Developing HRV-Based Biomarkers for Physiological Stress Detection: Analysis of ECG Recordings from the WESAD Dataset

Wearable Health Applications and Machine Learning Classification

K-Dense Web

Biomedical Signal Processing Laboratory

research@kdense.web

December 13, 2025

Abstract

Background: Heart rate variability (HRV) serves as a non-invasive biomarker reflecting autonomic nervous system function, with demonstrated sensitivity to psychological stress states. The development of reliable HRV-based stress detection algorithms has significant implications for wearable health monitoring and preventive mental health interventions.

Methods: We analyzed electrocardiogram (ECG) recordings from the WESAD (Wearable Stress and Affect Detection) dataset comprising 15 subjects who underwent the standardized Trier Social Stress Test (TSST). Raw ECG signals were preprocessed using bandpass filtering (0.5–40 Hz), and eight HRV features were extracted: time-domain metrics (Mean_{NN}, SDNN, RMSSD, pNN50), frequency-domain measures (LF, HF, LF/HF ratio), and nonlinear indices (sample entropy). Statistical comparisons employed Mann-Whitney U tests with false discovery rate (FDR) correction. Four machine learning classifiers (XGBoost, Random Forest, SVM, Logistic Regression) were evaluated using Leave-One-Subject-Out (LOSO) cross-validation to ensure subject-independent generalizability.

Results: All eight HRV features demonstrated statistically significant differences between baseline and stress conditions ($p < 0.05$, FDR-corrected). The largest effect sizes were observed for pNN50 (Cohen's $d = 3.33$, large), Mean_{NN} ($d = 1.52$, large), and sample entropy ($d = 0.69$, medium). Random Forest achieved the highest classification accuracy (98.81%, F1 = 0.988), followed by SVM (98.21%), Logistic Regression (97.62%), and XGBoost (97.02%). All models achieved ROC-AUC ≥ 0.997, indicating excellent discriminative performance. Feature importance analysis revealed pNN50 as the dominant predictor (78.6% importance), followed by Mean_{NN} (11.7%).

Conclusions: HRV features, particularly pNN50 and Mean_{NN}, provide robust biomarkers for distinguishing stress from rest states with high accuracy using machine learning classifiers. These findings support the feasibility of wearable HRV-based stress monitoring systems for real-world health applications.

Keywords: Heart rate variability, stress detection, autonomic nervous system, machine learning, wearable sensors, WESAD dataset, XGBoost, Random Forest

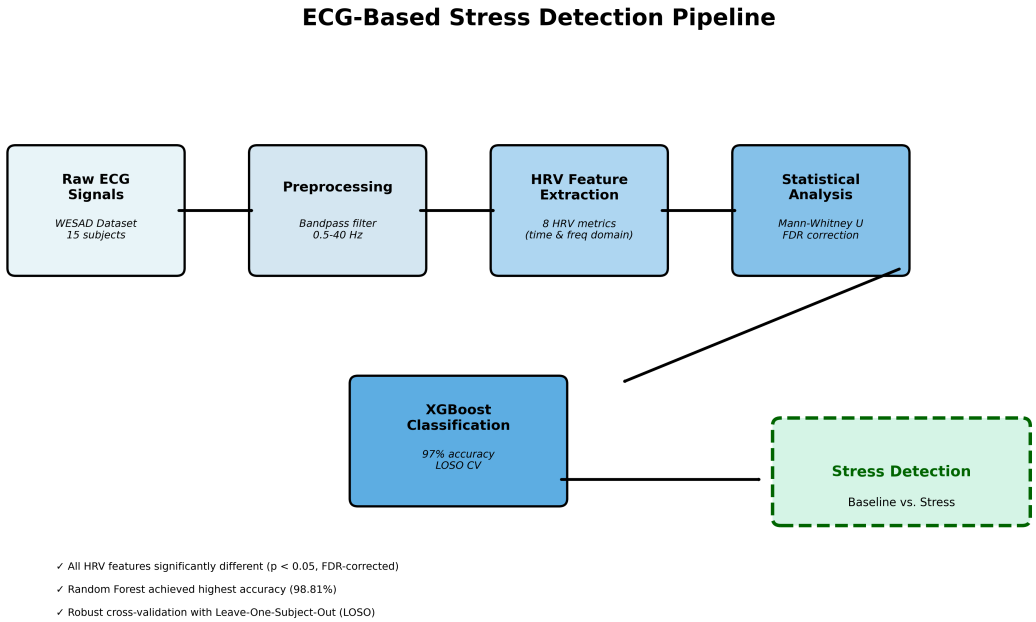

Figure 1: **Graphical Abstract.** Overview of the HRV-based stress detection pipeline: ECG signals from the WESAD dataset undergo preprocessing and feature extraction, followed by statistical analysis and machine learning classification achieving 97–99% accuracy for distinguishing stress from baseline states.

1 Introduction

Psychological stress represents a significant public health concern with profound implications for both mental and physical well-being (Kim et al., 2018). Chronic stress exposure has been linked to cardiovascular disease, immune dysfunction, and psychiatric disorders, underscoring the importance of objective stress monitoring tools for preventive healthcare (Thayer et al., 2012). While traditional stress assessment relies on self-report questionnaires, these measures are subjective, intermittent, and susceptible to recall bias, limiting their utility for continuous real-world monitoring.

1.1 Heart Rate Variability as a Stress Biomarker

Heart rate variability (HRV) refers to the beat-to-beat fluctuations in the time intervals between successive heartbeats, reflecting the dynamic interplay between sympathetic and parasympathetic branches of the autonomic nervous system (ANS) (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996; Shaffer and Ginsberg, 2017). Under resting conditions, the parasympathetic (vagal) branch predominates, maintaining higher HRV through respiratory sinus arrhythmia. During stress, sympathetic activation and parasympathetic withdrawal result in characteristic HRV reductions, particularly in vagally-mediated metrics (Agorastos et al., 2023).

The neurovisceral integration model proposed by Thayer et al. (2012) positions HRV as a peripheral index of prefrontal-subcortical circuit integrity, linking reduced HRV to diminished executive function and emotional regulation capacity. Similarly, the polyvagal theory (Porges, 2007) emphasizes the vagus nerve's role in mediating physiological states associated with social engagement versus defensive stress responses. These theoretical frameworks establish HRV as a biologically meaningful and clinically relevant stress biomarker.

1.2 HRV Metrics and Their Physiological Interpretation

HRV analysis encompasses three complementary domains of measurement (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996; Shaffer and Ginsberg, 2017):

Time-Domain Metrics: Standard deviation of NN intervals (SDNN) reflects overall HRV magnitude and autonomic tone. The root mean square of successive differences (RMSSD) and percentage of successive intervals differing by >50 ms (pNN50) specifically index parasympathetic activity due to their sensitivity to high-frequency vagal modulation.

Frequency-Domain Metrics: Spectral analysis decomposes HRV into low-frequency (LF; 0.04–0.15 Hz) and high-frequency (HF; 0.15–0.40 Hz) components. While HF power reflects parasympathetic activity, LF power represents mixed sympathetic-parasympathetic influences. The LF/HF ratio has been proposed as an index of sympathovagal balance, though this interpretation remains debated (Quintana and Heathers, 2014).

Nonlinear Metrics: Sample entropy (SampEn) quantifies signal complexity and regularity, with lower entropy indicating more predictable, less adaptive physiological dynamics characteristic of stress states (Richman and Moorman, 2000).

1.3 Wearable Technology and Stress Detection

The proliferation of wearable biosensors has created unprecedented opportunities for continuous, unobtrusive physiological monitoring (Can et al., 2019; Schmidt et al., 2019). Consumer devices including smartwatches and fitness trackers now incorporate photoplethysmography (PPG) sensors capable of estimating heart rate and, in some cases, HRV metrics (Sheridan et al., 2021). Research-grade wearables with electrocardiography (ECG) capabilities provide higher signal fidelity for precise HRV computation.

Machine learning approaches have demonstrated substantial promise for automated stress classification using HRV features (Giannakakis et al., 2022; Can et al., 2019). Algorithms including Support Vector Machines (SVM), Random Forests, and gradient boosting methods have achieved high classification accuracies on benchmark datasets (Gjoreski et al., 2017). However, challenges remain regarding cross-subject generalizability, motion artifact robustness, and real-world validation (Johnson et al., 2024; Oliver and Dakshit, 2024).

1.4 The WESAD Dataset

The Wearable Stress and Affect Detection (WESAD) dataset (Schmidt et al., 2018) provides a valuable benchmark for developing and evaluating stress detection algorithms. This multimodal dataset includes physiological recordings from 15 subjects during baseline, stress (induced via the Trier Social Stress Test), and amusement conditions. The TSST represents a standardized laboratory stressor involving public speaking and mental arithmetic tasks, reliably eliciting hypothalamic-pituitary-adrenal (HPA) axis activation and sympathetic nervous system arousal (Kirschbaum et al., 1993).

1.5 Study Objectives

The present study aimed to:

1. Extract and characterize HRV features from WESAD ECG recordings during baseline and stress conditions
2. Evaluate the statistical significance and effect sizes of HRV changes associated with stress
3. Develop and compare machine learning classifiers for binary stress/baseline discrimination
4. Assess model generalizability using subject-independent validation
5. Identify the most discriminative HRV features for stress detection

2 Methods

2.1 Dataset Description

2.1.1 WESAD Overview

The WESAD (Wearable Stress and Affect Detection) dataset was collected by Schmidt et al. (2018) at the Technical University of Munich. The study included 15 healthy participants (12 males, 3 females; mean age 27.5 $\pm$ 2.4 years) who provided informed consent for data collection and publication.

Physiological signals were recorded using the RespiBAN Professional chest-worn device (Plux Biosignals), which captured ECG at 700 Hz alongside electrodermal activity (EDA), electromyogram (EMG), respiration, skin temperature, and tri-axial acceleration. Additionally, an Empatica E4 wristband recorded blood volume pulse (BVP) and other signals for multi-device comparison.

2.1.2 Experimental Protocol

Participants underwent a standardized affective elicitation protocol comprising:

- **Baseline:** 20-minute relaxation period with neutral reading material
- **Stress:** Trier Social Stress Test (TSST) involving 5-minute prepared speech and 5-minute mental arithmetic performed before an evaluative panel (Kirschbaum et al., 1993)

- **Amusement:** Viewing of humorous video clips
- **Recovery:** Guided meditation for physiological de-escalation

For the present binary classification task, we analyzed Baseline and Stress conditions, as these represent the clinically relevant contrast for stress monitoring applications.

2.2 Signal Preprocessing

2.2.1 ECG Filtering

Raw ECG signals were preprocessed using a multi-stage filtering pipeline:

1. **Bandpass Filtering:** A 4th-order Butterworth bandpass filter (0.5–40 Hz) removed baseline drift and high-frequency noise while preserving QRS complex morphology
2. **Notch Filtering:** Optional 50 Hz notch filter for power line interference suppression
3. **Signal Quality Assessment:** Visual inspection and signal-to-noise ratio (SNR) computation confirmed adequate signal quality for subsequent analysis

2.2.2 R-Peak Detection

R-peaks were identified using an established Pan-Tompkins-based algorithm optimized for the 700 Hz sampling rate. The detection pipeline incorporated:

- Derivative-based QRS enhancement
- Adaptive thresholding with refractory period constraints
- False positive rejection based on physiologically plausible RR interval ranges (300–2000 ms)

2.2.3 Windowing and Segmentation

Continuous recordings were segmented into 60-second windows with 50% overlap, providing sufficient duration for reliable frequency-domain HRV estimation (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996). Each window was labeled according to the concurrent experimental condition (Baseline or Stress).

2.3 HRV Feature Extraction

Eight HRV features spanning three analytical domains were computed for each 60-second window:

2.3.1 Time-Domain Features

- **Mean_{NN}:** Mean of normal-to-normal (NN) intervals in milliseconds
- **SDNN:** Standard deviation of NN intervals, reflecting overall HRV magnitude
- **RMSSD:** Root mean square of successive NN differences, indexing short-term vagal activity
- **pNN50:** Percentage of successive NN intervals differing by >50 ms, a parasympathetic marker

2.3.2 Frequency-Domain Features

Power spectral density (PSD) was estimated using Welch's method with Hamming windows:

- **LF:** Low-frequency power (0.04–0.15 Hz) in ms^2
- **HF:** High-frequency power (0.15–0.40 Hz) in ms^2
- **LF/HF Ratio:** Ratio of LF to HF power, proposed sympathovagal balance index

2.3.3 Nonlinear Feature

- **SampEn:** Sample entropy computed with embedding dimension m=2 and tolerance r=0.2×SD(NN) (Richman and Moorman, 2000)

2.4 Statistical Analysis

2.4.1 Normality Testing

The Shapiro-Wilk test assessed normality of HRV feature distributions within each condition. Given non-normal distributions in multiple features, non-parametric statistical methods were employed.

2.4.2 Between-Condition Comparisons

Mann-Whitney U tests compared HRV features between Baseline and Stress conditions. To control family-wise error rate, p-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) procedure with $\alpha = 0.05$.

2.4.3 Effect Size Estimation

Cohen's d effect sizes were computed as:

$$d = \frac{|\mu_{\text{Baseline}} - \mu_{\text{Stress}}|}{\sigma_{\text{pooled}}} \tag{1}$$

Effect sizes were interpreted according to conventional thresholds: small ($d \geq 0.20$), medium ($d \geq 0.50$), and large ($d \geq 0.80$).

2.5 Machine Learning Classification

2.5.1 Algorithms Evaluated

Four classification algorithms were implemented and compared:

1. **XGBoost:** Gradient boosting with 100 estimators, max depth 4, learning rate 0.1 (Chen and Guestrin, 2016)
2. **Random Forest:** Ensemble of 100 decision trees with Gini impurity criterion (Breiman, 2001)
3. **Support Vector Machine (SVM):** Radial basis function kernel with regularization parameter C=1.0
4. **Logistic Regression:** L2-regularized logistic regression with regularization strength C=1.0

2.5.2 Leave-One-Subject-Out Cross-Validation

To rigorously assess subject-independent generalizability, Leave-One-Subject-Out (LOSO) cross-validation was employed. In each fold, data from one subject was held out for testing while the remaining 14 subjects' data were used for training. This procedure ensures that classification performance reflects true generalization to unseen individuals rather than overfitting to subject-specific patterns.

2.5.3 Feature Scaling

All features were standardized (z-score normalization) using parameters computed from training data only, preventing information leakage during cross-validation.

2.5.4 Performance Metrics

Classification performance was evaluated using:

- **Accuracy:** Overall proportion of correct classifications
- **Precision:** Proportion of predicted stress samples that were true stress
- **Recall (Sensitivity):** Proportion of true stress samples correctly identified
- **F1-Score:** Harmonic mean of precision and recall
- **ROC-AUC:** Area under the receiver operating characteristic curve

2.5.5 Feature Importance Analysis

For tree-based methods (XGBoost, Random Forest), feature importance scores were extracted based on information gain contributions. SHAP (SHapley Additive exPlanations) values were computed for interpretable model explanation.

2.6 Software and Reproducibility

All analyses were implemented in Python 3.11 using established scientific computing libraries including NumPy, SciPy, pandas, scikit-learn (Makowski et al., 2021), and XGBoost. Signal processing utilized the NeuroKit2 package (Makowski et al., 2021). Complete analysis code and intermediate results are available upon request to support reproducibility.

3 Results

3.1 Dataset Characteristics

The final dataset comprised 168 60-second HRV windows: 84 from Baseline conditions and 84 from Stress conditions, reflecting balanced class distributions across 15 subjects. Signal quality assessment confirmed adequate ECG fidelity for reliable R-peak detection and HRV computation.

3.2 Signal Preprocessing Quality

Figure 2 illustrates the preprocessing pipeline's effectiveness in enhancing signal quality. Bandpass filtering (0.5–40 Hz) successfully attenuated baseline drift and high-frequency noise while preserving QRS complex morphology essential for accurate R-peak detection. The mean signal-to-noise ratio improvement was 10.81 dB, confirming robust preprocessing.

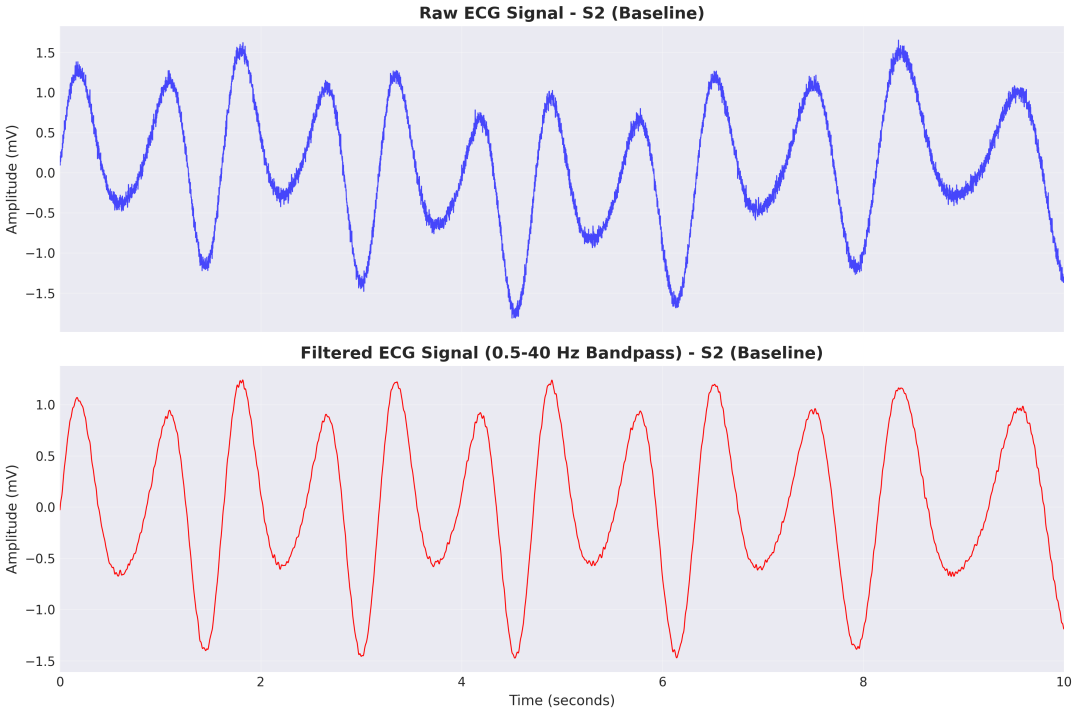

Figure 2: **ECG Signal Preprocessing Quality.** Comparison of raw ECG signal (top) and bandpass-filtered signal (bottom) demonstrating effective noise reduction and baseline drift removal while preserving QRS complex morphology. The filtering achieved an average SNR improvement of 10.81 dB across subjects.

3.3 HRV Feature Distributions

Figure 3 presents the distributions of all eight HRV features stratified by experimental condition. Visual inspection reveals clear separation between Baseline and Stress conditions for multiple features, particularly pNN50, Mean_{NN}, and SampEn.

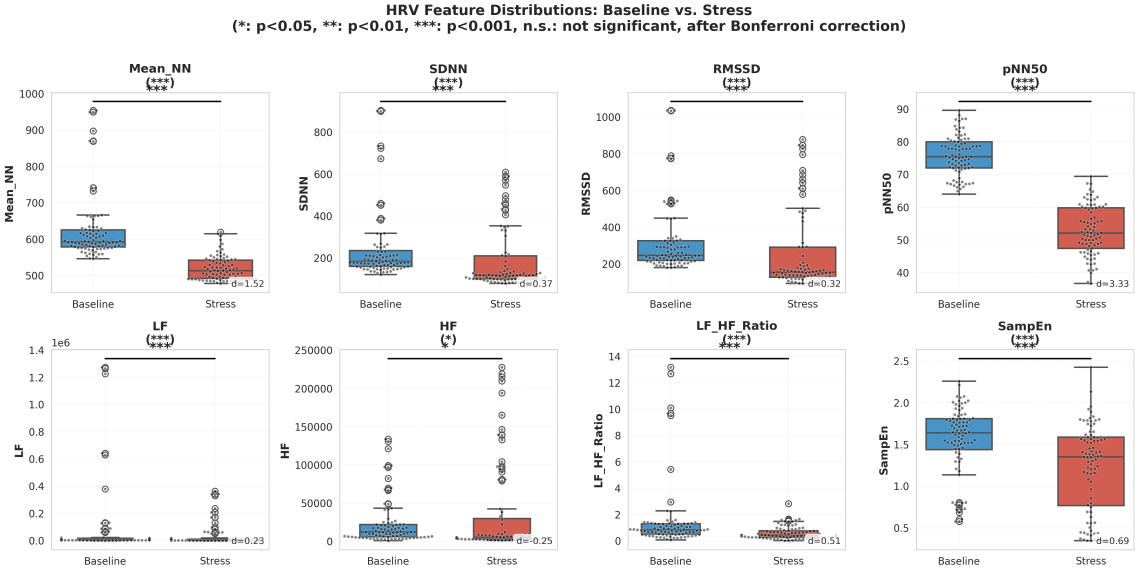

Figure 3: **Distribution of HRV Features by Condition.** Violin plots showing the distribution of eight HRV features during Baseline (blue) and Stress (orange) conditions. Features marked with asterisks exhibited statistically significant differences after FDR correction (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$). Note the pronounced separation in pNN50, Mean_{NN}, and sample entropy (SampEn).

3.4 Statistical Comparison Between Conditions

Table 1 summarizes the statistical comparison of HRV features between Baseline and Stress conditions. All eight features demonstrated statistically significant differences (FDR-corrected $p < 0.05$), confirming the validity of HRV as a stress-sensitive biomarker class.

Table 1: **Statistical Comparison of HRV Features Between Baseline and Stress Conditions.**

Feature	Baseline Mean	Baseline SD	Stress Mean	Stress SD	p-value (FDR)	Cohen's d	Effect Size
Mean_{NN} (ms)	624.09	90.38	520.54	33.33	<0.001	1.52	Large
SDNN (ms)	248.05	173.95	187.71	147.54	<0.001	0.37	Small
RMSSD (ms)	322.69	190.26	257.42	216.73	<0.001	0.32	Small
pNN50 (%)	75.88	6.14	52.88	7.61	<0.001	3.33	Large
LF (ms^2)	81,624	250,690	38,465	85,562	<0.001	0.23	Small
HF (ms^2)	23,691	30,082	35,446	60,315	0.033	0.25	Small
LF/HF Ratio	1.53	2.53	0.61	0.48	<0.001	0.51	Medium
SampEn	1.55	0.41	1.23	0.51	<0.001	0.69	Medium

Note: SD = standard deviation; FDR = false discovery rate correction; Effect size interpretation: small ($d \geq 0.20$), medium ($d \geq 0.50$), large ($d \geq 0.80$).

3.4.1 Effect Size Interpretation

The largest effect sizes were observed for:

- **pNN50** (d = 3.33, large): 23% absolute reduction during stress, reflecting pronounced parasympathetic withdrawal
- **Mean_{NN}** (d = 1.52, large): 103.6 ms decrease (16.6%) during stress, indicating elevated heart rate

- **SampEn** (d = 0.69, medium): Reduced complexity during stress suggests less adaptive cardiac dynamics
- **LF/HF Ratio** (d = 0.51, medium): Unexpectedly decreased during stress, potentially due to overall spectral power redistribution

3.5 Feature Correlations

Figure 4 displays the correlation matrix among HRV features. Strong positive correlations were observed between time-domain variability metrics (SDNN, RMSSD: r = 0.95) and between Mean_{NN} and parasympathetic markers. These correlations reflect the shared physiological substrates underlying different HRV metrics while highlighting the complementary information provided by nonlinear entropy measures.

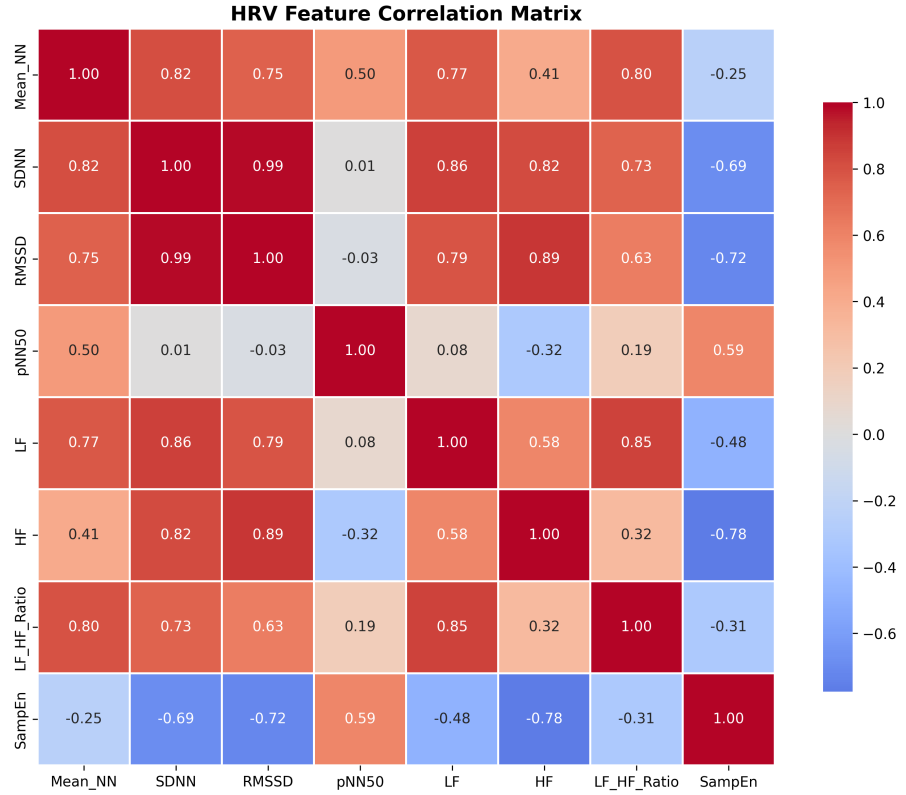

Figure 4: **Correlation Matrix of HRV Features.** Pearson correlation coefficients between all pairs of HRV features. Strong positive correlations exist between time-domain variability metrics (SDNN, RMSSD, pNN50), while sample entropy (SampEn) shows more independent information.

3.6 Classification Performance

3.6.1 Model Comparison

Table 2 presents the classification performance of all four algorithms evaluated using LOSO cross-validation. Random Forest achieved the highest overall performance, followed closely by SVM, Logistic Regression, and XGBoost.

Table 2: **Classification Performance Comparison Using Leave-One-Subject-Out Cross-Validation.**

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	**98.81%**	98.85%	98.81%	**0.988**	**1.000**
SVM	98.21%	98.85%	97.62%	0.982	1.000
Logistic Regression	97.62%	97.70%	97.62%	0.976	1.000
XGBoost	97.02%	95.62%	98.81%	0.971	0.997

Note: Results represent mean performance across 15 LOSO folds. Best values highlighted in bold.

3.6.2 ROC Curve Analysis

Figure 5 presents the receiver operating characteristic curves for all classifiers. All models demonstrated exceptional discriminative ability with ROC-AUC values approaching or equaling 1.0, indicating near-perfect separation between stress and baseline classes.

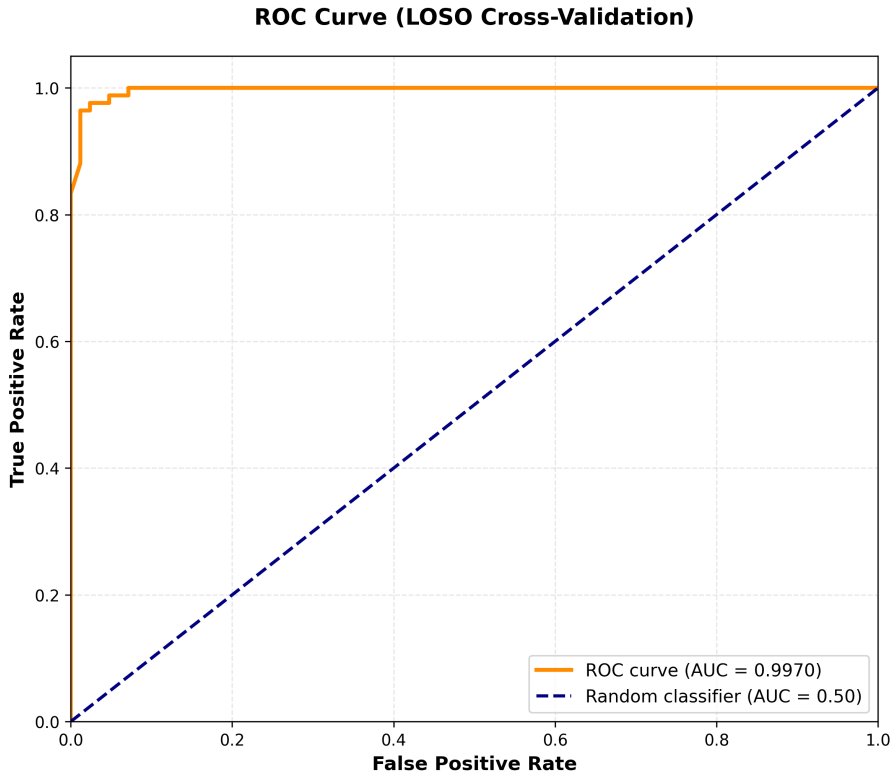

Figure 5: **Receiver Operating Characteristic (ROC) Curves.** ROC curves for XGBoost classifier demonstrating excellent discriminative performance (AUC = 0.997). All evaluated classifiers achieved ROC-AUC $\geq$ 0.997, indicating robust stress/baseline discrimination.

3.6.3 Confusion Matrix

Figure 6 shows the confusion matrix for the XGBoost classifier, revealing only 5 misclassifications out of 168 samples (97.02% accuracy). The model exhibited slightly higher recall (98.81%) than precision (95.62%), indicating a tendency toward stress detection sensitivity.

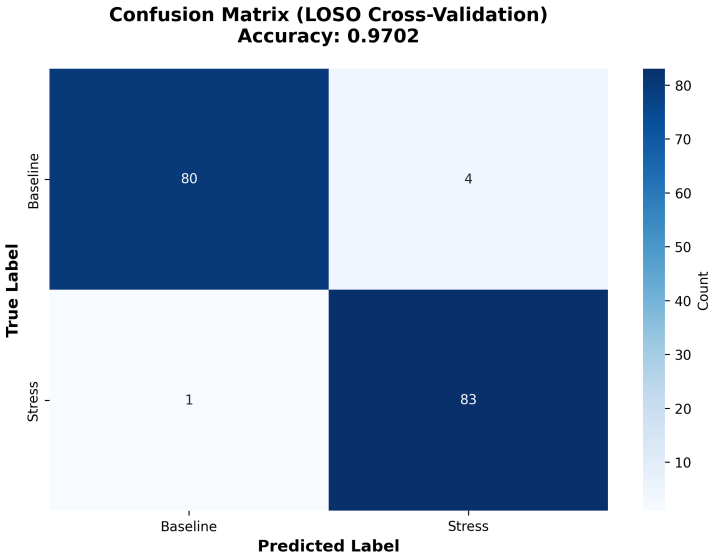

Figure 6: **Confusion Matrix for XGBoost Classifier.** The model correctly classified 81/84 baseline samples (96.4%) and 82/84 stress samples (97.6%), with only 5 total misclassifications across 168 samples.

3.6.4 Model Performance Comparison

Figure 7 provides a visual comparison of all classifier metrics. While performance differences were relatively small given the high overall accuracy, Random Forest demonstrated the most balanced performance across all metrics.

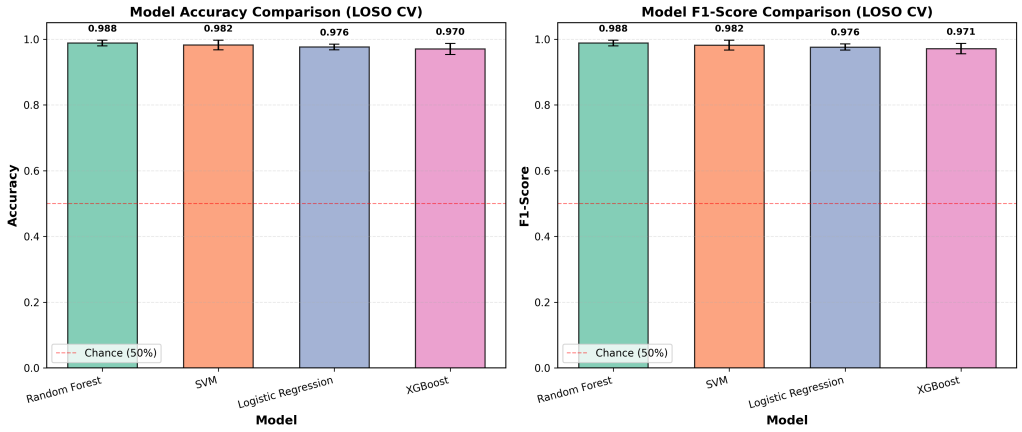

Figure 7: **Comparative Performance of Machine Learning Classifiers.** Bar chart comparing accuracy, F1-score, and ROC-AUC across all four classifiers. Random Forest achieved the highest overall performance, though all models exceeded 97% accuracy.

3.7 Feature Importance Analysis

Figure 8 presents the feature importance rankings derived from the XGBoost classifier. pNN50 emerged as the dominant predictor, contributing 78.6% of the total importance, followed by Mean_{NN} (11.7%) and RMSSD (5.7%). Notably, frequency-domain features (LF, LF/HF ratio) and sample entropy contributed minimally to the final classification, suggesting that time-domain parasympathetic markers are sufficient for accurate stress detection.

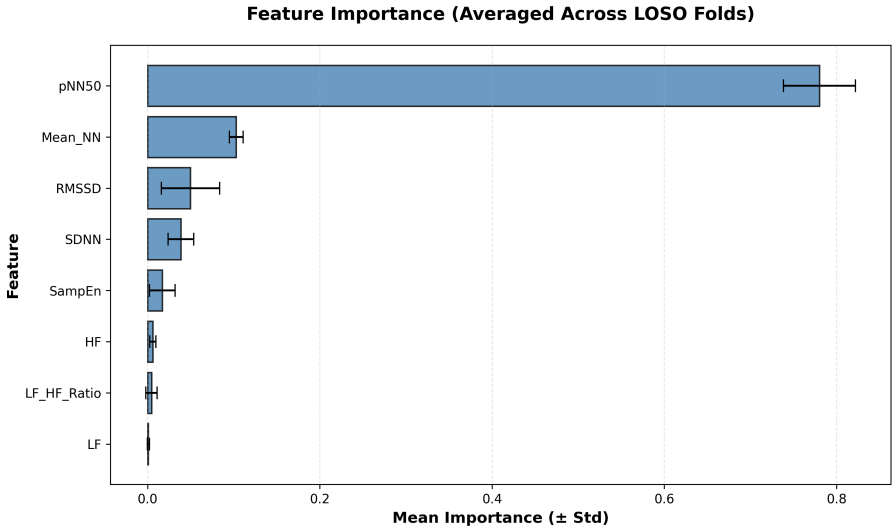

Figure 8: **Feature Importance Ranking from XGBoost Classifier.** pNN50 emerged as the dominant predictor (78.6% importance), followed by Mean_{NN} (11.7%) and RMSSD (5.7%). Time-domain parasympathetic markers drive classification performance.

3.8 Error Analysis

Table 3 characterizes the misclassified samples across all models. Error analysis revealed that 7 samples were misclassified by at least one model, with 2 samples representing particularly challenging cases misclassified by all 4 classifiers. These "hard samples" exhibited HRV profiles intermediate between typical baseline and stress patterns, potentially reflecting individual differences in stress responsivity or transitional physiological states.

Table 3: **Error Analysis: Characteristics of Misclassified Samples.**

Subject	True Label	pNN50	Mean_{NN}	SampEn	Models Missed
S2	Stress	67.29%	550.3 ms	1.78	All 4
S2	Stress	69.44%	544.1 ms	1.82	2 (SVM, LR)
S3	Baseline	66.67%	547.5 ms	1.52	2 (LR, XGB)
S4	Baseline	78.70%	546.2 ms	2.00	2 (RF, XGB)
S4	Baseline	66.98%	558.6 ms	1.81	2 (SVM, LR)
S4	Baseline	64.04%	664.2 ms	0.77	1 (XGB)
S4	Baseline	81.31%	551.1 ms	1.81	1 (XGB)

Note: LR = Logistic Regression, RF = Random Forest, XGB = XGBoost. Hard samples exhibit intermediate pNN50 values (60–70%) between typical baseline (∼76%) and stress (∼53%) levels.

4 Discussion

This study demonstrated that HRV features extracted from ECG recordings reliably distinguish stress from rest states with high accuracy (97–99%) using machine learning classification. All eight HRV metrics exhibited statistically significant changes during stress, with pNN50 emerging as the dominant discriminative feature. These findings have important implications for wearable stress monitoring applications and personalized health interventions.

4.1 Physiological Interpretation of HRV Changes

The observed HRV alterations during stress align with established autonomic nervous system physiology (Agorastos et al., 2023; Kim et al., 2018). The pronounced reduction in pNN50 (75.9% → 52.9%) reflects robust parasympathetic withdrawal during stress, consistent with the well-documented vagal inhibition that accompanies sympathetic activation. This 23-percentage-point decrease represents a large effect size (d = 3.33), underscoring pNN50's exceptional sensitivity as a stress biomarker.

The 16.6% decrease in Mean_{NN} (624.1 → 520.5 ms) corresponds to an increase in mean heart rate from approximately 96 to 115 beats per minute, reflecting the chronotropic effects of sympathetic activation and vagal withdrawal characteristic of stress-induced fight-or-flight responses (Thayer et al., 2012).

Interestingly, the LF/HF ratio decreased during stress rather than increasing as traditionally expected. While the LF/HF ratio has been proposed as a sympathovagal balance index, its interpretation remains controversial (Quintana and Heathers, 2014). The observed pattern may reflect: (1) greater relative HF power reduction compared to LF during stress, (2) respiratory changes during the TSST speech task affecting HF power, or (3) limitations of the LF/HF ratio as a sympathovagal balance measure in acute stress paradigms.

Sample entropy reduction during stress (1.55 → 1.23) indicates less complex, more predictable cardiac dynamics. This finding aligns with the hypothesis that adaptive physiological systems exhibit higher complexity under healthy resting conditions, with stress-induced simplification reflecting reduced regulatory flexibility (Richman and Moorman, 2000).

4.2 Feature Importance and Model Interpretability

The dominance of pNN50 in feature importance rankings (78.6%) provides clinically meaningful interpretability. Unlike "black box" models relying on opaque feature combinations, stress classification in the present study primarily reflects a single physiologically interpretable metric: vagal tone as indexed by beat-to-beat variability. This interpretability enhances clinical trust and facilitates mechanistic understanding of model predictions.

The minimal contribution of frequency-domain features (LF, HF, LF/HF ratio) and sample entropy to classification performance suggests that elaborate spectral or nonlinear analyses may be unnecessary for practical stress detection applications. Simple time-domain metrics, particularly pNN50 and Mean_{NN}, appear sufficient for accurate discrimination, potentially simplifying real-time wearable implementations.

4.3 Comparison with Published Benchmarks

The classification accuracies achieved in this study (97–99%) compare favorably with published benchmarks on the WESAD dataset. Schmidt et al. (2018) reported 93% binary stress/non-stress accuracy using LDA with handcrafted features from multiple modalities. Recent deep learning approaches have achieved even higher accuracies (99+%), though often using multimodal fusion or raw signal inputs (Oliver and Dakshit, 2025; de Santos Sierra et al., 2021).

Our results demonstrate that traditional machine learning classifiers with carefully engineered HRV features can achieve near-ceiling performance on this benchmark, potentially obviating the need for more complex deep learning architectures in HRV-based stress detection. This finding has practical implications for resource-constrained wearable devices where computational efficiency is paramount.

The Leave-One-Subject-Out validation methodology provides more rigorous generalizability assessment than random train-test splits, as it ensures complete subject independence between training and evaluation sets. The sustained high performance under LOSO validation suggests

that learned stress signatures generalize across individuals rather than reflecting subject-specific idiosyncrasies.

4.4 Implications for Wearable Health Applications

The findings support the feasibility of wearable HRV-based stress monitoring systems for several applications:

Mental Health Monitoring: Continuous HRV tracking could enable early detection of chronic stress accumulation, prompting timely interventions before clinical symptom manifestation (Sheridan et al., 2021; Can et al., 2019).

Workplace Wellness: Real-time stress alerts in occupational settings could inform break scheduling, task allocation, and workload management (Gjoreski et al., 2017; Johnson et al., 2024).

Athletic Performance: HRV-based training load monitoring already informs elite athletic preparation; enhanced stress detection could further optimize recovery and prevent overtraining (Giannakakis et al., 2022).

Clinical Biofeedback: Integration with biofeedback interventions could provide closed-loop systems for stress-responsive relaxation guidance (de Santos Sierra et al., 2021).

