Stress and Recovery Tracking Using Consumer Wearables: A Pilot Study

*Course: 17-320 / 17-720 — Machine Learning and Sensing for Healthcare

1st Yuki Yang
Carnegie Mellon University
Pittsburgh, PA, USA
yukiyang@andrew.cmu.edu

2nd Alfredo Romo Osorno
Carnegie Mellon University
Pittsburgh, PA, USA
aromooso@andrew.cmu.edu

3rd Fernando Ruiloba Portilla
Carnegie Mellon University
Pittsburgh, PA, USA
fruiloba@andrew.cmu.edu

***Abstract*—Continuous and unobtrusive stress monitoring has the potential to transform mental-health assessment by providing early indicators of burnout and impaired recovery. While consumer wearables such as smartwatches and rings offer widespread access to physiological and behavioral sensing, their real-world reliability for stress tracking remains insufficiently understood. In this pilot study, we evaluate the feasibility of using multi-modal data from consumer devices, including heart rate variability (HRV), sleep architecture, and proprietary daily stress scores to detect and predict day-level stress. We construct a unified dataset combining three months of longitudinal Garmin and Apple Watch recordings, combined with the SWELL public stress dataset and a controlled "jump-scare" experiment designed to isolate contextual stressors. Using this dataset, we train both classification and regression models to assess stress states and quantify continuous stress levels. Our results show that wearables can capture meaningful fluctuations in physiological stress: binary stress-day classification achieved high recall ($> 87\%$) despite modest accuracy, and a Random Forest regressor predicted daily stress scores with an MAE of approximately 3 points. Importantly, contextual information dramatically improved predictive performance in controlled settings, underscoring the ambiguity of HRV signals when interpreted in isolation. These findings demonstrate both the promise and limitations of consumer wearables for real-world stress tracking and highlight the need for context-aware modeling to unlock accurate and reliable stress monitoring in everyday environments.**

***Index Terms*—mHealth, wearable sensing, stress detection, machine learning, HRV**

I. Introduction

Chronic stress and individuals' capacity for recovery play a pivotal role in both mental and physical health. Traditional stress assessment approaches, such as self-report questionnaires or occasional clinical visits are inherently limited: they are episodic rather than continuous, rely on subjective recall, and often fail to capture daily or intra-day fluctuations in stress and recovery. As a result, subtle but persistent patterns of stress accumulation, or deficits in recovery, may go undetected. In contrast, consumer-grade wearable devices, such as smartwatches, smart rings, and fitness trackers have become increasingly common. These devices enable continuous, unobtrusive monitoring of physiological signals (e.g., heart rate variability, resting heart rate) and behavioral / lifestyle metrics like sleep duration and quality. This technological trend offers a promising opportunity for long-term, real-world monitoring of stress and recovery dynamics, potentially bridging gaps left by traditional assessments [1,2].

In this study, we aim to investigate whether consumer-grade wearables can be used to reliably detect and predict daily stress and recovery levels in real-world settings. Specifically, we address the following research questions:

1) Can daily aggregated features derived from wearable data (e.g., HRV, sleep metrics) reliably classify days into "high-stress" vs "low/normal-stress"?
2) Beyond binary classification, can wearable-derived features support prediction of *continuous* stress levels (i.e., a regression approach)?
3) Does the integration of multiple data sources, physiological (HRV), behavioral (sleep), and subjective self-report enhance prediction performance, compared to using HRV alone?

The main contributions of this work include:

- Construction of a unified daily-scale dataset combining multi-device, multi-modal data (HRV, sleep, self-report) from consumer wearables.
- Implementation of both *classification* (stress-day detection) and *regression* (continuous stress prediction) models, providing more granular and practical utility for stress monitoring in everyday life.
- Use of daily-scale aggregation over longitudinal periods, offering real-world ecological validity, rather than short-term or lab-based snapshots common in prior stress studies.
- Exploration of the *feasibility and limitations* of consumer-grade wearables for continuous stress and recovery tracking, providing empirical evidence for both strengths and constraints.

The remainder of this report is organized as follows. In Section 2, we review relevant literature on wearable-based stress monitoring, including sensor validity, common methodologies, and limitations. Section 3 describes the system design, data collection procedures, and feature engineering pipeline.

Section 4 presents our experimental results and evaluation, assessing both classification and regression performance. Finally, we discuss implications, limitations, and future directions.

II. Related Work

A. Wearable Sensor–Based Stress and Health Monitoring: Overview

Recent years have witnessed a growing body of research leveraging wearable sensors to monitor physiological and behavioral signals for stress, mental health, and general well-being tracking. A recent systematic review synthesized findings across over 60 peer-reviewed studies from 2016 to 2025, covering a variety of wearable devices (smartwatches, wristbands, rings) and sensor modalities such as heart rate, heart rate variability (HRV), electrodermal activity (EDA), skin temperature, and sleep metrics [1]. These studies commonly target applications such as stress detection, anxiety monitoring, resilience tracking, and general wellness management.

B. Validity and Reliability of Wearable-Derived Physiological Signals (HRV, PPG, etc.)

For wearable-based stress monitoring to be meaningful, the physiological signals measured by consumer devices must be sufficiently accurate. Several studies have validated HRV (and related metrics) derived from photoplethysmography (PPG) or wearable sensors against gold-standard electrocardiogram (ECG)-based metrics. For instance, a recent investigation demonstrated that certain wearables produce resting HRV measurements that correlate with clinical health and well-being indicators, particularly when recordings are taken during sleep or upon waking (conditions that minimize motion and external confounders) [3]. Another validation study comparing multiple popular wearables found that HR and HRV measurements show variable agreement levels across devices: devices like finger-worn rings often outperform wrist-worn smartwatches under nocturnal or resting conditions, while wrist devices may show degraded reliability when users are active or during daytime activities [4,5].

These findings suggest that while consumer wearables can produce HRV metrics of reasonable fidelity under resting or sleep conditions, the accuracy for HRV during daily living (especially when motion or activities are involved) remains a concern [6]. Such variability raises caution in interpreting wearable-derived HRV as a direct proxy for autonomic nervous system (ANS) state without careful context consideration.

C. Empirical Studies: Stress Detection / Prediction Using Wearables + Machine Learning

Beyond physiological measurement validation, a number of empirical studies have attempted to detect or predict stress (or stress-related states) using wearable-derived data combined with machine learning (ML) or statistical modeling. For example, recent research using wearable biosensors in a large cohort (hundreds of participants) applied recurrent neural networks (e.g., LSTM) to HRV time series and demonstrated associations between wearable HRV and self-reported stress, anxiety, and general health status [7]. Another study specifically designed a PPG-based wearable stress detection device: after validating the PPG-derived RR intervals against ECG references, the authors trained classification models that achieved promising stress-detection performance under controlled conditions [8].

More recently, a survey of wearable-based stress detection techniques identified HRV, PPG, EDA as the most commonly used signals, and machine learning models (ranging from random forests to deep neural networks) as mainstream modeling approaches [9]. However, many of these studies suffer from limitations such as small sample sizes, short monitoring periods, and reliance on lab-based stress induction protocols, which limit their ecological validity.

D. Multi-Modal Longitudinal / Real-World Monitoring: Sleep, HRV, EMA, and Recovery

Recognizing the limitations of lab-based studies, a subset of work has focused on real-world, longitudinal monitoring by combining physiological data (HRV, HR), behavioral data (sleep), and ecological momentary assessment (EMA) or self-reported stress. Such multi-modal, real-life studies offer higher ecological validity and capture the day-to-day variability of stress and recovery. For example, one study demonstrated that wearable-derived nocturnal HRV and resting heart rate (RHR) correlate with certain mental health and behavioral indicators over extended monitoring periods — suggesting potential of wearables as general health biomarkers [3]. However, variability across individuals, missing data, and inconsistent associations (e.g., sleep quality correlates only weakly with subjective stress) remain common challenges.

E. Challenges, Limitations, and Gaps in Current Literature

Despite progress, existing literature reveals several recurrent limitations:

- Many validation studies restrict consideration to resting or sleep conditions; accuracy degrades in daily living with motion and activity.
- Sample sizes are often small, and monitoring periods short; limiting longitudinal and large-scale inference.
- Heterogeneity in devices, sensor modalities, data preprocessing, and HRV computation undermines comparability across studies.
- Stress labeling is often inconsistent: some studies rely on self-report (subjective), others on physiological proxies complicating interpretation.
- Few studies integrate multi-modal data (physiology + sleep + self-report) over long-term real-world monitoring and adopt both classification and continuous prediction approaches.

F. Gap Assessment Positioning of Current Study

Given the limitations and heterogeneity in prior work, there is a clear need for studies that:

- Build *unified, multi-device, multi-modal datasets* covering HRV, sleep, and self-report over longitudinal daily-scale monitoring.

- Evaluate both *classification* (stress-day detection) and *regression* (continuous stress level prediction) to support flexible applications.
- Examine real-world feasibility including accuracy, noise, missing data, and device variability rather than idealized lab conditions.

Our current study aims precisely to address these gaps, offering empirical evidence for the feasibility and limitations of wearable-based stress and recovery monitoring in real-life contexts.

III. Methods

To evaluate the feasibility of consumer-grade wearables for stress tracking, we designed a comprehensive data processing pipeline that ingests raw sensor data, extracts physiological features, and applies machine learning models for both classification and regression tasks.

A. Data Collection and Sources

We utilized a multi-source approach to gather physiological and behavioral data:

1) **Primary Wearable Data (Longitudinal):** We collected three months of empirical data using Garmin and Apple Watch devices. The primary data streams included raw Heart Rate Variability (HRV), resting heart rate, and sleep stages. Data was exported in the proprietary binary `.FIT` format, necessitating specialized extraction tools.
2) **Public Dataset (SWELL-KW):** To augment our analysis with labeled stress events, we incorporated the SWELL Knowledge Work dataset. This dataset provides multimodal data (computer logging, facial expressions, body postures) and physiological signals (HRV, skin conductance) from 25 subjects performing knowledge work under varying stressors (e.g., email interruptions, time pressure).
3) **Controlled Validation Dataset (Jump Scare Experiment):** To test the hypothesis that contextual features improve stress prediction, we created a validation dataset involving subjects watching 12 horror movies. We recorded HRV differences, the count of "jump scares" (contextual stressor), and perceived stress levels.

B. Data Processing Pipeline

The raw data processing involved several engineering steps to create a unified analysis-ready dataset:

1) Extraction and Parsing: We utilized `FitCSVTool.jar` to convert proprietary Garmin `.FIT` files into readable `.CSV` formats. Custom Python scripts were developed to parse dynamic schemas, map field values, and handle timestamp conversions.

2) Preprocessing and Cleaning: Data cleaning involved handling null values and converting timestamps to a uniform datetime format. We performed daily aggregation to merge disparate data streams, specifically aligning HRV metrics and Sleep metrics on a daily axis to correspond with the proprietary "Daily Stress Score" provided by the devices.

3) Feature Engineering: We extracted and engineered specific features known to correlate with autonomic nervous system (ANS) activity:

- **HRV Metrics:** Daily Mean, Median, Minimum, Maximum, and Standard Deviation (SD). HRV serves as the gold standard for ANS balance.
- **Sleep Metrics:** Total sleep hours, Deep Sleep percentage, and REM sleep percentage. These metrics were chosen as sleep composition often dictates recovery capacity more accurately than duration alone.
- **Target Variable:** The "Daily Stress Score" (0-100) estimated by the wearable's proprietary algorithm served as our ground truth for modeling.

C. Machine Learning Models

We implemented two distinct modeling approaches using `scikit-learn`:

1) **Binary Classification (Stress Detection):** We trained a Logistic Regression model to classify days into "High Stress" vs. "Low Stress/Recovery." Features were standardized using `StandardScaler`.
2) **Regression (Continuous Prediction):** We employed a Random Forest Regressor to predict the continuous Daily Stress Score. This model was selected for its ability to handle non-linear relationships and robustness with smaller datasets.
3) **Contextual Validation:** For the Jump Scare experiment, we compared two Linear Regression models—one using HRV alone and one combining HRV with the "Jump Scare Count"—to quantify the value of situational context.

IV. Results

A. Classification Performance

The binary classification models aimed to distinguish between high-stress and low-stress days.

- **Garmin Data:** The Logistic Regression model achieved an overall accuracy of 60% with a recall of 0.875.
- **Apple Watch Data:** Performance was comparable, with an accuracy of 59% and a recall of 0.82.

While the overall accuracy leaves room for improvement, the high recall indicates that the system is highly sensitive to stress; it rarely misses a high-stress day, making it an effective screening tool for potential burnout risk.

B. Regression Analysis

The Random Forest regressor, trained to predict the specific daily stress score (0–100), achieved an R^2 Score of 0.514 and a Mean Absolute Error (MAE) of 3.02.

- **Visual Analysis:** The regression predictions successfully captured the general trend of stress fluctuations over the three-month period, typically predicting the daily score within a margin of ± 3 points.
- **Feature Importance:** Our exploratory analysis revealed that simple sleep duration was not predictive of stress.

However, sleep composition (specifically the ratio of Deep Sleep to REM Sleep) added valuable context to the recovery models, correlating more strongly with the subsequent day's stress score.

C. Impact of Contextual Features

Our controlled validation study (Jump Scare Experiment) demonstrated the critical role of context in stress prediction. We compared two Leave-One-Out Cross-Validation (LOOCV) models:

- **Model A (HRV Only):** Performed poorly with an R^2 of -0.64, indicating HRV changes alone could not fully explain the variance in perceived stress during the movies.
- **Model B (HRV + Context):** By adding the context of "Jump Scare Count," the model achieved strong predictive power with an R^2 of 0.90 and an MAE of ≈ 2.2.

V. Discussion

A. Robustness of Physiological Biomarkers

Our findings validate that Heart Rate Variability (HRV) is a robust biomarker for passive stress detection in real-world settings. The strong correlation found in our regression analysis confirms that consumer wearables can capture the physiological "cost" of daily stressors. Furthermore, our results highlight the nuance of sleep tracking; total sleep time is a poor predictor of recovery compared to sleep architecture (Deep vs. REM cycles), suggesting that future interventions should focus on sleep quality optimization rather than just duration.

B. The Necessity of Context

The stark contrast in performance between the HRV-only model and the HRV+Context model in our Jump Scare study (R^2 of -0.64 vs. 0.90) is a significant finding. It suggests that physiological signals alone are often ambiguous; a drop in HRV could signal distress (fear) or eustress (excitement). Integrating situational features potentially derived from calendar data, location, or app usage is essential for disambiguating these states and building high-fidelity stress prediction systems.

C. Viability as a Screening Tool

Despite a modest classification accuracy of $\approx 60\%$, the system's high recall ($> 87\%$) demonstrates its viability as a burnout screening tool. In a mental health context, false negatives (missing a high-stress day) are often more dangerous than false positives. Our model effectively flags the majority of high-stress days, which could trigger "preventative" interventions (e.g., mindfulness prompts) that are beneficial even if the user is not critically stressed.

D. Limitations

- **Data Parsing & Engineering:** A significant portion of the project timeline was consumed by engineering challenges related to proprietary binary formats (.FIT), reducing time available for model tuning.
- **Sample Size:** Merging disparate data streams from multiple sources resulted in a limited number of perfectly overlapping "complete" days, constraining the complexity of models we could train without overfitting.
- **Ablation Studies:** Due to time constraints, our feature ablation study was not exhaustive; other combinations of sensor inputs might yield higher accuracy.

VI. Conclusion

This pilot study demonstrates that consumer-grade wearables are a feasible platform for continuous, longitudinal stress monitoring. We successfully engineered a pipeline to transform raw, proprietary sensor data into predictive insights, achieving a regression model capable of tracking stress trends with reasonable precision (MAE ≈ 3) and a classification system with high sensitivity (Recall $> 87\%$).

Our work highlights two critical paths for the future of mHealth:

1) **Context Integration:** Physiological data must be augmented with situational context (as proven by our Jump Scare validation) to achieve high predictive accuracy.
2) **Advanced Modeling:** Future work should move beyond daily aggregation and leverage Deep Learning, specifically Long Short-Term Memory (LSTM) networks, to exploit the sequential nature of time-series data. Additionally, incorporating Electrodermal Activity (EDA) and Skin Temperature would provide a more holistic view of the autonomic nervous system.

Ultimately, this system serves as a proof-of-concept for a proactive mental health tool that moves beyond simple tracking to provide timely, data-driven warnings for burnout prevention.

References

[1] Kapogianni, N.-A., Sideraki, A., & Anagnostopoulos, C.-N. (2025). Using Smartwatches in Stress Management, Mental Health, and Well-Being: A Systematic Review. *Algorithms*, 18(7), 419. https://doi.org/10.3390/a18070419

[2] Altini, M., & Kinnunen, H. (2021). The Promise of Sleep: A Multi-Sensor Approach for Accurate Sleep Stage Detection Using the Oura Ring. *Sensors*, 21(13), 4302. https://doi.org/10.3390/s21134302

[3] Jerath, R., Syam, M., & Ahmed, S. (2023). The Future of Stress Management: Integration of Smartwatches and HRV Technology. *Sensors*, 23, 7314. https://doi.org/10.3390/s23177314

[4] Pinge, A., et al. (2024). Detection and Monitoring of Stress Using Wearables: A Systematic Review. *Frontiers in Computer Science*. https://doi.org/10.3389/fcomp.2024.1478851

[5] Herberger, S., et al. (2025). Performance of wearable finger-ring trackers for diagnostic sleep staging: A validation study. *Scientific Reports*. (online first) https://doi.org/10.1038/s41598-025-93774-z