
Report

Capturing the Stress Jump: Predicting Physiological Reactivity from Empatica E4 Signals

Rebecca Metallo: *BSc in Mathematical and Computing Sciences for Artificial Intelligence*

Giulia Caffi: *MSc in Medicine and Surgery*

Anna Uberti: *BSc in Medical Radiology Techniques*

r.metallo@studenti.unisr.it
g.caffi@studenti.unisr.it
a.uberti1@studenti.unisr.it

Abstract

BACKGROUND: Psychological stress is associated with characteristic autonomic responses that can be captured through wearable physiological sensors. Recent advances in wearable technology enable continuous monitoring of physiological signals, raising interest in their potential to quantify individual stress reactivity in controlled experimental settings.

OBJECTIVES: This study aims to investigate whether within-subject changes in stress relative to baseline can be predicted using short segments of physiological signals recorded by the Empatica E4 wristband during a Stress Induction Protocol.

METHODS: Multimodal physiological signals were processed to extract summary statistics and heart rate variability features. Random Forest, Gradient Boosting, and XGBoost models were applied using both regression and classification formulations, with hold-out and Leave-One-Subject-Out validation.

RESULTS: Regression-based approaches showed limited and unstable performance in predicting continuous stress reactivity, even after baseline normalisation. Reformulating the task as a binary classification problem resulted in more stable, with accuracy and discriminative ability only slightly above chance. Feature importance analyses indicated a predominance of anthropometric over physiological stress-related features.

Keywords: Stress, Wearable Devices, Physiological Signals, Heart Rate Variability, Empatica E4.

1 Introduction and Dataset Description

Psychological stress elicits a well-characterised autonomic response, primarily involving sympathetic activation. This response is associated with increased heart rate, reduced heart rate variability, elevated electrodermal activity, and peripheral vasoconstriction leading to lower skin temperature [1, 2]. These physiological changes have long been used as indirect markers of acute stress in both laboratory and ambulatory settings.

Recent advances in wearable technology allow continuous monitoring of such signals in real-world conditions. Devices such as the Empatica E4 wristband provide ac-

cess to multimodal physiological data, raising the question of whether these signals contain sufficient information to quantify individual stress reactivity.

This project examines whether physiological data from the Empatica E4, combined with a limited set of subject-related variables, can really capture individual stress reactivity. The analysis is guided by the following research question:

Is it possible to predict how much an individual becomes stressed relative to their own baseline level using physiological signals recorded during the Stress Induction Protocol?

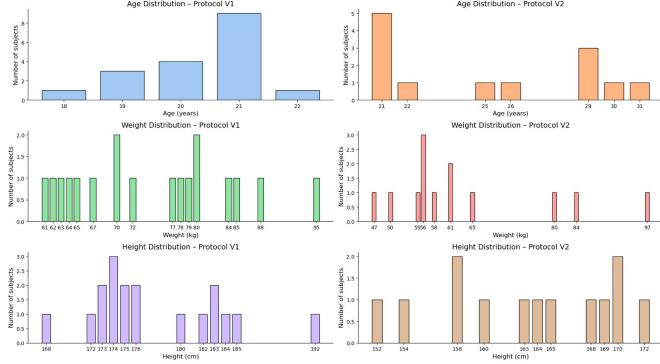


Figure 1: Exploratory Data Analysis' Trends

2 Materials and Methods

Although data were collected from **Stress**, **Aerobic**, and **Anaerobic** protocols, for this project only the **Stress** session was analysed. The analysis was complemented by the files `stress_level_v1.csv`, `stress_level_v2.csv`, and `subject_info.csv`, which provide information on stress levels, age, gender, height, weight, and physical activity.

Exploratory Data Analysis

The descriptive analysis highlighted clear differences between the two versions of the protocol (See Figure 4 for details in protocols).

Participants in Protocol V2 ranged from 18 to 31 years of age, whereas those in Protocol V1 were younger (18–22 years). The sex distribution also differed substantially: Protocol V1 included only male participants, while Protocol V2 consisted mostly of females. (Figure 1)

Anthropometric measures reflected this composition. Female participants typically weighed 47–65 kg and measured 152–172 cm in height, while male participants weighed 61–88 kg and measured 165–192 cm.

