UniSR
Università Vita-Salute San Raffaele

POLITECNICO MILANO 1863

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

Wearable Devices’ Course Project

MSc in Health Informatics

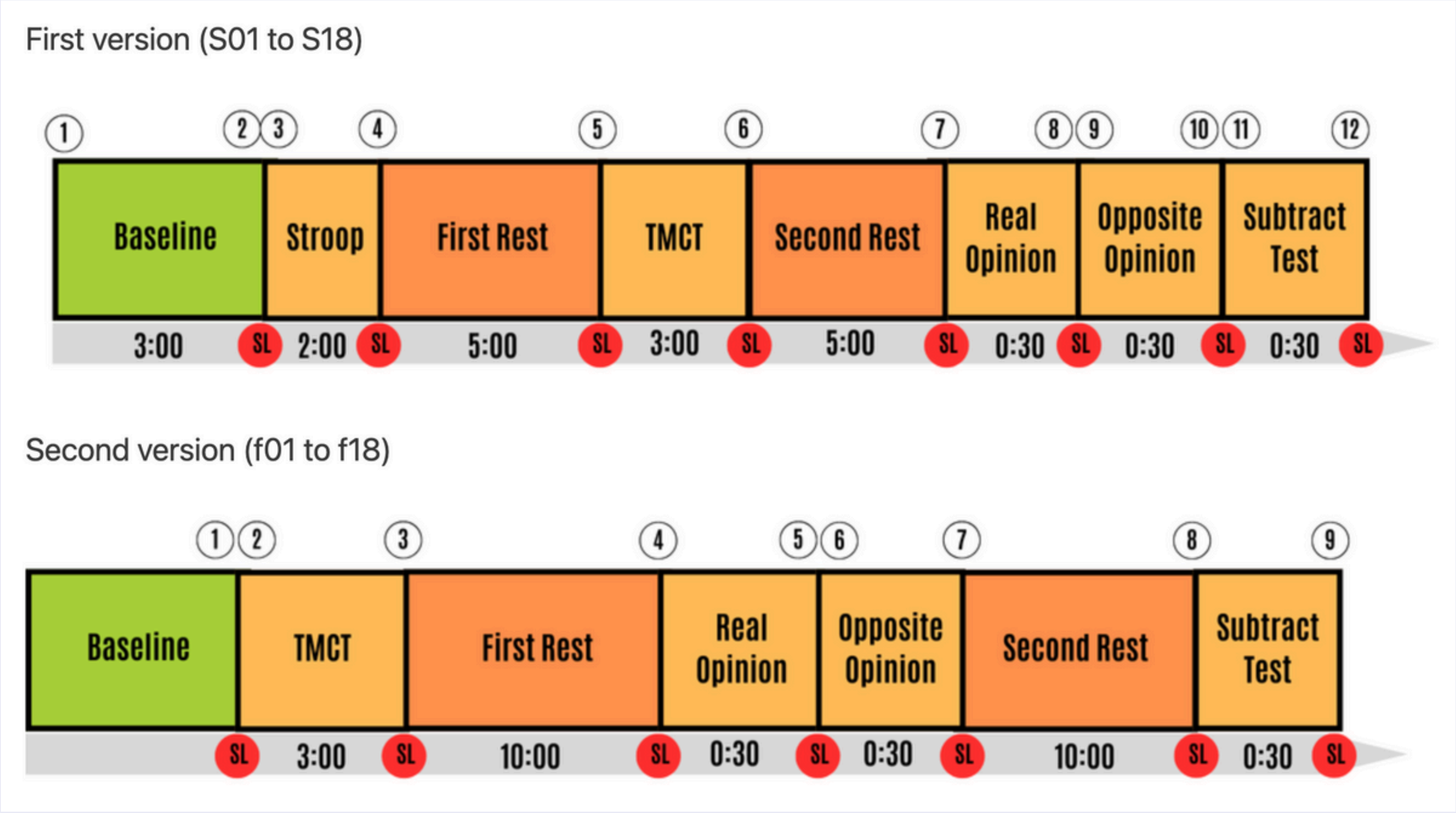
A.Y 2025-26

Caffi Giulia
Metallo Rebecca
Uberti Anna

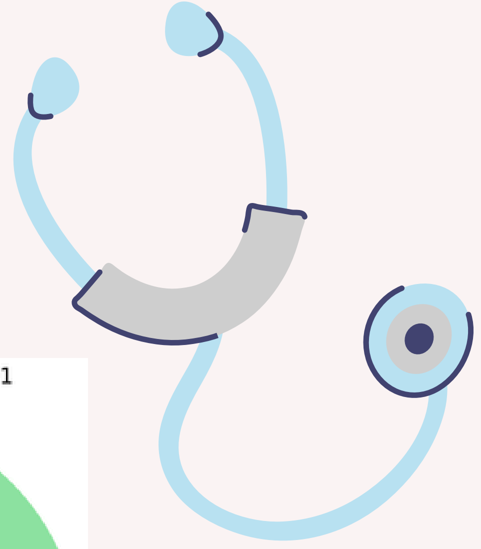
How we choose our idea?

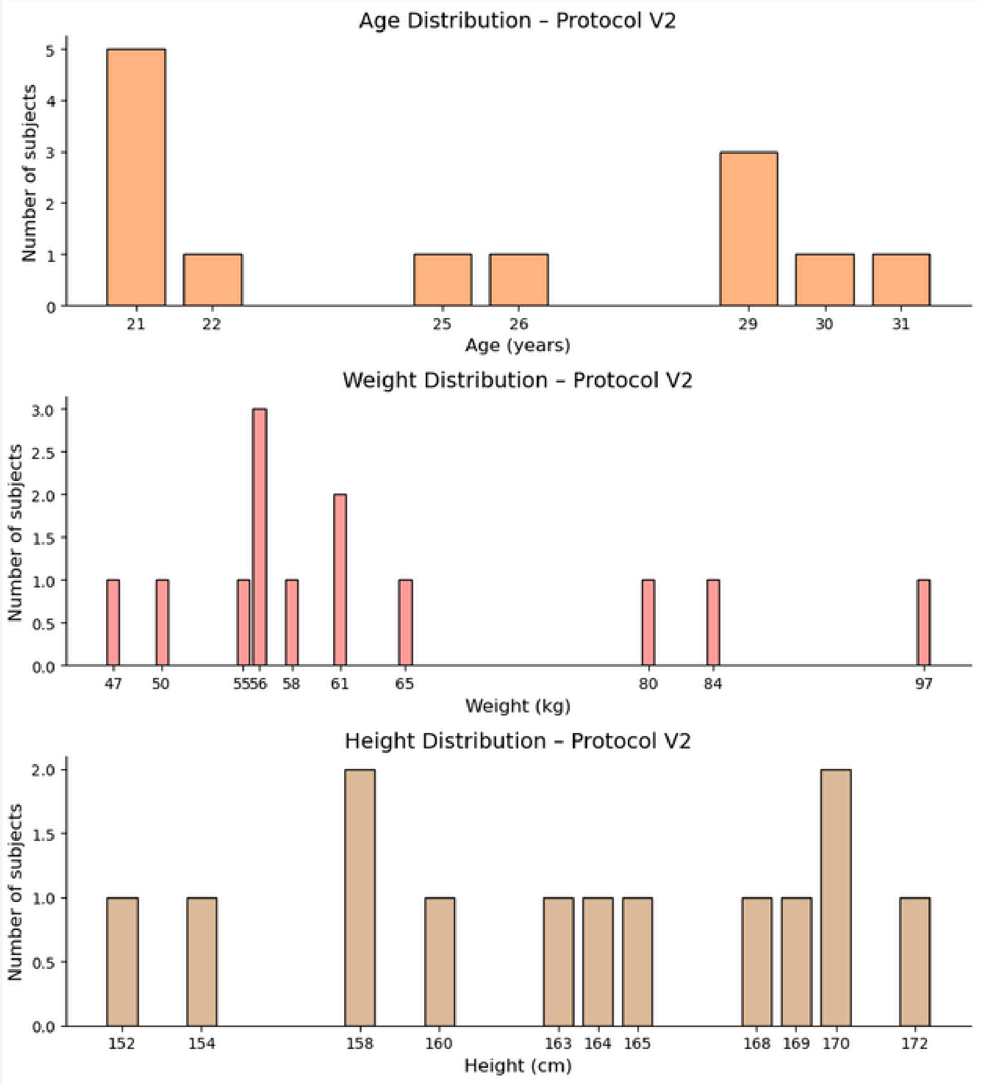
We were interested in understanding how much a person's physiological state changes when facing a stressful task during the day, **rather than limiting** the problem to a simple *stress vs. no-stress classification*.

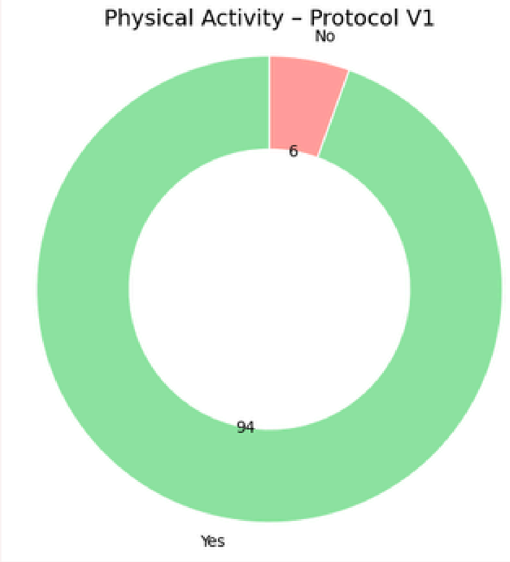
To ensure cleaner data and more reliable measurements, we focused specifically on the **STRESS protocol**, which provided higher-quality signals and a more structured experimental design for our analysis.

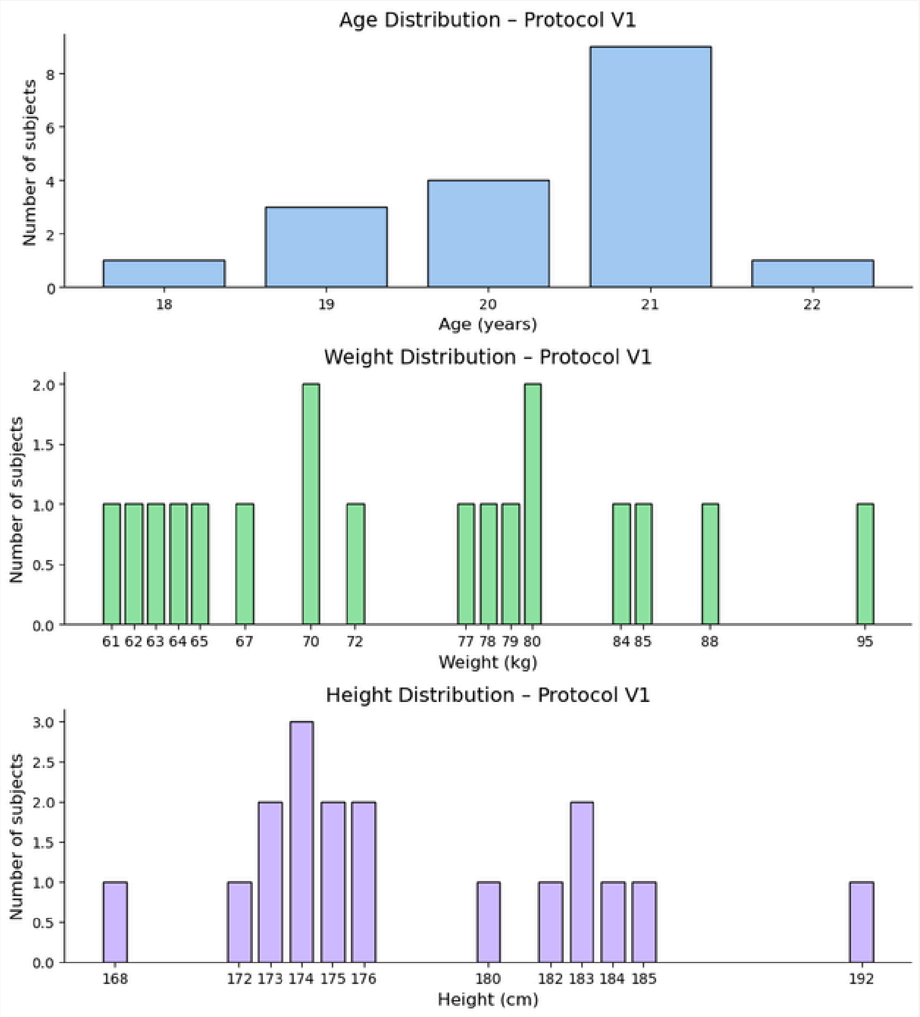


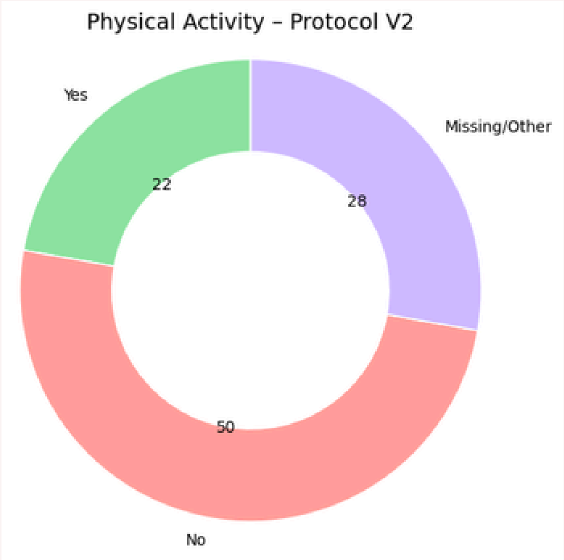
EDA : Exploratory Data Analysis











The exploratory analysis confirmed the patterns already observed during the initial inspection of the dataset.

Age range:

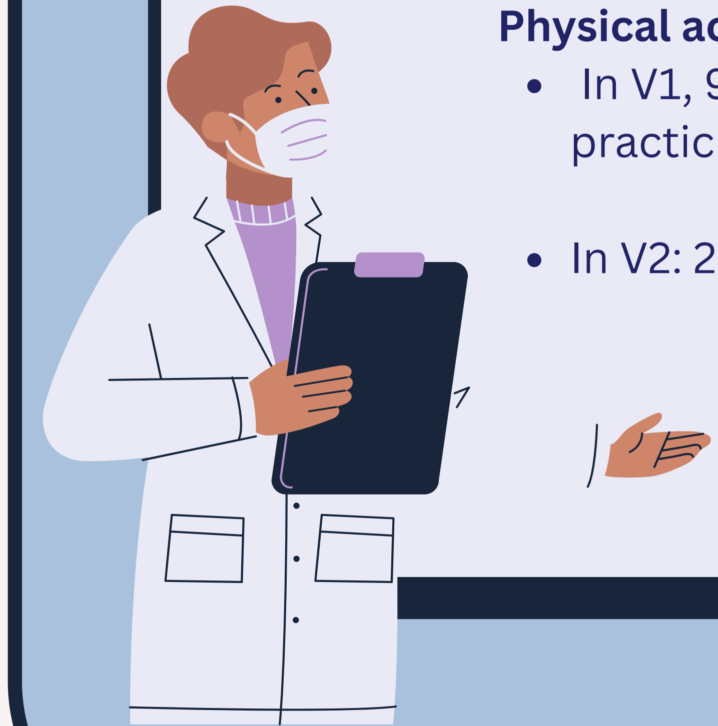
- Protocol V2: participants aged 18–31
- Protocol V1: participants aged 18–22

Sex distribution:

- V1 is composed entirely of male participants
- V2 includes mostly female participants

Anthropometric measures reflect this distribution:

- Weight: females ~ 47–65 kg (with a few outliers); males ~ 61–88 kg
- Height: females ~ 152–172 cm; males ~ 165–192 cm



Physical activity :

- In V1, 94% of participants report practicing regular physical activity
- In V2: 22% yes, 50% no, 28% unknown

Construction of Dataset

Step 1 — Signal Integration and Segmentation

We merged **STRESS protocol, stress_level_v1, stress_level_v2, and subject_info**.
All physiological signals were carefully segmented to include only the effective phases of each protocol, removing unrelated portions of the recordings.

Step 2 — Data Quality Filtering

Participants showing **problematic measurements** or more than **15% missing data** were excluded.
Removed subjects included: “S02”, “F07”, “f14_a”, “f14_b”, “f14”, “f15”, “f16”, “f17”, “f18”.

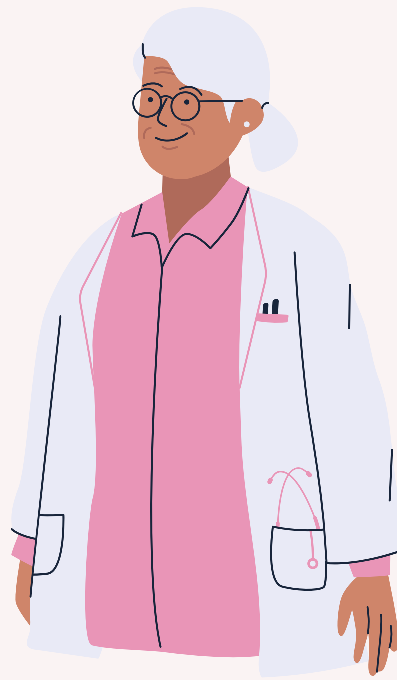
Step 3 — Feature Extraction per Phase

For each participant and each physiological signal (TEMP, EDA, HR, ACC, BVP), we computed the **mean** and **standard deviation** for every phase of the protocol.
V1 and V2 were treated separately to respect the structure of each protocol.

Step 4 — Target Construction (Δ-Stress)

We defined our prediction target using stress_level values.
The variable **delta_stress** was created as:
Δ-stress = stress_level − baseline_stress.

1st Model : Random Forest Regressor



mRMR to discard correlated and not informative features

1. "EDA_std",
2. "HR_std",
3. "TEMP_mean",
4. "ACC_norm_mean",
5. "Height (cm)",
6. "Weight (kg)",
7. "Physical_Activity"

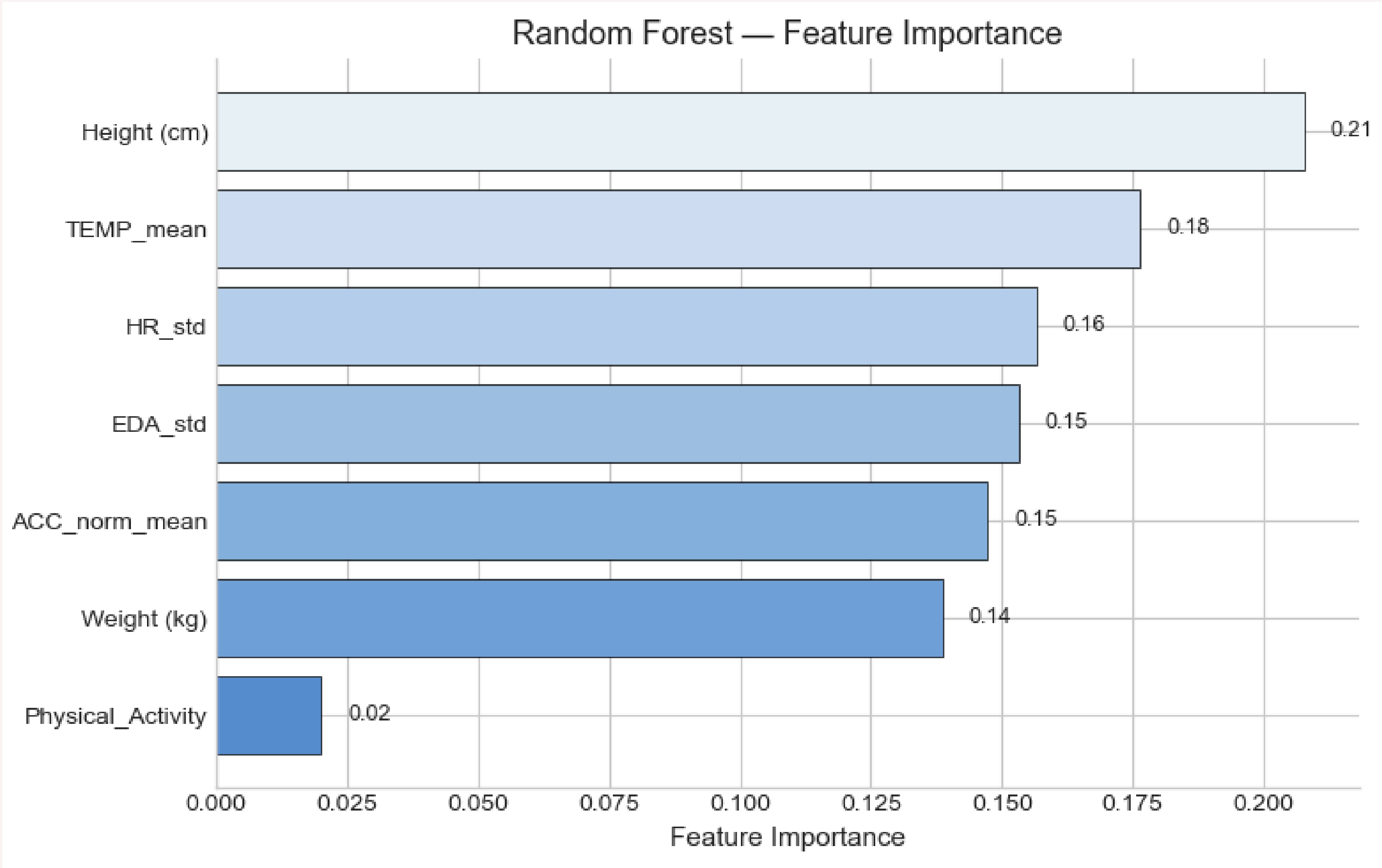
y = Δ-stress

RandomizedSearchCV (20 iterations, 3-fold CV)

Tuned Hyperparameters : n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features

Best params: {'n_estimators': 800, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': None}

Test R^2: 0.176
Test RMSE: 1.637
Test MAE: 1.103



Our model explains only the 18% of variability in the delta_stress

2nd Model : Random Forest Regressor + IBI signals

*“Utilizing these unobtrusive technologies allows for the measurement of inter-beat intervals (IBIs), which, in turn, allows for the calculation of heart rate variability (HRV). Compared to average heart rate (HR), **HRV provides more detailed insights** into cardiac and neurological functions “[4]*

This highlighted the importance of incorporating **HRV-related metrics** in our model.
Therefore, we took into account IBI values and extracted HRV features for each protocol phase:

- **IBI_mean** – average inter-beat interval
- **IBI_std** – variability of IBI
- **RMSSD** – root mean square of successive differences
- **pNN50** – proportion of IBI differences > 50 ms

mRMR to discard correlated and not informative features

"EDA_std",
"HR_std",
"TEMP_mean",
"ACC_norm_mean",
"Height (cm)",
"IBI_mean",
"RMSSD",
"Weight (kg)",
"Physical_Activity"

y = Δ-stress

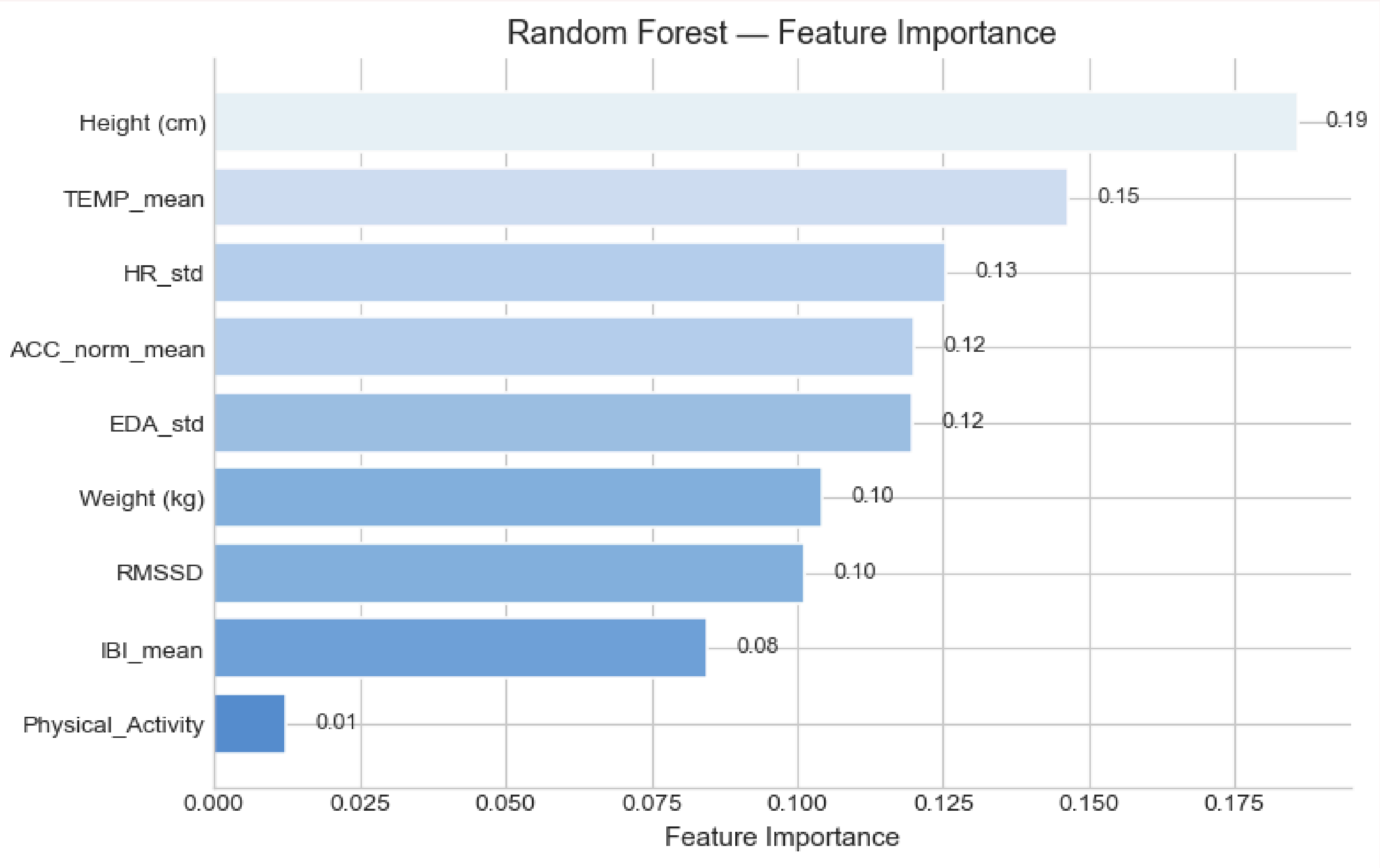
2nd Model : Random Forest Regressor + IBI signals

RandomizedSearchCV (20 iterations, 3-fold CV)

Tuned Hyperparameters : n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features

Best params: {'n_estimators': 800, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': None}

Test R^2: 0.197
Test RMSE: 1.623 Test MAE: 1.079



Our model now explains the 20% of variability in the delta_stress

3rd Model : Gradient Boosting Regressor

mRMR to discard correlated and not informative features

"EDA_std",
"HR_std",
"TEMP_mean",
"ACC_norm_mean",
"Height (cm)",
"IBI_mean",
"RMSSD",
"Weight (kg)",
"Physical_Activity"

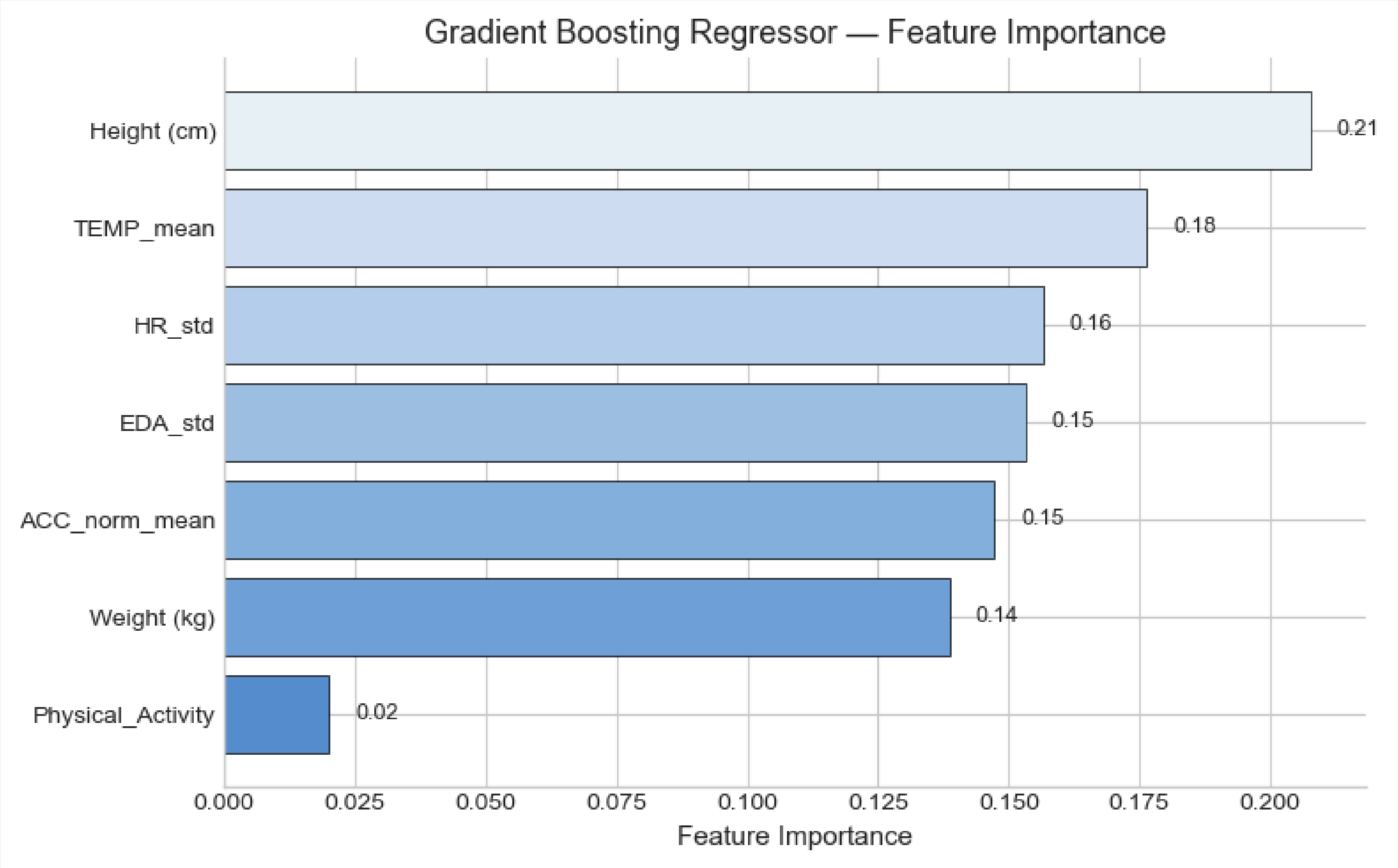
y = Δ-stress

RandomizedSearchCV (20 iterations, 3-fold CV)

Tuned Hyperparameters : n_estimators, learning_rate, max_depth, min_samples_leaf, min_samples_split, max_features, subsample

Best params: {
'subsample': 1.0,
'n_estimators': 200,
'min_samples_split': 6,
'min_samples_leaf': 3,
'max_features': 'log2',
'max_depth': 5,
'learning_rate': 0.01}

Test R^2: 0.121
Test RMSE: 1.690
Test MAE: 1.125



Our model explains the 12% of variability in the delta_stress

4th Model : Gradient Boosting Regression over a New Dataset

Given the limited performance of previous models, the feature engineering strategy was revised.

What’s new?

In protocols, **Baseline** is a dedicated stage used
to capture each subject’s resting physiological state.

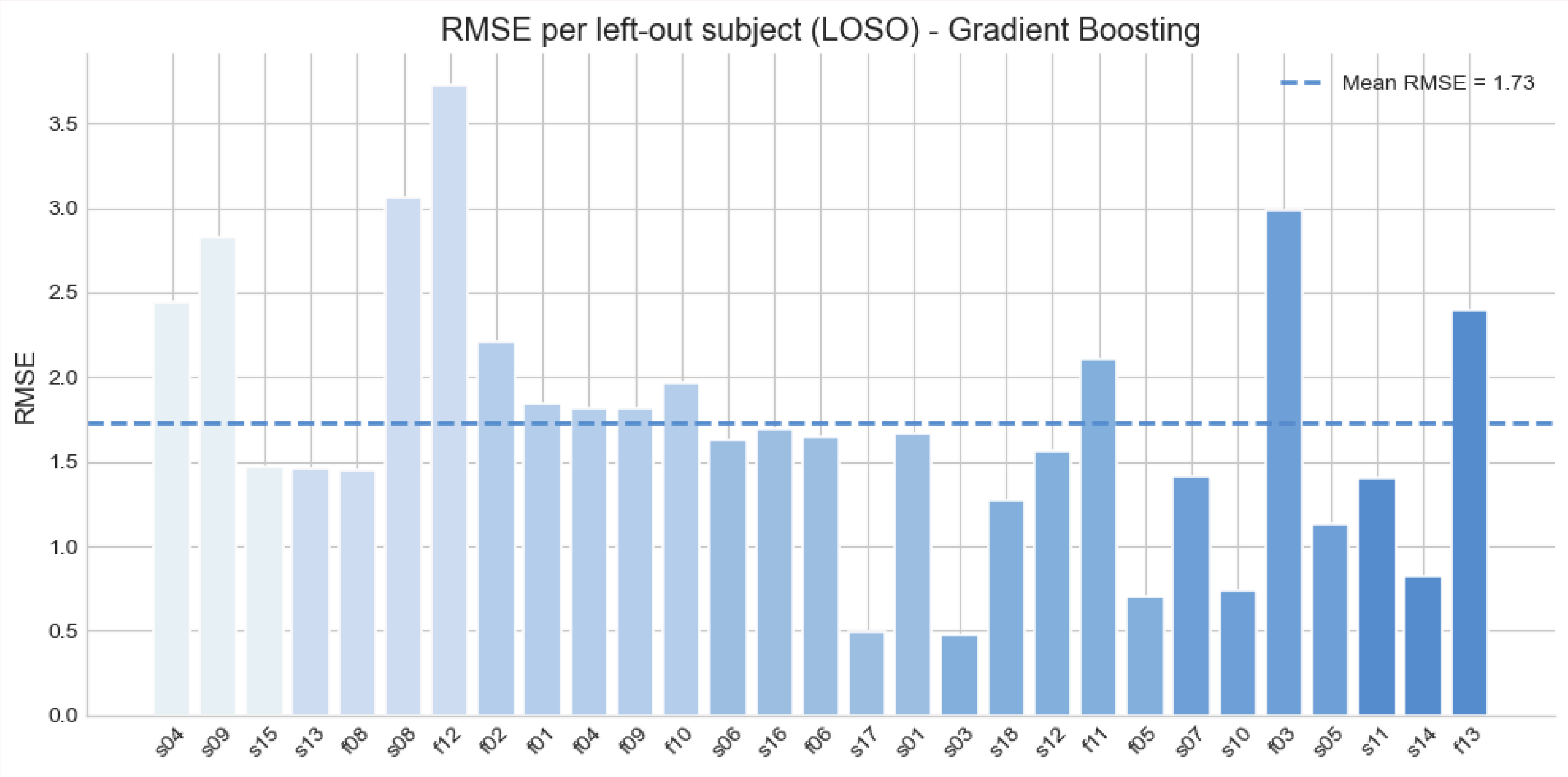
For each phase of the stress protocol, we computed **delta-features**,
defined as the **difference** between the **phase-value** and the **baseline value**

Two additional HRV indices were derived:

- **CVRR**: ratio between RMSSD and mean IBI, capturing relative beat-to-beat variability.
- **CVSD**: ratio between RMSSD and IBI standard deviation, reflecting variability dispersion.

Model performance was evaluated using **Leave-One-Subject-Out** (LOSO) cross-validation
to ensure subject-independent assessment.

4th Model : Gradient Boosting Regression over a New Dataset



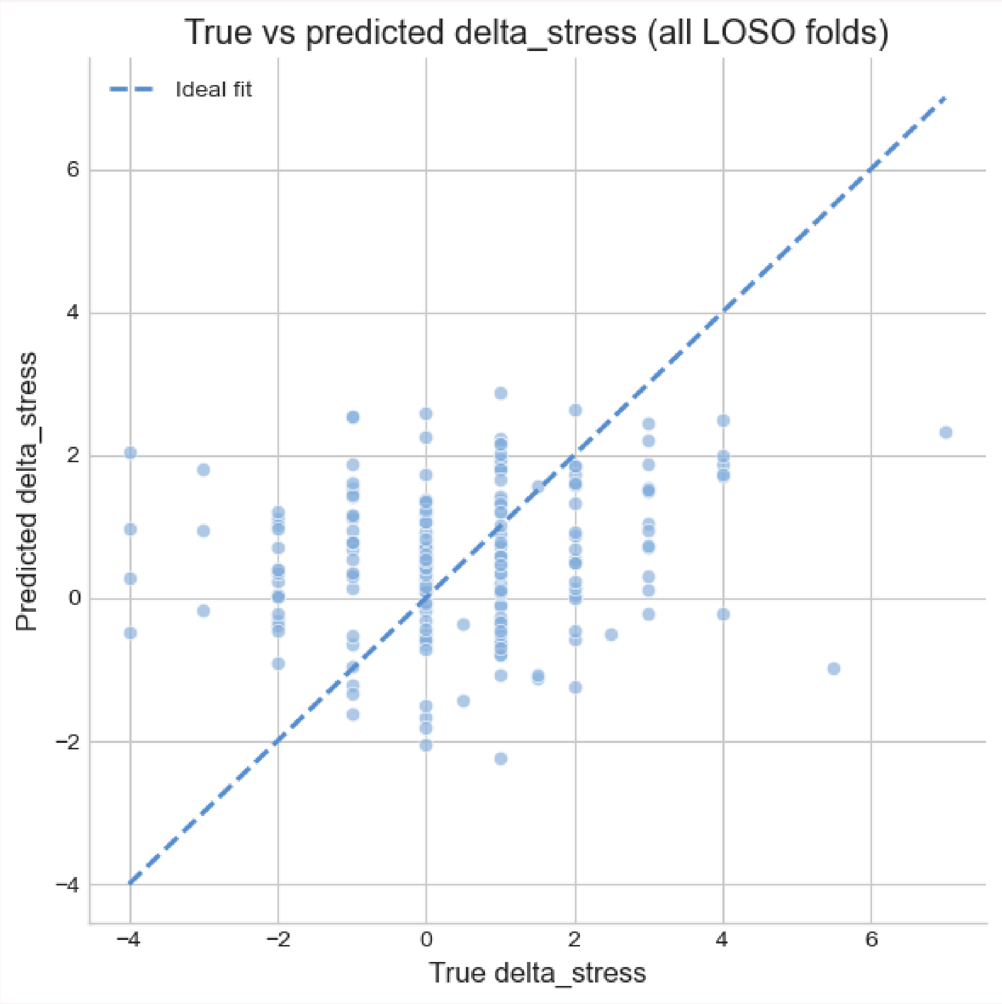
"delta_EDA_std", "delta_HR_std", "delta_TEMP_mean", "delta_ACC_norm_mean", "delta_IBI_mean", "delta_RMSSD", "CVRR", "CVSD", "Height (cm)", "Weight (kg)", "Physical_Activity"

y = Δ-stress

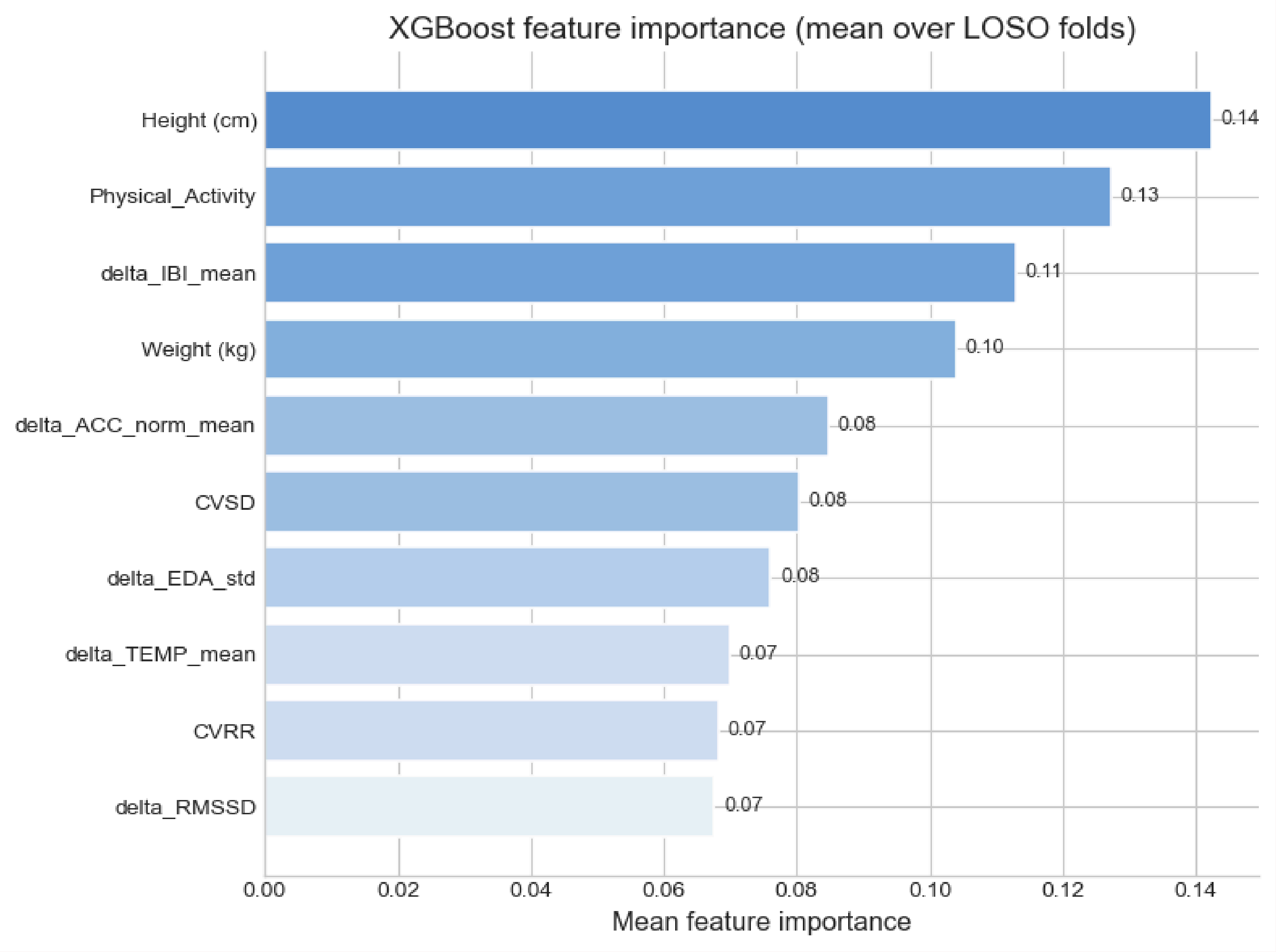
RandomizedSearchCV
(20 iterations, 3-fold CV inside each LOSO training fold)

