Social, Physical, and Cognitive Stressor Identification using Electrocardiography-derived Features and Machine Learning from a Wearable Device

Maxine He, Jonathan Cerna, Abdul Alkurdi, Jennifer Zhao, Prakhar Gupta, Jay Clore, Richard Sowers, Elizabeth T. Hsiao-Wecklser, and Manuel E. Hernandez

Abstract— **Anxiety is a prevalent and detrimental mental health condition that affects young adults, particularly those in underrepresented minorities in sciences, technology, engineering, and mathematics (STEM) disciplines. The ability to predict anxiety would help create individualized treatment. There is a need for objective and non-invasive continuous monitoring tools that allow for the prediction of anxiety. However, the generalizability of physiological changes across various stressors and participants must first be examined. The aim of this work is to examine the relationship of different stressors on heart rate variability in combination with machine learning (ML) models to assess binary and multi-class classification performance using electrocardiography (ECG) derived features from a wearable device. Five healthy young adults (4 female) from STEM disciplines performed baseline tasks (a guided meditation, cold pressor test, and resting state with eyes open or closed) and social, cognitive, and motor-cognitive challenges while wearing a Hexoskin smartshirt and an E4 wristband. The effect of stressor type on ECG-derived features was evaluated using a one-way ANOVA, while the performance of binary and multi-class ML classifiers of stressor type was evaluated. Nine out of 19 ECG features were significantly altered by stressor type. Binary classification accuracy of 77.1% and multiclass classification accuracy of 45.7% was achieved using a support vector machine (SVM) architecture. These results contribute to our understanding of individual anxiety detection using ML and have potential implications for using similar monitoring tools to predict anxiety using wearable devices.**

Wearables, mental health, machine learning

I. Introduction

Anxiety, the most common mental health condition in the United States [1], has been shown to have detrimental effects on physical and cognitive performance particularly in women and underrepresented minorities [2], [3], [4]. Thus, there is a widespread need for approaches to effectively promote psychophysiological well-being. Self-reported measures have been used as a gold standard for evaluating anxiety, but they are not feasible for continuous monitoring and are unable to provide a measure of event-contingent changes. Remote monitoring tools, such as wearable devices (e.g., smart watches, wristbands, shirts, and chest bands [5], [6]) can provide objective and non-invasive continuous recording of physiological signals that facilitate prediction of anxiety. Heart rate variability (HRV), extracted from cardiac measurements, reflects the modulation of heart rate by autonomic systems and has shown to be a strong stress/anxiety detection measure [7]. The combination of continuous recording with advanced data analytics (i.e., machine learning [ML] algorithms) enables the generalizing of stress responses in groups of subjects. However, it is also important to address the inter-individual variability and subjectiveness of stress and anxiety perception to improve personalized classification. We aim to (1) explore the changes in state anxiety and various HRV features in response to multiple acute stressors (i.e., cognitive, social, and physical), and (2) leverage strongly correlated HRV features as predictive physiological signals, aided by several ML models, to help further assess inter-individual classification and distinguish between multiple stressors with binary and multi-class labels. We hypothesize that ML models will yield high prediction accuracy and delineate HRV patterns across different stressors. Overall, this work will help begin to bridge the gap in understanding the personalization of anxiety detection using ML models.

II. Methods

A. Participants

Five adults (21.2±1.5 years of age, 4 females) are included in these preliminary findings. Inclusion criteria are a) 18-30 years of age; b) enrollment in a STEM degree program; c) willingness to wear a smartwatch, and smart shirt for 2 weeks; d) use a smartphone; e) ability to read English.

B. Protocol

Participants underwent baseline conditions in the following order: 5-minute guided meditation, 1-minute cold-pressor test, 1-minute eyes-closed resting (EC), 1-minute eyes-open resting (EO). Following baseline conditions, participants were exposed to the following in order: Trier Social Stress Test (TSST; social challenge), seated Stroop task (SST; cognitive challenge), and walking Stroop task (WST; motor-cognitive challenge). More specifically, the TSST includes a preparatory stage (1.5 minute) to prepare a short speech for an imaginary job interview, a 3-minute speech to a confederate representing the 'professor' conducting the interview. Following the speech task, participants were given a 5-minute arithmetic task, ending with a 1.5-minute wind-down. Both the SST and WST consisted of four trials: color matching, word matching (congruent), color matching with words in different colors (incongruent), and switching between colors and words upon presentation of a cue (task-switching). The difference between the SST and WST was that, in the WST condition, participants walked on a treadmill at a self-selected safe pace allowing them to still respond to the stimuli presented on the screen. Lastly, a reliable and validated smart shirt was utilized to collect electrocardiogram (ECG) signal during the entire testing session (Carré Technologies Inc., Montreal, Quebec, Canada) [8], [9], [10]. A smart wristband (E4, Empatica Inc., Boston, MA, United States) was also used to provide a tag of the start and end of each task condition to segment continuous recordings of the Hexoskin ECG data. To evaluate participant-reported affect and stress levels, we used digitized forms before and after each experimental condition using self-reported state anxiety items of the State-Trait Anxiety Inventory (STAI-Y6) [11].

C. Feature Extraction and Statistical Analysis

Raw ECG from Hexoskin recordings was visually checked for any noisy segments and was processed using Kubios Scientific Software (Kubios Oy, Finland) to extract relevant HRV features [12]. Timestamps tagged by E4 for each task condition were used to slice the continuous ECG data and HRV analysis was applied to each segment. Nineteen HRV features and accompanying respiratory frequency were extracted for each task condition in 5 participants (Table I).

Post-test STAI-Y6 were scored for all task conditions except EC and EO for each individual. A one-way ANOVA was conducted to test if there were significant changes in stress levels between task conditions. A high STAI-Y6 score indicates a higher level of perceived stress. Similarly, each of the 19 HRV features and ECG-derived respiratory frequency was plotted in box plots to visualize the variance across participants and conditions, and a Shapiro-Wilk test was applied for each feature set to ensure the assumption of Gaussian distribution was met. A repeated measure one-way ANOVA was then performed for every feature during the seven task conditions. Post-hoc for pair-wise comparisons with Tukey's comparison test was conducted to identify the groups of tasks differing from each other.

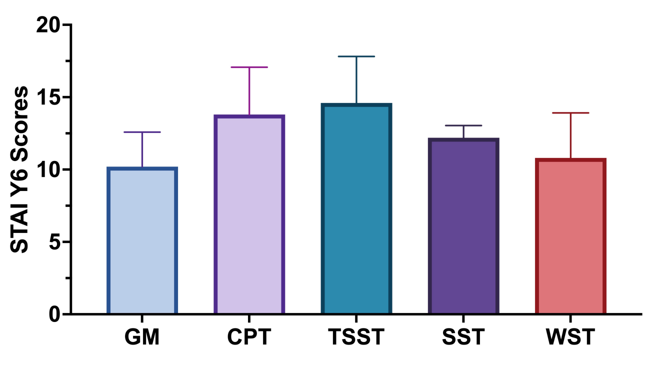

Fig. 1: Changes in post-test STAI-Y6 scores for each task condition. STAI-Y6 scores were not collected for EC and EO conditions. GM: Guided Meditation; CPT: Cold Pressor Test; TSST: Trier Social Stress Test; SST: Seated Stroop Test; WST: Walking Stroop Test.

D. Classifier Models

We implemented binary and multi-class classification using multiple ML models to explore the potential of classifying task conditions using HRV features. For the binary classification, the task conditions were grouped into either 'non-stress' or 'stress' based on the nature of the task. Specifically, GM, EC, and EO were considered 'non-stress', and the other four tasks, (CPT, TSST, SST, and WST) were considered 'stress'. The ML models used were Support Vector Machine (SVM) with a linear kernel, Random Forest Trees (RF), Naive Bayes (NB), k-Nearest Neighbor (k-NN) and were trained with 5-fold cross-validation (CV). These task-wise classifiers focused on task-specific characteristics based on HRV features irrespective of participants. The sample from the same participant could end up in both training and testing sets, causing over-optimistic results. Therefore a participant-wise SVM classifier was also implemented to create a more generalizable model. Since each participant completed multiple task conditions, we adopted a grouped 5-fold cross-validation (GKCV) method for all classifiers such that the HRV features from all task conditions of the same participant were grouped together. They would appear only in the training or test set (not both) in order to avoid data leakage. This enables the assessment of how well the model performs if a new set of participant data was given. To address the issues of imbalance between non-stress (3) and stress (4) tasks, synthetic minority oversampling technique (SMOTE) was also used to over-sample the non-stress class.

III. Results

The mean and standard deviation of STAI-Y6 scores for each task condition is illustrated in Figure 1. Although no significant difference across task conditions was found, a trend of increased stress levels compared to GM is observed. Significant ANOVA results were observed in mean RR (p = 0.0134), min HR (p = 0.0257), max HR (0.0008), mean HR (p = 0.0145), approximate entropy (p = 0.0005), $\alpha 1$ (p = 0.0363), $\alpha 2$ (p = 0.014), recurrence rate (p = 0.0085), and respiratory frequency (0.0162). No other statistically significant task differences were observed. Based on post-hoc comparisons after applying Tukey correction, max HR drops during the EC condition compared to GM and CPT (p = 0422, p = 0.0194, respectively), while it increases in the TSST and WST conditions contrasted to EC (p = 0.0107, p = 0.0315, respectively). EO also had lower max HR compared to CPT (p = 0.0254). Furthermore, mean HR is lower during SST compared to both TSST (p = 0.0369) and WST (p = 0.0337).

For non-linear features, $\alpha 2$ exhibits an increase in TSST, SST, and WST compared to the EC condition (p = 0.0131, p = 0.0369, p = 0.0222, respectively), while WST also had higher $\alpha 2$ than EO condition (p = 0.0471). Compared to GM and WST, EO has a lower recurrence rate (p = 0.0258, p = 0.0964, respectively). SST also has a lower recurrence rate compared to WST (p = 0.0265). Interestingly, the baseline conditions exhibit an overall reduction of ApEn in contrast to each of the three stressors conditions. Specifically, GM, CPT, EC, and EO are all lower in ApEn when compared to TSST, SST, and WST (GM vs. TSST: p = 0.0236; GM vs. SST: p = 0.0037; GM vs. WST: p = 0.0025; CPT

TABLE I: Features extracted from ECG data from wearable device [12].

Time-domain features			Frequency-domain features			Non-linear features		
Feature	Unit	Description	Feature	Unit	Description	Feature	Unit	Description
Mean RR	ms	Mean values of RR intervals	LF	ms^2	Absolute powers of low-frequency band (0.04-0.12 Hz); represent autonomic activity	SD1	ms	SD of Poincare plot perpendicular to the line-of-identity; represents long-term variability
SDNN	ms	Standard deviation of Normal-to-Normal intervals	HF	ms^2	Absolute powers of high-frequency band (0.12-0.4 Hz); represents mostly parasympathetic activity	SD2	ms	SD of Poincare plot along the line of identity; represents long-term variability
Min HR	bpm	minimum value of heart rate	LF (n.u.)	no unit	Normalized LF power	SD2/SD1	no unit	ratio between SD2 and SD1
Max HR	bpm	maximum value of heart rate	HF (n.u.)	no unit	Normalized HF power	DFA $\alpha 1$	no unit	Detrended fluctuation analysis; represents short-term fluctuation
Mean HR	bpm	mean value of heart rate	LF/HF	no unit	ratio of LF to HF power; represents autonomic system activity balance	DFA $\alpha 2$	no unit	Detrended fluctuation analysis; represents long-term fluctuation
RMSSD	ms	root mean square of differences between successive RR interval				ApEn	no unit	Approximate entropy; represents the complexity and regularity of HRV patterns
						SampEn	no unit	Sample entropy; similar to ApEn that also measures the complexity and regularity of HRV patterns
						REC	no unit	Recurrence plot; measures recurrent patterns and dynamic changes

vs. TSST: p = 0.0167; CPT vs. SST: p = 0.0014; CPT vs. WST: p = 0.0032; EC vs. TSST: p = 0.0374; EC vs. SST: p = 0.0057; EC vs. WST: p = 0.0108; EO vs. TSST: p = 0.0190; EO vs. SST: p = 0.0022; EO vs. WST: p = 0.0041). Within baseline conditions, GM revealed high ApEn values compared to CPT and EO conditions. Lastly, the respiratory frequency is higher in TSST compared to GM (p = 0.0485), in SST compared to CPT (p = 0.0296), and in WST compared to EC (p = 0.0123) and EO (p = 0.0093). The binary and multi-class classification results are shown in Table II to compare the performance of different classifiers. Although SVM performed worse in the task-wise binary classification, it had the highest performance metrics for participant-wise binary (accuracy = 77.1%) and multi-class classification (accuracy = 45.7%). kNN was the best classifier for task-wise classification, while its performance dropped for other types of classification, and RF and NB had similar performance scores in three classification situations.

TABLE II: Average performance metrics for binary and multiclass task-wise and participant-wise classifier using 5-fold cross-validation and grouped 5-fold cross-validation.

Task-wise binary classification				
Classifier	Accuracy	Precision	Recall	F1 Score
SVM	0.714	0.733	0.800	0.760
RF	0.743	0.814	0.700	0.755
k-NN	**0.800**	**0.790**	**0.900**	**0.817**
NB	0.743	0.728	0.850	0.740
Participant-wise binary classification				
Classifier	Accuracy	Precision	Recall	F1 Score
SVM	**0.771**	**0.724**	**0.742**	**0.692**
RF	0.742	0.712	0.733	0.696
k-NN	0.743	0.644	0.717	0.657
NB	0.686	0.632	0.675	0.611
Task-wise multiclass classification				
Classifier	Accuracy	Precision	Recall	F1 Score
SVM	**0.229**	**0.205**	**0.188**	**0.188**
RF	0.171	0.144	0.161	0.140
k-NN	0.057	0.042	0.067	0.047
NB	0.114	0.079	0.126	0.090
Participant-wise multiclass classification				
Classifier	Accuracy	Precision	Recall	F1 Score
SVM	**0.457**	**0.317**	**0.457**	**0.359**
RF	0.371	0.200	0.371	0.245
k-NN	0.314	0.143	0.314	0.189
NB	0.343	0.187	0.343	0.228

IV. Discussion

This study examined changes in STAI and HRV feature distributions across multiple stressors and the feasibility of using combining ML models to assess binary and multi-class classification performance using a wearable device.

It is noteworthy that no task effects in STAI scores were found, and WST and GM were found to have similar mean scores. This could be due to the difference in sensitivity of STAI scores and HRV features in response to stressors. HRV is a direct physiological measure of the autonomic system, which may be more sensitive to slight changes in stress responses, whereas STAI is subjective to perceived stress level [13]. Second, the STAI-Y6 is given at the end of the task condition, so the scores may not capture the transient responses during the task, and immediate changes in HRV features are not necessarily reflected in STAI-Y6. Lastly, it is worth mentioning given the small sample size in our study, large inter-individual variance was expected and can confound the trend reported by these preliminary results. This variance could be further influenced by inconsistent correlations between self-perceived stress levels, coping style, and physiological responses across individuals [14], [15].

Moreover, this study presents preliminary results of using HRV features to classify non-stress vs. stress conditions with different classification methods. The task-wise classification focuses on distinguishing the task conditions by identifying task-specific characteristics without considering individual differences. The performance of the task-wise classification is notable in that the linear SVM achieved an accuracy of 80%, suggesting that the model is capable of recognizing patterns in HRV features that distinguish non-stress and stress states. The participant-wise binary classification, on the other hand, takes individual differences into account to differentiate between non-stress and stress conditions, enabling a more generalized model for new individuals' data that has not been seen before. As expected, it performed better than task-wise multiclass classification to discern different stressors.

There are several limitations to this study. First, the preliminary study includes only 5 participants that are not representative of the general population, which can lead to model overfitting. Second, the HRV features are extracted from the entire duration of the task, which presents static pictures, lacking temporal dynamics and transient changes. Moreover, only a single modality, cardiac activity, was used. Stress responses, however, are well-coordinated activities mediated by the autonomic system that affect the whole body. Using HRV alone may not comprehensively register how humans react to different stressors. Lastly, the ground-truth of classification is solely based on the assumed nature of stressors (i.e., we explicitly assume GM, EC and EO induce low-stress, and CPT, TSST, SST, and WST as high-stress conditions). The assumption may not hold true for all participants due to differences in perception and appraisal of tasks, warranting future investigation on the correlation between perceived stress and HRV measures.

V. Conclusions and Future Work

The preliminary results provide insight into HRV patterns in response to different stressors and the potential for using participant-wise classifications to distinguish various types of stressors. Future research with larger sample sizes could help clarify and refine the model to enhance performance across all individuals. Using a time-varying analysis that incorporates sliding windows to capture temporal dynamics would also likely better capture the alteration of HRV over time [16].

Acknowledgment

We would like to thank the members of the Mobility and Fall Prevention Research Lab for their assistance with data collection and all of the subjects who participated in this study. This project has been funded by the Jump ARCHES endowment through the Health Care Engineering Systems Center.

References

[1] T. R. Frieden, D. W. H. Jaffe, S. B. Thacker, R. L. Moolenaar, J. C. Martinroe, S. R. Spriggs, T. M. Starr, Q. M. Doan, P. H. King, W. L. Roper, D. Holtzman, G. K. J. Iglehart, D. G. Maki, W. P. Quinlisk, D. Moines, I. L. P. Remington, W. K. B. Rimer, C. Hill, N. V. J. Rullan, S. Juan, P. W. Schaffner, T. A. Schuchat, G. E. D. Snider, and G. W. J. Ward, "Morbidity and mortality weekly report mental illness surveillance among adults in the united states centers for disease control and prevention mmwr editorial and production staff mmwr editorial board," 2011.

[2] J. A. Chalmers, D. S. Quintana, M. J. Abbott, and A. H. Kemp, "Anxiety disorders are associated with reduced heart rate variability: A meta-analysis," *Frontiers in Psychiatry*, vol. 5, 2014.

[3] K. M. Lukasik, O. Waris, A. Soveri, M. Lehtonen, and M. Laine, "The relationship of anxiety and stress with working memory performance in a large non-depressed sample," *Frontiers in Psychology*, vol. 10, 1 2019.

[4] M. Berger and Z. Sarnyai, ""more than skin deep": Stress neurobiology and mental health consequences of racial discrimination," pp. 1–10, 1 2015.

[5] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Computer Methods and Programs in Biomedicine*, vol. 108, pp. 1287–1301, 12 2012.

[6] B. A. Hickey, T. Chalmers, P. Newton, C. T. Lin, D. Sibbritt, C. S. McLachlan, R. Clifton-Bligh, J. Morley, and S. Lal, "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review," 5 2021.

[7] P. Schmidt, A. Reiss, R. Duerichen, and K. V. Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection." Association for Computing Machinery, Inc, 10 2018, pp. 400–408.

[8] C. M. Smith, S. N. Chillrud, D. W. Jack, P. Kinney, Q. Yang, and A. M. Layton, "Laboratory validation of hexoskin biometric shirt at rest, submaximal exercise, and maximal exercise while riding a stationary bicycle," *Journal of Occupational and Environmental Medicine*, vol. 61, pp. E104–E111, 4 2019.

[9] J. Montes, J. C. Young, R. Tandy, and J. W. Navalta, "Reliability and validation of the hexoskin wearable bio-collection device during walking conditions," pp. 806–816, 2018. [Online]. Available: http://www.intjexersci.com

[10] D. Bläsing, A. Buder, J. E. Reiser, M. Nisser, S. Derlien, and M. Vollmer, "Ecg performance in simultaneous recordings of five wearable devices using a new morphological noise-to-signal index and smith-waterman-based rr interval comparisons," *PLoS ONE*, vol. 17, 10 2022.

[11] K. A. Knowles and B. O. Olatunji, "Specificity of trait anxiety in anxiety and depression: Meta-analysis of the state-trait anxiety inventory," 12 2020.

[12] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, 2017. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2017.00258

[13] R. K. Dishman, Y. Nakamura, M. E. Garcia, R. W. Thompson, A. L. Dunn, and S. N. Blair, "Heart rate variability, trait anxiety, and perceived stress among physically fit men and women," *International journal of psychophysiology*, vol. 37, no. 2, pp. 121–133, 2000.

[14] B. Fuller, "The effects of stress-anxiety and coping styles on heart rate variability," *International Journal of Psychophysiology*, vol. 12, no. 1, pp. 81–86, 1992.

[15] A. V. Klinkenberg, U. M. Nater, A. Nierop, A. Bratsikas, R. Zimmermann, and U. Ehlert, "Heart rate variability changes in pregnant and non-pregnant women during standardized psychosocial stress," *Acta obstetricia et gynecologica Scandinavica*, vol. 88, no. 1, pp. 77–82, 2009.

[16] C. F. Goss and E. B. Miller, "Dynamic metrics of heart rate variability," *arXiv preprint arXiv:1308.6018*, 2013.