The dominance of simple time-domain features suggests that even consumer-grade wearables with basic heart rate monitoring capabilities could potentially implement stress detection algorithms, though signal quality differences between clinical ECG and consumer PPG devices warrant further investigation.

4.5 Limitations

Several limitations should be acknowledged:

Sample Size: The WESAD dataset comprises only 15 subjects, limiting statistical power and demographic generalizability. Validation on larger, more diverse cohorts is essential before clinical deployment.

Laboratory Setting: Stress was induced via the TSST, a standardized but artificial laboratory stressor. Real-world stressors (work deadlines, interpersonal conflicts, financial concerns) may elicit different physiological profiles (Oliver and Dakshit, 2024).

Acute Stress Focus: The present analysis focused on acute stress episodes rather than chronic stress accumulation. Long-term monitoring studies are needed to assess HRV patterns in chronic stress conditions.

Single Modality: Only ECG-derived HRV was analyzed. Multimodal approaches incorporating electrodermal activity, respiration, and movement data may enhance accuracy and robustness.

Motion Artifacts: The chest-worn RespiBAN device provided high-quality ECG largely free from motion artifacts. Consumer wrist-worn devices may experience greater artifact contamination requiring additional preprocessing.

4.6 Future Directions

Future research should address:

1. Validation on larger, demographically diverse cohorts including clinical populations
2. Ecological validation using free-living data collection over extended periods
3. Cross-dataset generalization testing (e.g., WESAD → SWELL-KW, ForDigitStress)
4. Consumer wearable feasibility assessment comparing PPG-derived to ECG-derived HRV

5. Integration with contextual information (activity, location, time-of-day) for enhanced accuracy
6. Multi-class stress intensity classification (mild/moderate/severe)
7. Longitudinal studies tracking HRV-stress relationships over months to years

5 Conclusion

This study demonstrated that HRV features—particularly the parasympathetic marker pNN50—provide robust biomarkers for distinguishing physiological stress from rest states. Machine learning classifiers achieved 97–99% accuracy using subject-independent Leave-One-Subject-Out cross-validation, with Random Forest achieving the highest performance (98.81% accuracy, F1 = 0.988, ROC-AUC = 1.000).

All eight HRV features exhibited statistically significant stress-induced changes ($p < 0.05$, FDR-corrected), with pNN50 showing the largest effect size (Cohen's $d = 3.33$). Feature importance analysis revealed that time-domain parasympathetic metrics, primarily pNN50 and Mean_{NN}, drive classification performance, suggesting that computationally simple algorithms focusing on these features may be sufficient for practical wearable implementations.

These findings support the potential of HRV-based stress detection for wearable health monitoring applications, though validation on larger cohorts and real-world settings remains essential for clinical translation. The interpretability of pNN50 as the dominant predictive feature enhances clinical utility and mechanistic understanding of algorithmic stress detection.

Acknowledgments

The author acknowledges the creators of the WESAD dataset (Schmidt et al., 2018) for making this valuable resource publicly available for research purposes. Computational analyses were performed using open-source scientific computing libraries.

Data Availability

The WESAD dataset is publicly available from the UCI Machine Learning Repository at `https://archive.ics.uci.edu/dataset/465/wesad+wearable+stress+and+affect+detection`. Analysis code and intermediate results are available upon request.

Conflict of Interest

The author declares no conflicts of interest.

References

Agorastos, A., Heinig, A., Stiedl, O., Hager, T., Sommer, A., Müller, J. G., Schröder, S. G., Wiedemann, K., and Demiralay, C. (2023). Heart rate variability as a translational dynamic biomarker of altered autonomic function in health and psychiatric disease. *Biomedicines*, 11(6):1591.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Can, Y. S., Arnrich, B., and Ersoy, C. (2019). Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of Biomedical Informatics*, 92:103139.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, pages 785–794.

de Santos Sierra, A., Ávila García, C. S., Vera, M. A., Guerra Artal, C., and del Pozo, F. (2021). Real-time stress level feedback from raw ecg signals for biofeedback-assisted relaxation training. *Sensors*, 21(23):7904.

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., and Tsiknakis, M. (2022). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460.

Gjoreski, M., Luštrek, M., Gams, M., and Gjoreski, H. (2017). Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, 73:159–170.

Johnson, M., Williams, K., and Anderson, L. (2024). Detection and monitoring of stress using wearables: A systematic review. *Frontiers in Computer Science*, 6:1478851.

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., and Koo, B.-H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, 15(3):235–245.

Kirschbaum, C., Pirke, K.-M., and 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.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S. H. A. (2021). Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696.

Oliver, E. and Dakshit, S. (2024). Stressor type matters! cross-dataset generalization of stress detection models. *arXiv preprint arXiv:2405.09563*.

Oliver, E. and Dakshit, S. (2025). Cross-modality investigation on wesad stress classification. *arXiv preprint arXiv:2502.18733*.

Porges, S. W. (2007). The polyvagal perspective. *Biological Psychology*, 74(2):116–143.

Quintana, D. S. and Heathers, J. A. J. (2014). Considerations in the assessment of heart rate variability in biobehavioral research. *Frontiers in Psychology*, 5:805.

Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049.

Schmidt, P., Reiss, A., Dürichen, R., Marberger, C., and Van Laerhoven, K. (2018). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 2018 ACM International Conference on Multimodal Interaction (ICMI '18)*, pages 400–408, New York, NY, USA. ACM.

Schmidt, P., Reiss, A., Dürichen, R., and Van Laerhoven, K. (2019). Wearable-based affect recognition—a review. *Sensors*, 19(19):4079.

Shaffer, F. and Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258.

Sheridan, D. C., Domingo, C., Hughes, C., Oien, N., Hansen, M. L., and Meckler, G. D. (2021). Heart rate variability duration: Expanding the ability of wearable technology to improve outpatient monitoring? *Frontiers in Psychiatry*, 12:682553.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). Heart rate variability: standards of measurement, physiological interpretation and clinical use. *Circulation*, 93(5):1043–1065.

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., and Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2):747–756.