Physical activity levels also varied across protocols: in Protocol V1, 94% of subjects reported engaging in regular physical activity, compared with 22% in Protocol V2, where 50% declared no activity and 28% did not provide this information. (Figure 2)

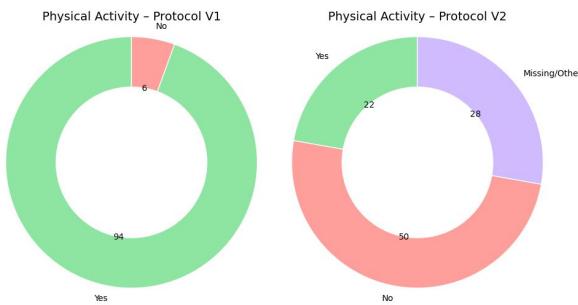


Figure 2: Physical Activity Levels

Self-reported stress ratings showed also substantial variability across protocol phases and between participants (Figure 3). As expected, the TMCT consistently elicited the largest increase in perceived stress in both versions of

the protocol, whereas Baseline and Rest phases produced lower ratings [3].

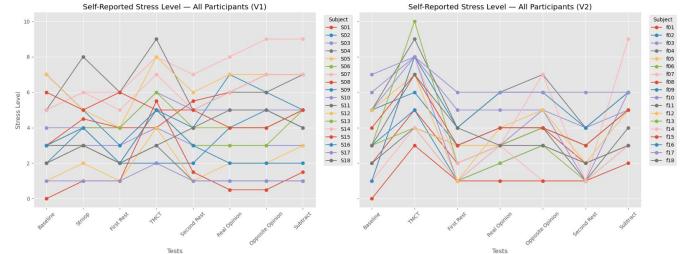


Figure 3: Self-Reported Stress Ratings

Feature Engineering

In Figure 4 are shown the two versions of the Stress Protocol included in the dataset. Each protocol consists of a sequence of cognitive and social stressors alternated with rest periods, during which participants rated their perceived stress level (SL). These time-stamped phases serve as reference markers for the segmentation of the physiological signals.

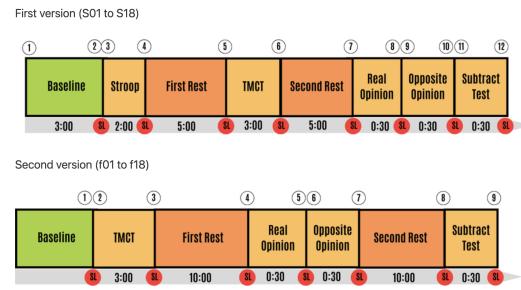


Figure 4: Stress Protocols

To construct the dataset used for modelling, each Empatica E4 signal (EDA, BVP, ACC, TEMP, HR) (See Table 1) was aligned to the corresponding protocol timeline. Signals were then segmented into windows matching the exact duration of each task (Baseline, Stroop/TMCT, Rest periods, Opinion tasks, Subtraction).

Non-relevant portions of the signals were removed, and for each window within each stage of each protocol the mean and standard deviation of every channel were computed.

Participants with more than 15% missing data, as well as those showing evident measurement artefacts, were excluded from the analysis. Specifically, the following subjects were removed: S02, f07, f14_a, f14_b, f14, f15, f16, f17, and f18. The resulting initial dataset did not include IBI values for participants.

In this project, two main datasets were employed.

- The first corresponds to the feature set described above, based solely on summary statistics (mean and standard deviation) of each physiological signal.
- A second dataset was later constructed to incorporate more physiologically informative markers, motivated by the need for more accurate indicators. This extended dataset includes properly segmented IBI

values and a set of Heart Rate Variability (HRV) metrics derived from them (e.g. RMSSD, pNN50, CVRR, CVSD) (See Table 2) [4], [5].

For both datasets, feature selection was guided by the *Minimum Redundancy Maximum Relevance* (mRMR) criterion, ensuring the retention of features with high predictive relevance while avoiding highly correlated or redundant variables.

The choice of tree-based models is consistent with recent evidence showing that Random Forest and XGBoost perform well in stress detection and monitoring tasks [10].

Model 1: Random Forest Regressor on the Initial Dataset

The first model employs a Random Forest Regressor trained on a reduced feature set obtained through mRMR, consisting of: *EDA_std*, *HR_std*, *TEMP_mean*, *ACC_norm_mean*, *Height (cm)*, *Weight (kg)*, *Physical_Activity*.

To optimise the estimator, a RandomizedSearchCV (20 iterations, 3-fold cross-validation) was applied over a pre-defined hyperparameter space (*n_estimators*, *max_depth*, *min_samples_leaf*, *min_samples_split*, *max_features*).

The objective was to predict the continuous change in stress (Δstress).

Despite modelling nonlinear interactions, the regression performance remained very limited.

Model 2: Random Forest Regressor with IBI-Derived Features

The second model extends the initial approach by integrating inter-beat interval (IBI) information and standard heart rate variability (HRV) metrics [6]. After mRMR selection, the feature set included: *EDA_std*, *HR_std*, *TEMP_mean*, *ACC_norm_mean*, *IBI_mean*, *RMSSD*, *Height (cm)*, *Weight (kg)*, *Physical_Activity*. These variables provide a richer description of autonomic cardiac dynamics compared to Model 1.

As in the previous model, hyperparameter optimisation was performed using a RandomizedSearchCV (20 iterations, 3-fold cross-validation) over *n_estimators*, *max_depth*, *min_samples_leaf*, *min_samples_split*, *max_features*.

The objective remained the prediction of the continuous stress reactivity measure (Δstress).

Although the inclusion of IBI-derived features adds physiological depth, the improvement in performance was marginal.

Model 3: Gradient Boosting Regressor on the Same IBI-Enhanced Dataset

The third model retains exactly the same feature set introduced in Model 2, and changes only the learning algorithm. A Gradient Boosting Regressor was tested to assess whether a more flexible, stage-wise ensemble method could extract nonlinear patterns that Random Forest had failed to capture [7], [8]. Hyperparameters were optimised through a RandomizedSearchCV (20 iterations, 3-fold CV).

Contrary to expectations, performance decreased compared to Model 2.

Model 4: Gradient Boosting Regressor on a Baseline-Normalized Dataset

This model introduces a more refined feature representation: for every physiological variable, the value recorded during each task phase is compared with the subject's own baseline (i.e., the Baseline stage of the protocol). These baseline-normalized deltas aim to capture individual physiological reactivity rather than absolute signal levels. . Normalising physiological signals to each subject's baseline has been proposed as a strategy to reduce inter-subject variability and improve robustness in stress detection [9]. Additional HRV-derived features (CVRR and CVSD) are also included to quantify beat-to-beat variability.

A Gradient Boosting Regressor is trained to estimate the continuous change in stress (Δstress), using Leave-One-Subject-Out validation to assess generalisation. Despite the more principled feature engineering, the model shows poor predictive performance, indicating that within-subject normalization alone is insufficient to produce a reliable regression of stress reactivity.

Model 5: XGBoost Classification on the Baseline-Normalized Dataset

Given the poor reliability of continuous regression in the previous models, the task is reframed as a binary classification problem indicating whether stress increased compared to baseline ($\Delta\text{stress} > 0$). The same baseline-normalized dataset used in Model 4 is retained, including delta features, HRV indices (CVRR, CVSD), and anthropometric variables.

An XGBoost classifier is trained to discriminate between “stress increase” and “no increase”. This reformulation is also consistent with recent wearable-stress literature, where stress is most commonly modelled as a two-class classification problem (stressed vs. not stressed) using physiological signals from wearable devices [10], [11].

This algorithm relies on gradient-boosted decision trees and is well suited for capturing nonlinear interactions in small-to-medium physiological datasets. Model performance is assessed using Leave-One-Subject-Out cross-validation to ensure generalization to unseen individuals.

Compared to regression-based approaches, this formulation yields more stable and interpretable results, although discriminative performance remains modest across subjects.

3 Results

Overall, the proposed models showed limited ability to predict individual stress reactivity from wearable-derived physiological signals.

A comparative summary of model performance across different feature sets and validation strategies is reported in Table 3.

Across all modelling approaches, predictive performance

remained modest, indicating that the extracted features capture only weakly the physiological determinants of Δ stress.

- Model 1 – Random Forest Regressor (summary statistics):** The baseline regression model achieved marginal predictive performance. Feature importance analysis indicates a strong dominance of anthropometric variables (Height and Weight), whereas classical stress-related physiological features contributed minimally, suggesting limited informative value in the initial feature set.

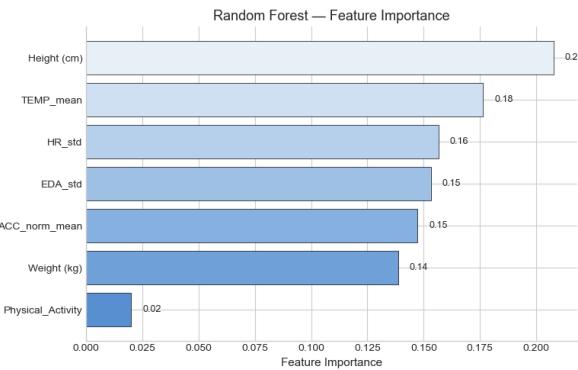


Figure 5: Feature importance for Model 1 (Random Forest on summary statistics).

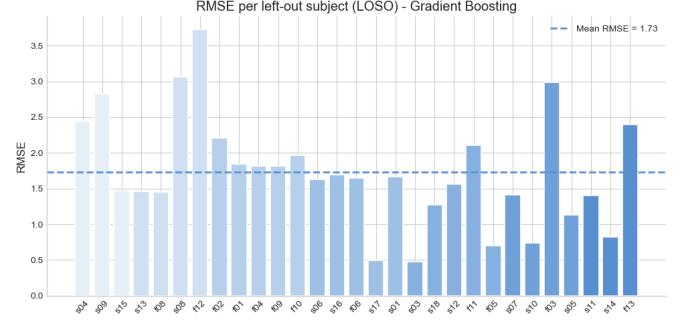


Figure 6: Distribution of RMSE across subjects for Model 4 (LOSO validation).

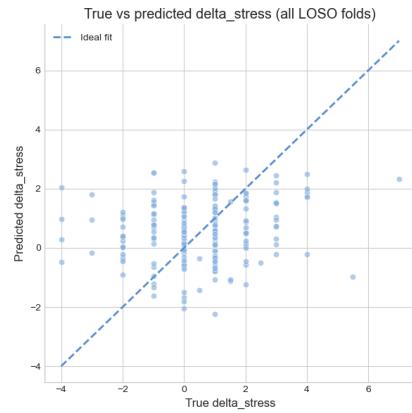


Figure 7: True versus predicted Δ stress for Model 4.

- Model 2 – Random Forest Regressor with HRV features:** The inclusion of inter-beat interval-derived HRV metrics resulted in only a slight improvement over Model 1. Despite their established physiological relevance, HRV features provided limited additional predictive power, likely due to short window durations and noise in wrist-based IBI measurements.

- Model 3 – Gradient Boosting Regressor with HRV features:** Replacing Random Forest with a Gradient Boosting Regressor did not improve performance and led to further degradation. This result indicates that the observed limitations are not algorithm-dependent but primarily related to insufficient information content in the available features.

- Model 4 – Gradient Boosting Regressor on baseline-normalized features:** Baseline normalization was introduced to explicitly target within-subject stress reactivity. However, Leave-One-Subject-Out validation yielded poor and highly variable results, with no stable predictive structure across individuals.

- Model 5 – XGBoost classification on baseline-normalized features:** Reformulating the task as a binary classification problem resulted in more stable, though still modest, performance. Discriminative ability remained only slightly above chance. Feature importance analysis suggests that the classifier relies primarily on anthropometric and lifestyle variables rather than on acute physiological changes.

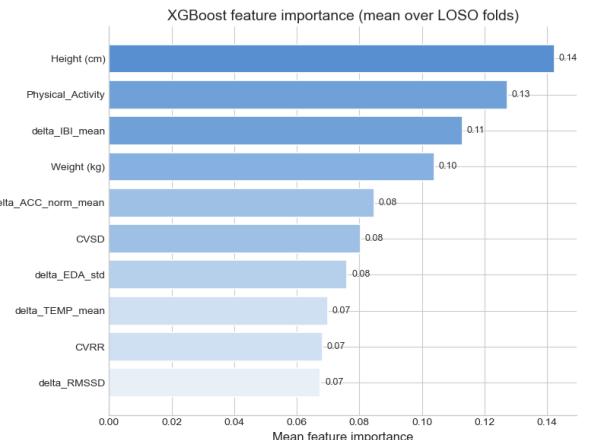


Figure 8: Feature importance for Model 5 (XGBoost classifier).

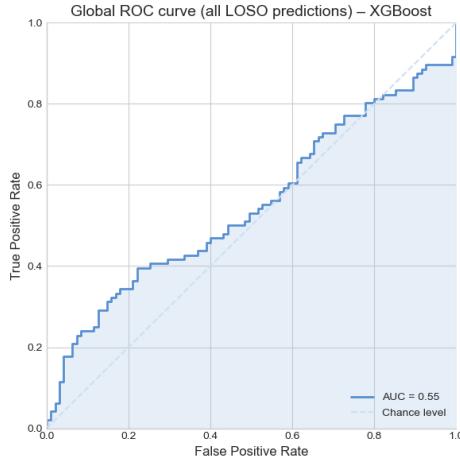


Figure 9: ROC curve for Model 5.

4 Discussion

The results of this project indicate that predicting within-subject changes in stress (Δ stress) from short segments of physiological data acquired with a wrist-worn device is highly challenging.

Physiological Interpretation

The features included in the analyses reflect canonical physiological responses to acute stress. Electrodermal activity typically increases under cognitive or emotional load due to sympathetic activation, while heart rate and inter-beat intervals shorten and heart rate variability indices such as RMSSD and pNN50 decrease, indicating reduced parasympathetic modulation [1]. Peripheral skin temperature may decrease as a consequence of vasoconstriction and anthropometric and lifestyle variables were retained because they influence baseline autonomic tone and inter-individual variability. Overall, the selected features have a clear theoretical basis and are consistent with both classical psychophysiological literature and recent wearable-based stress studies [2].

Sources of Limited Generalisation

Despite this solid physiological rationale, none of the models generalised well across participants. Several factors likely contributed to this limitation. First, stress-related physiological changes were often small relative to background variability, reducing the signal-to-noise ratio. Second, Empatica E4 signals—particularly BVP-derived heart rate and HRV—are sensitive to noise and movement artefacts, and the short duration of task segments further undermines the reliability of HRV estimates [4]. In addition, marked heterogeneity between protocol versions in terms of age, sex distribution, and physical activity levels likely introduced latent inter-subject variability. Although age and sex were excluded from the final feature set, their indirect effects on baseline physiology and stress responses may still be reflected in the recorded signals, limiting model generalisation.

Finally, Δ stress was derived from self-reported ratings, which are subjective and coarse, limiting the quality of the target variable.

Methodological Bias Toward Anthropometric Features

A recurrent and unexpected finding was the strong importance assigned to anthropometric variables, particularly Height. This effect is unlikely to reflect a genuine physiological mechanism of stress reactivity. Instead, it is more plausibly explained by methodological factors: anthropometric variables exhibit greater variability than physiological features. When physiologically meaningful features are weak or noisy, tree-based models tend to rely on these more stable but semantically irrelevant variables. This behaviour indicates that the models primarily capture dataset-specific structure rather than true stress-related physiological changes.

Limits of Baseline Normalisation and HRV Metrics

Baseline-normalised features were introduced to isolate within-subject physiological changes and reduce inter-individual variability [9]. However, Leave-One-Subject-Out validation remained unstable, with R^2 values frequently negative. Several factors may account for this result: stress phases may have been too short to elicit measurable autonomic responses; baseline periods may not represent a true resting state; and HRV estimates derived from short, noisy IBI sequences are inherently unreliable. Moreover, psychological, contextual, or hormonal factors—unobserved by peripheral sensors—may play a substantial role in shaping subjective stress responses [10].

Regression Versus Classification

Reframing the task as a binary classification problem improved stability compared with continuous regression, suggesting that the direction of stress change is more detectable than its magnitude. However, classification performance remained close to chance, indicating that the available physiological signals provide only limited discriminative information about subjective stress fluctuations.

Implications and Future Directions

Overall, these findings highlight the limitations of using short wearable recordings to model fine-grained changes in subjective stress. Future work would benefit from longer and more stable recordings, the inclusion of additional physiological modalities such as respiration, and modelling approaches that explicitly exploit temporal dynamics rather than summary statistics alone. Improving the reliability of the target variable and ensuring more balanced experimental designs may further enhance model robustness.

5 Conclusion

This study examined whether physiological signals recorded by a wrist-worn wearable device could predict individual stress reactivity during a controlled Stress Protocol. Across multiple modelling approaches, predictive performance remained limited, indicating that the information captured by the Empatica E4—at the signal quality and time resolution available in this dataset—is in-

sufficient to reliably estimate changes in subjective stress. These findings highlight the inherent complexity of modelling stress from peripheral physiology and point to the

need for richer features, more robust ground-truth measures, and more controlled experimental conditions in future work.

