Interpretable Prediction of Short-Term HRV Dynamics from Music Acoustics

Hayato Yasuda
a-hayato@eagle.sophia.ac.jp
Sophia University
Tokyo, Japan

Akiko Takahashi
akiko.xmas@gmail.com
Sophia University
Tokyo, Japan

Tad Gonsalves
t-gonsal@sophia.ac.jp
Sophia University
Tokyo, Japan

Yusuke Fukazawa
fukazawa@sophia.ac.jp
Sophia University
Tokyo, Japan

Abstract

Personalized mental health support systems may benefit from sensing user states during everyday supportive activities such as music listening. This paper presents an exploratory single-listener study on whether acoustic features can predict short-term changes in Root Mean Square of Successive Differences (RMSSD), a heart rate variability index related to parasympathetic activity, during music listening. Using acoustic descriptors extracted from music segments, we trained a Random Forest classifier to distinguish relative increases and decreases in short-term RMSSD dynamics under leave-one-song-out evaluation. The model showed modest predictive performance in the main rolling-window setting, while a stricter non-overlapping sensitivity analysis did not show comparable performance. SHAP-based interpretation suggested that variability in timbre, spectral spread, loudness, and rhythmic activity contributed to the predictions. These results suggest that music acoustics may contain limited but interpretable information about short-term physiological dynamics in this dataset. Rather than demonstrating robust prediction of independent future responses, this work provides an initial step toward interpretable, user-state-aware music-listening support for personalized mental health applications.

Keywords

music listening, heart rate variability, RMSSD, acoustic features, physiological response prediction, interpretable machine learning

1 Introduction

Music-based interventions have been studied for stress reduction and mental health support, with systematic reviews reporting effects on physiological and psychological stress-related outcomes [3, 4] and depressive symptoms [13]. Recent advances in wearable sensing and machine learning create opportunities to assess physiological responses during music listening and to use such feedback in future adaptive support systems. Heart rate variability (HRV) is widely used as a non-invasive index of autonomic nervous system activity [5], and HRV parameters have been used in wearable-based monitoring of stress and anxiety [7]. Among time-domain HRV measures, RMSSD is commonly interpreted as an index related to parasympathetic heart rate modulation [12]. Modeling short-term RMSSD dynamics during music listening may therefore help characterize physiological responses relevant to user-state-aware music-listening support.

Despite growing evidence for music-based interventions, several gaps remain. Many studies evaluate music effects at the group, session, or song level, such as pre-post changes in stress, mood, or physiological indices. