

Investigating the Effect of ECG Recording Duration on HRV Reliability and Stress Classification

1st Ahmed H. Aly

University of Pisa

Pisa, Italy

a.alys2@studenti.unipi.it

2nd Vittorio Meini

National Research Council of Italy

Pisa, Italy

vittoriomeini@cnr.it

3rd Lucia Billeci

National Research Council of Italy

Pisa, Italy

lucia.billeci@cnr.it

Abstract—Continuous stress monitoring using wearable devices has gained increasing attention because of its potential for everyday health assessment. Electrocardiogram (ECG)-based stress detection commonly relies on heart rate variability (HRV) analysis, which traditionally requires relatively long ECG recordings. However, wearable devices often acquire shorter recordings, making it important to determine whether they provide reliable HRV features while maintaining accurate stress classification.

This study investigates the effect of ECG recording duration on the reliability of time-domain HRV features and HRV-based stress classification using the WESAD dataset. ECG signals were segmented into 30 s, 120 s, and 300 s windows, and the extracted HRV features were evaluated using Pearson's correlation coefficient, the Intraclass Correlation Coefficient (ICC), and the Mean Absolute Error (MAE). Seven machine learning classifiers were then evaluated under binary and three-class stress classification settings.

The results show that reducing the recording duration generally decreases HRV feature agreement. Nevertheless, Mean HR and Mean RR remained relatively stable, whereas SDNN was the most sensitive to shorter recordings. Despite the reduced feature agreement, HRV features extracted from 30 s ECG recordings achieved classification performance comparable to that obtained using 120 s and 300 s recordings, and in the binary classification task higher than that of longer recordings. These findings demonstrate that short ECG recordings provide sufficient time-domain HRV information for reliable stress monitoring.

Index Terms—Stress detection, Electrocardiogram, Heart rate variability, Wearable healthcare, Machine learning, WESAD

I. INTRODUCTION

Stress is a common physiological and psychological response to challenging situations. Although short-term stress is a normal adaptive mechanism, prolonged or chronic stress can negatively affect both physical and mental health. It has been associated with an increased risk of cardiovascular diseases, anxiety, depression, and a reduced quality of life [1], [2]. These adverse effects have motivated increasing interest in developing reliable methods for continuously monitoring stress during daily life.

Recent advances in wearable technology have made stress monitoring increasingly feasible in daily life. Smartwatches and portable ECG sensors can now acquire physiological signals with minimal disruption to users' everyday activities. Previous studies have demonstrated the potential of wearable devices for stress detection in real-world environments while highlighting the challenges of obtaining reliable physiological measurements outside controlled laboratory settings [3].

Among the physiological signals used for stress assessment, the electrocardiogram (ECG) has attracted considerable attention because it provides a non-invasive measure of cardiac activity. Stress alters the balance of the autonomic nervous system (ANS), typically increasing sympathetic activity while reducing parasympathetic activity. These physiological changes affect heart rate (HR) and heart rate variability (HRV), making HRV one of the most widely used indicators of autonomic regulation and psychological stress [4], [5]. Consequently, many ECG-based stress detection methods rely on HRV features combined with machine learning algorithms to distinguish between different stress conditions. Although recent studies have demonstrated that the complete ECG waveform can also be used for stress recognition [6], [7], HRV-based approaches remain the most widely adopted because of their physiological interpretability and computational efficiency.

Despite these advances, ECG recording duration remains an important practical challenge for wearable stress monitoring. Conventional HRV analysis recommends using approximately five minutes of ECG data to obtain reliable measurements [5]. While this recommendation is appropriate in clinical and laboratory environments, wearable devices often acquire much shorter recordings because of practical limitations. Whether these shorter recordings preserve sufficient information for reliable HRV analysis and stress classification remains an open question.

Previous studies have primarily investigated the validity of HRV features extracted from short or ultra-short ECG recordings by comparing them with conventional five-minute recordings. Their findings indicate that several time-domain HRV features remain relatively consistent across different recording durations, whereas frequency-domain measures become less reliable as the recording length decreases [8]–[11]. However, considerably less attention has been given to understanding how shorter ECG recordings influence the performance of machine learning models for stress classification. More importantly, it remains unclear whether a reduction in the statistical reliability of HRV features necessarily leads to lower stress classification performance [12]. Addressing this question is essential for the development of practical wearable stress monitoring systems.

To address this gap, this paper investigates how ECG recording duration influences both the reliability of time-

domain HRV features and the performance of HRV-based stress classification. Using the WESAD dataset [13], short ECG recordings are compared with conventional recordings to determine whether reductions in HRV feature reliability translate into lower classification accuracy. By jointly analyzing HRV feature reliability and stress classification performance, this work provides a comprehensive assessment of the suitability of short ECG recordings for wearable stress monitoring applications.

The main contributions of this work are summarized as follows:

- A systematic evaluation of the effect of ECG recording duration on the reliability of time-domain HRV features using multiple statistical agreement measures.
- An investigation of the relationship between HRV feature reliability and stress classification performance to determine whether improvements in feature agreement necessarily translate into higher classification accuracy.
- A comprehensive evaluation of seven machine learning classifiers under both binary and three-class stress classification settings using ECG recordings of different durations.
- Practical insights into the feasibility of short ECG recordings for wearable ECG-based stress monitoring, highlighting the trade-off between acquisition time and stress detection performance.

II. MATERIALS AND METHODS

A. Dataset

The experiments were conducted using the publicly available Wearable Stress and Affect Detection (WESAD) dataset [13], which was developed for multimodal stress and affect recognition. The dataset contains physiological and motion signals collected from 15 healthy participants (12 males and 3 females) with a mean age of 27.5 ± 2.4 years. During the experimental protocol, each participant underwent four affective conditions: baseline, stress, amusement, and meditation.

Although WESAD provides multiple physiological modalities, this study focuses exclusively on the electrocardiogram (ECG) signal acquired using the RespiBAN chest device. The ECG was recorded at a sampling frequency of 700 Hz, providing sufficient temporal resolution for accurate R-peak detection and subsequent heart rate variability (HRV) analysis.

B. ECG Preprocessing

Prior to feature extraction, the ECG recordings were preprocessed to improve signal quality and ensure reliable HRV estimation. The preprocessing pipeline consisted of signal cleaning, window segmentation, and RR interval extraction.

The raw ECG signals were first processed using the NeuroKit2 [14] ECG cleaning pipeline. A bandpass filter with a passband of approximately 0.5–45 Hz was applied to suppress baseline wander, motion artifacts, and high-frequency noise while preserving the QRS complexes required for accurate R-peak detection.

The filtered ECG recordings were segmented into non-overlapping windows of 30 s, 120 s, and 300 s. Since the physiological state can change over time, the labels assigned to each window were carefully examined. A window was retained only if at least 90% of its samples shared the same label; otherwise, it was discarded to avoid introducing ambiguously labeled samples into the analysis.

Finally, R-peaks were detected from each retained ECG segment using the NeuroKit2 ECG processing pipeline. The RR intervals were then computed as the time differences between successive R-peaks and used to derive the HRV features. The extracted time-domain HRV features comprised the mean RR interval (Mean RR), mean heart rate (Mean HR), the standard deviation of normal-to-normal intervals (SDNN), the root mean square of successive differences (RMSSD), and the percentage of successive RR intervals differing by more than 50 ms (pNN50). These features were computed according to the standard definitions implemented in NeuroKit2 and were subsequently used in both the statistical analysis and the machine learning models.

C. Label Definition

Two labeling strategies were considered to evaluate the effect of ECG recording duration under different classification scenarios.

The first strategy defines a binary classification problem in which the stress recordings form the positive class, while the baseline, amusement, and meditation recordings are combined into a single non-stress class. This setting reflects the practical objective of distinguishing stressful from non-stressful conditions.

The second strategy formulates a three-class classification problem consisting of stress, non-stress, and meditation. In this case, the non-stress class includes only the baseline and amusement recordings, whereas meditation is treated as a separate class because of its distinct physiological characteristics and increased parasympathetic activity. Evaluating both labeling strategies enables the influence of recording duration to be examined under different levels of classification complexity.

D. Statistical Analysis

The statistical analysis was performed to evaluate how recording duration influences the reliability of time-domain HRV features. Features extracted from the 30 s and 120 s recordings were compared with those obtained from the 300 s recordings, which served as the reference because five-minute ECG recordings are conventionally recommended for HRV analysis [4], [5].

Feature agreement was quantified using Pearson's correlation coefficient (r), the Intraclass Correlation Coefficient (ICC), and the Mean Absolute Error (MAE). Pearson's correlation coefficient measures the linear association between feature values, ICC evaluates their consistency, and MAE quantifies the average absolute difference between corresponding measurements.

To account for physiological differences across affective states, the analysis was conducted separately for each class. This approach allows the influence of recording duration to be evaluated independently of the underlying physiological condition.

E. Machine Learning Analysis

The extracted time-domain HRV features were subsequently used to investigate the effect of recording duration on stress classification under both binary and three-class labeling strategies. Seven supervised machine learning classifiers were evaluated: Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost).

Model performance was assessed using Leave-One-Subject-Out (LOSO) cross-validation. In each iteration, data from one participant were reserved for testing, while the remaining participants were used for training. This subject-independent evaluation protocol provides a realistic estimate of model generalization and reduces the risk of overfitting to individual participants. Unless otherwise specified, all classifiers were trained using their default hyperparameter settings.

Performance was evaluated using Accuracy, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). The reported results correspond to the average performance across all LOSO folds and were used to examine the effect of ECG recording duration on stress classification.

III. RESULTS

A. Feature Correlation Analysis

Before evaluating feature reliability and classification performance, an exploratory correlation analysis was conducted to examine the relationships among the extracted time-domain HRV features. Understanding these relationships helps identify potentially redundant features while highlighting those that may provide complementary physiological information.

Figure 1 presents the Pearson correlation matrix of the extracted HRV features. Several feature pairs exhibit moderate to strong correlations, reflecting the fact that they characterize related aspects of cardiac autonomic regulation. As expected, Mean RR and Mean HR show a strong inverse relationship because heart rate is directly derived from the RR interval. Similarly, RMSSD and pNN50 are strongly positively correlated, indicating that both features capture short-term beat-to-beat variability associated with parasympathetic activity.

In contrast, SDNN shows weaker correlations with several of the remaining features, suggesting that it captures complementary information related to overall heart rate variability rather than short-term fluctuations alone. Overall, the correlation matrix indicates that the selected feature set contains both correlated and complementary information, supporting its use in the subsequent statistical and machine learning analyses.

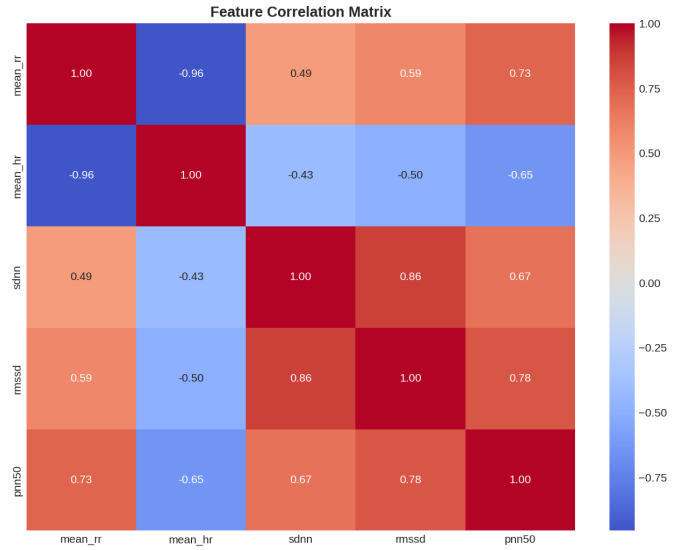


Fig. 1. Pearson correlation matrix of the extracted time-domain HRV features from 30 s ECG recordings.

B. Statistical Agreement Analysis

To evaluate the effect of ECG recording duration on HRV feature reliability, features extracted from shorter recordings were compared with those obtained from the conventional 300 s recordings. Agreement was assessed using Pearson's correlation coefficient (r), the Intraclass Correlation Coefficient (ICC), and the Mean Absolute Error (MAE). The results for the meditation, non-stress, and stress conditions are presented in Tables I –II–III.

For the meditation condition (Table II), RMSSD exhibited the strongest agreement across recording durations, achieving the highest Pearson correlation and ICC values for both the 30 s versus 120 s and 30 s versus 300 s comparisons. Mean RR and pNN50 also demonstrated relatively high agreement, whereas SDNN showed noticeably lower correlation and ICC values together with larger MAE values. Although Mean HR produced lower correlation coefficients than RMSSD, it consistently achieved the smallest MAE, indicating that absolute heart rate estimates remained relatively stable even when shorter recordings were used.

Table III summarizes the agreement between HRV features extracted from ECG recordings of different durations for the non-stress condition. Overall, this condition exhibited the highest agreement among the three physiological states. Mean RR showed the strongest consistency, with Pearson correlation coefficients of 0.920 and 0.915 and ICC values of 0.918 and 0.911 for the 30 s versus 120 s and 30 s versus 300 s comparisons, respectively. pNN50 also demonstrated excellent agreement across both comparisons. Although Mean HR produced slightly lower correlation coefficients, it consistently yielded the lowest MAE values (2.944 and 3.436), indicating stable heart rate estimation across different recording durations. In contrast, SDNN exhibited the lowest Pearson correlation and ICC values together with the largest MAE, confirming that it

TABLE I
AGREEMENT OF HRV FEATURES ACROSS ECG RECORDING DURATIONS FOR THE STRESS CONDITION.

Feature	30 s vs 120 s			30 s vs 300 s		
	<i>r</i>	ICC	MAE	<i>r</i>	ICC	MAE
Mean HR	0.924	0.922	5.295	0.901	0.897	6.357
Mean RR	0.902	0.899	37.219	0.877	0.868	44.666
pNN50	0.805	0.801	5.912	0.747	0.730	6.639
RMSSD	0.606	0.598	19.941	0.673	0.635	22.465
SDNN	0.641	0.581	24.396	0.551	0.397	37.962

TABLE II
AGREEMENT OF HRV FEATURES ACROSS ECG RECORDING DURATIONS FOR THE MEDITATION CONDITION.

Feature	30 s vs 120 s			30 s vs 300 s		
	<i>r</i>	ICC	MAE	<i>r</i>	ICC	MAE
Mean HR	0.685	0.655	4.300	0.522	0.427	8.400
Mean RR	0.815	0.804	44.067	0.699	0.635	80.300
pNN50	0.812	0.807	8.086	0.713	0.677	11.025
RMSSD	0.872	0.868	14.953	0.807	0.803	18.465
SDNN	0.725	0.711	21.171	0.542	0.479	33.873

is more sensitive to reduced recording duration.

The results for the stress condition are presented in Table I. Mean HR demonstrated the strongest agreement across recording durations, with Pearson correlation coefficients of 0.924 and 0.901, ICC values of 0.922 and 0.897, and the lowest MAE values (5.295 and 6.357). Mean RR also remained highly consistent, although its agreement was slightly lower than that of Mean HR. Compared with the non-stress condition, pNN50 and RMSSD showed a more noticeable reduction in agreement, while SDNN again exhibited the weakest agreement, with the lowest ICC values and the largest MAE. These findings indicate that HRV features are generally less consistent under stress than under non-stress conditions, particularly for variability-based measures.

Across all three affective conditions, the agreement between HRV features generally decreased as the difference in recording duration increased. Nevertheless, the magnitude of this reduction varied across features. Mean HR and Mean RR remained relatively robust regardless of recording length, whereas SDNN consistently exhibited the greatest sensitivity to shorter ECG recordings.

C. Machine Learning Analysis

The influence of ECG recording duration on stress classification was evaluated using seven supervised machine learning algorithms under both binary and three-class labeling strategies. Figures 2–4 compare the performance of all evaluated classifiers, while Table IV summarizes the best-performing model for each recording duration.

For the binary classification task, most classifiers achieved consistently high performance across all recording durations. As shown in Fig. 2, Gradient Boosting obtained the highest Accuracy (88.4%) and also achieved the best F1-score and AUC using the 30 s recordings. Similar performance was observed for the 120 s recordings (Fig. 3), where Gradient Boosting again outperformed the remaining classifiers with an Accuracy of 87.3%. For the 300 s recordings (Fig. 4), KNN achieved the highest Accuracy (81.3%), although the overall

differences among the evaluated models remained relatively small.

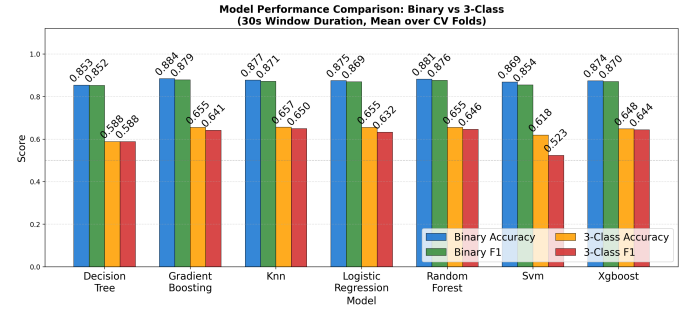


Fig. 2. Performance of the evaluated machine learning models using HRV features extracted from 30-second ECG recordings.

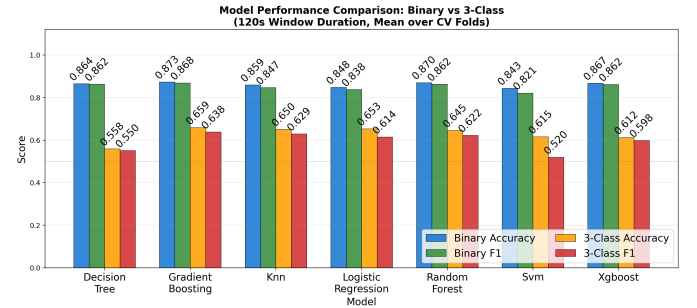


Fig. 3. Performance of the evaluated machine learning models using HRV features extracted from 120-second ECG recordings.

Compared with the binary setting, the three-class classification task resulted in lower Accuracy and F1-score across all recording durations because of the increased difficulty of simultaneously distinguishing stress, non-stress, and meditation. KNN achieved the highest Accuracy (65.7%) using the 30 s recordings, Gradient Boosting performed best for the 120 s recordings (65.9%), and Logistic Regression obtained the highest Accuracy (67.4%) for the 300 s recordings.

TABLE III
AGREEMENT OF HRV FEATURES ACROSS ECG RECORDING DURATIONS FOR THE NON-STRESS CONDITION.

Feature	30 s vs 120 s			30 s vs 300 s		
	<i>r</i>	ICC	MAE	<i>r</i>	ICC	MAE
Mean HR	0.880	0.878	2.944	0.871	0.864	3.436
Mean RR	0.920	0.918	30.556	0.915	0.911	35.272
pNN50	0.915	0.911	6.212	0.898	0.889	7.088
RMSSD	0.847	0.840	11.947	0.828	0.817	13.459
SDNN	0.640	0.597	20.528	0.451	0.451	27.021

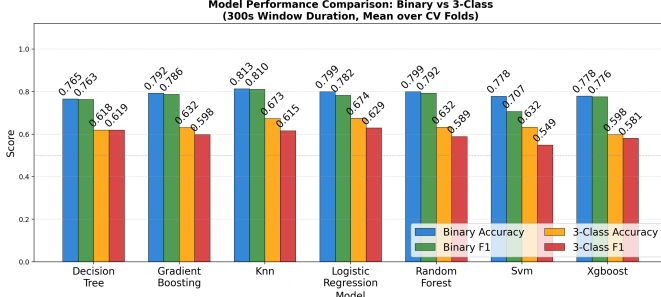


Fig. 4. Performance of the evaluated machine learning models using HRV features extracted from 300-second ECG recordings.

Table IV summarizes the best-performing classifier for each recording duration and classification task. An important observation is that the shortest ECG recordings produced classification performance comparable to, and in the binary classification task even better than, the longer recordings. Although the statistical agreement analysis showed that several HRV features became less consistent as recording duration decreased, this reduction did not translate into lower classification performance.

IV. DISCUSSION

This study examined the influence of ECG recording duration on both the reliability of time-domain HRV features and the performance of HRV-based stress classification. The statistical analysis showed that shortening the ECG recording generally reduced the agreement between HRV features extracted from short and conventional recordings. However, the degree of sensitivity varied across features. Mean HR and Mean RR remained highly consistent across all affective conditions, whereas SDNN was consistently the most affected by shorter recording durations. These findings are consistent with previous studies reporting that some time-domain HRV features remain reliable even when derived from short ECG recordings, while others require longer recordings to provide stable estimates [8]–[11].

Despite the reduction in feature agreement, the machine learning experiments demonstrated that shorter recordings did not lead to poorer stress classification performance. In the binary classification task, the highest performance was achieved using 30 s ECG recordings, while the three-class classification results remained comparable across all evaluated recording durations. These findings suggest that improvements

in the statistical reliability of HRV features do not necessarily translate into higher classification accuracy. Instead, the information required for stress discrimination appears to be preserved even in relatively short ECG segments.

One possible explanation is that short ECG recordings capture the transient cardiovascular responses associated with acute stress while reducing the influence of slower physiological variations that may not contribute to stress discrimination. In addition, segmenting the recordings into shorter windows increases the number of available training samples, which may improve the ability of machine learning models to learn discriminative patterns. Although this study was not designed to isolate the contribution of each factor, these observations indicate that classification performance depends on more than the statistical stability of individual HRV features.

The results also showed that binary classification consistently outperformed the three-class classification task regardless of the recording duration or classifier. This outcome is expected because distinguishing stress from non-stress is inherently less complex than simultaneously separating stress, non-stress, and meditation. Meditation is characterized by enhanced parasympathetic activity and shares physiological characteristics with relaxed states, increasing the overlap between classes and making the classification task more challenging.

From a practical perspective, these findings support the feasibility of using short ECG recordings in wearable stress monitoring systems. Shorter recordings reduce acquisition time and improve user convenience while maintaining classification performance comparable to that achieved using conventional recording durations. This trade-off is particularly important for wearable devices, where battery capacity, intermittent measurements, and user comfort often limit the duration of physiological recordings.

V. LIMITATIONS

Although the proposed framework demonstrated promising performance, several limitations should be acknowledged. First, the experiments were conducted using only the WESAD dataset, and therefore the findings should be validated using additional datasets collected under different recording protocols and real-world conditions. Second, only commonly used time-domain HRV features were investigated. Incorporating frequency-domain and nonlinear HRV measures may provide a more comprehensive understanding of the influence of recording duration. Third, only three recording durations (30 s, 120 s, and 300 s) were evaluated. Investigating additional

TABLE IV

BEST CLASSIFICATION PERFORMANCE OBTAINED FOR EACH RECORDING DURATION UNDER THE BINARY AND THREE-CLASS LABELING STRATEGIES.

Label	Duration	Training chunks	Best Model	Accuracy	F1-score	AUC
Three-class	30 s	1491	KNN	0.6566	0.6495	0.7820
Three-class	120 s	369	Gradient Boosting	0.6586	0.6381	0.7605
Three-class	300 s	144	Logistic Regression	0.6739	0.6289	0.7451
Binary	30 s	1491	Gradient Boosting	0.8840	0.8785	0.9133
Binary	120 s	369	Gradient Boosting	0.8726	0.8679	0.9001
Binary	300 s	144	KNN	0.8131	0.8104	0.8115

window lengths may help identify the shortest ECG recording that still provides reliable stress assessment. Finally, shorter ECG windows naturally generated a larger number of training samples than longer windows, which may have influenced the machine learning results independently of recording duration. Future work should investigate this effect using balanced sample sizes.

VI. CONCLUSION

This study investigated the effect of ECG recording duration on the reliability of time-domain HRV features and HRV-based stress classification using the WESAD dataset. Statistical analysis showed that shorter recordings generally reduced agreement with conventional recordings, with Mean HR and Mean RR remaining relatively robust, while SDNN was most sensitive to duration.

Despite this reduced agreement, machine learning results showed that shorter recordings did not compromise stress classification. In particular, 30 s ECG recordings achieved performance comparable to longer recordings and outperformed them in the binary classification task. These findings suggest that statistical agreement does not necessarily translate into higher classification accuracy.

Overall, short ECG recordings can provide sufficient HRV information for stress detection while reducing acquisition time, making them promising for wearable applications where comfort, battery life, and intermittent measurements are important.

Future work will validate the framework on additional real-world datasets, investigate more recording durations, and explore frequency-domain and nonlinear HRV features, as well as deep learning approaches, to further improve stress detection from short ECG recordings.

VII. FUNDING

This work was supported by the Consiglio Nazionale delle Ricerche (CNR) under the Project MU.S.A. - A Multifactorial intervention for Successful Aging (FOE 2021, 2022, 2023)

VIII. SOURCE CODE

The source code used to preprocess the ECG signals, extract the time-domain HRV features, and perform the statistical and machine learning analyses is publicly available at:

<https://github.com/ahmedhassan187/ECG-to-stress>

REFERENCES

- [1] B. S. McEwen, "Protective and Damaging Effects of Stress Mediators," *New England Journal of Medicine*, vol. 338, no. 3, pp. 171–179, 1998.
- [2] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and Health: Psychological, Behavioral, and Biological Determinants," *Annual Review of Clinical Psychology*, vol. 1, pp. 607–628, 2005.
- [3] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study," *Sensors*, vol. 19, no. 8, Art. no. 1849, 2019, doi: 10.3390/s19081849.
- [4] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Frontiers in Public Health*, vol. 5, Art. no. 258, 2017.
- [5] M. Malik *et al.*, "Heart Rate Variability: Standards of Measurement, Physiological Interpretation and Clinical Use," *European Heart Journal*, vol. 17, no. 3, pp. 354–381, 1996.
- [6] A. Tonacci, A. Dellabate, A. Dieni, L. Bachi, F. Sansone, R. Conte, and L. Billeci, "Can Machine Learning Predict Stress Reduction Based on Wearable Sensors' Data Following Relaxation at Workplace? A Pilot Study," *Sensors*, vol. 20, no. 12, Art. no. 3612, 2020, doi: 10.3390/s20123612.
- [7] K. Arquilla, A. K. Webb, and A. P. Anderson, "Utility of the Full ECG Waveform for Stress Classification," *Sensors*, vol. 22, no. 18, Art. no. 7034, 2022, doi: 10.3390/s22187034.
- [8] M. L. Muñoz, A. J. C. van Roon, H. Riese, C. Thio, E. Oostenbroek, I. Westrik, E. J. C. de Geus, R. H. Gansevoort, J. E. Lefrandt, I. M. Nolte, and H. Snieder, "Validity of (Ultra-)Short Recordings for Heart Rate Variability Measurements," *PLoS ONE*, vol. 10, no. 9, Art. no. e0138921, 2015, doi: 10.1371/journal.pone.0138921.
- [9] H. J. Baek, C.-H. Cho, J. Cho, and J.-M. Woo, "Reliability of Ultra-Short-Term Analysis as a Surrogate of Standard 5-Min Analysis of Heart Rate Variability," *Telemedicine and e-Health*, vol. 21, no. 5, pp. 404–414, 2015, doi: 10.1089/tmj.2014.0104.
- [10] U. Nussinovitch, K. P. Elishkevitz, K. Katz, M. Nussinovitch, S. Segev, B. Volovitz, and N. Nussinovitch, "Reliability of Ultra-Short ECG Indices for Heart Rate Variability," *Annals of Noninvasive Electrocardiology*, vol. 16, no. 2, pp. 117–122, 2011, doi: 10.1111/j.1542-474X.2011.00417.x.
- [11] L. Salahuddin, J. Cho, M. G. Jeong, and D. Kim, "Ultra Short Term Analysis of Heart Rate Variability for Monitoring Mental Stress in Mobile Settings," in *Proc. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2007, pp. 4656–4659.
- [12] P. B. B. de Santos Sierra, C. Ávila, J. Guerra-Casanova, and G. Bailón, "Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life," *BMC Medical Informatics and Decision Making*, vol. 19, Art. no. 12, 2019.
- [13] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proc. 20th ACM International Conference on Multimodal Interaction (ICMI)*, 2018, pp. 400–408.
- [14] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, 2021, doi: 10.3758/s13428-020-01516-y.