Tables

| Signal | Acronym | Description | Unit |
|------------------------|---------|---|--------------------|
| Electrodermal Activity | EDA | Skin conductance reflecting sympathetic nervous system activity | μS |
| Blood Volume Pulse | BVP | Optical signal related to blood flow and cardiac cycles | a.u. |
| Accelerometer | ACC | Triaxial body movement and physical activity | g |
| Skin Temperature | TEMP | Peripheral skin temperature measurement | $^{\circ}\text{C}$ |
| Heart Rate | HR | Instantaneous heart beats per minute derived from BVP | bpm |
| Inter-Beat Interval | IBI | Time interval between consecutive heartbeats | ms |

Table 1: Physiological signals acquired from wearable sensors and their meanings.

| Metric | Acronym | Description | Unit |
|------------------------|---------|----------------------------|------|
| Root Mean Square of SD | RMSD | Short-term HR variability | ms |
| Percentage of NN50 | pNN50 | Adjacent IBI diff > 50 ms | % |
| Coeff. of Variation RR | CVRR | Normalised RR variability | – |
| Coeff. of Variation SD | CVSD | Normalised successive diff | – |
| Std. Deviation NN | SDNN | Overall HR variability | ms |

Table 2: Heart Rate Variability (HRV) metrics derived from IBI.

| Model | Type | Validation | Performance Summary |
|-------|-----------------------------|------------|--|
| 1 | RF, summary stats | Hold-out | $R^2 = 0.176$, RMSE = 1.637 |
| 2 | RF + HRV | Hold-out | $R^2 = 0.191$, RMSE = 1.623 |
| 3 | GBR + HRV | Hold-out | $R^2 = 0.121$, RMSE = 1.690 |
| 4 | GBR baseline-normalised | LOSO | Global $R^2 = -0.120$ |
| 5 | XGBoost baseline-normalised | LOSO | Acc = 0.514, BalAcc = 0.524, AUC = 0.593 |

Table 3: Summary of model performance across different feature sets and validation strategies.

References

- [1] Boucsein, W. (2012). *Electrodermal Activity*. Springer.
- [2] Picard, R. W., Fedor, S., & Ayzenberg, Y. (2016). Multiple arousal theory and applications for stress monitoring. *IEEE Signal Processing Magazine*.
- [3] Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The “Trier Social Stress Test” – A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1–2), 76–81.
- [4] Bai, Z., Wu, P., Geng, F., Zhang, H., Chen, X., Du, L., Wang, P., Li, X., Fang, Z., & Wu, Y. (2024). HSF-IBI: A universal framework for extracting inter-beat interval from heterogeneous unobtrusive sensors. *Bioengineering*, 11(12), 1219.
- [5] Lima, R., Osorio, D., & Gamboa, H. (2019). Heart Rate Variability and Electrodermal Activity in Mental Stress Aloud: Predicting the Outcome. LASIAC / Universidade Nova de Lisboa.
- [6] Kokate, S., Kovatte, M., Mhatre, N., & Deshpande, H. (2023). Comparative Analysis of Machine Learning Methodologies for HRV-based Stress Detection. Thadomal Shahani Engineering College.
- [7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- [8] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54, 1937–1967.
- [9] Fruet, D., Barà, C., Pernice, R., Iovino, M., Faes, L., & Nollo, G. (2025). A Signal Normalization Approach for Robust Driving Stress Assessment Using Multi-Domain Physiological Data. *Eng*, 6(11), 288.

- [10] Pingue, A., Gad, V., Jaisighani, D., Ghosh, S., & Sen, S. (2024). Detection and monitoring of stress using wearables: a systematic review. *Frontiers in Computer Science*, 6.
- [11] Quadrini, M., Falcone, D., & Gerard, G. (2024). Comparison of Machine Learning Approaches for Stress Detection from Wearable Sensors Data. *CEUR Workshop Proceedings*, Vol. 3762.
- [12] Hongn, A., Bosch, F., Prado, L. E., Ferrández, J. M., & Bonomini, M. P. (2025). Wearable Physiological Signals under Acute Stress and Exercise Conditions. *Scientific Data*.