Tuned parameters:
n_estimators, learning_rate, max_depth, min_samples_split, min_samples_leaf, subsample, max_features

LOSO performance:
Mean R^2 = −1.65 ± 1.64
Mean RMSE = 1.72 ± 0.72
Mean MAE = 1.47 ± 0.64



5th Model : XGBoost Classification



Previous regression approaches did not achieve satisfactory predictive performance.

Therefore, the task was reframed as a **binary classification** problem, where the goal becomes predicting whether stress increases relative to each subject's baseline.
A binary outcome is defined as:

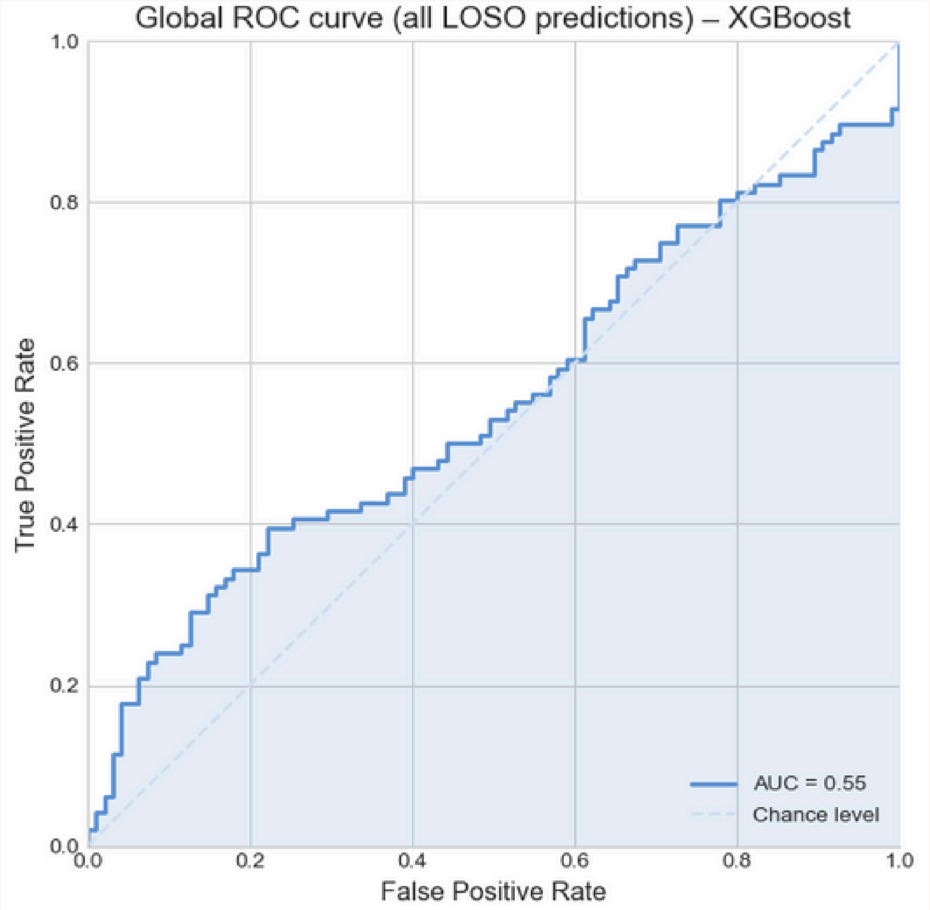
$$y = \begin{cases} 1 & \text{if } \Delta\text{stress} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Key results (LOSO)

Mean Accuracy: 0.519
Std Accuracy: 0.294
Mean Balanced Accuracy: 0.542
Std Balanced Accuracy: 0.300
Mean AUC: 0.593

Same dataset as in Model 4.

LOSO cross-validation .



Critical Considerations

Physiological effects are subtle
Although EDA, HRV, and skin temperature behave as expected under stress, their variations are often small compared to background variability, especially over short time windows.

Spurious importance of anthropometric variables
Height and weight appear highly predictive, likely because they capture dataset structure rather than true stress mechanisms. Baseline normalization alone was insufficient to ensure generalization.

Subjective and noisy target variable
Stress labels are based on self-reports, which are subjective and discretized, adding uncertainty to the learning process.

Limited sample size
The dataset includes an extremely small number of observations, which reduces drastically statistical robustness .

Conclusions and Future Directions

With the available dataset and wrist-based signals, reliable prediction of individual stress is not achievable.

This reflects intrinsic limitations of wearable data and stress labeling rather than a failure of machine learning.

Future work should focus on more robust signals, longer and temporal features, and improved ground truth definition.

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. 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] Pinge, 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.

Thank you for your attention!