J. Hallam, “Ranking vs. Preference: A Comparative Study of Self-reporting,” in *Affective Computing and Intelligent Interaction*, S. D’Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., Berlin, Heidelberg: Springer, 2011, pp. 437–446. doi: 10.1007/978-3-642-24600-5_47.

[50] S. A. Patil and A. N. Paithane, “AI-Driven Multimodal Stress Detection: A Comparative Study,” *Biomed. Pharmacol. J.*, vol. 18, no. March Spl Edition, pp. 245–255, Feb. 2025.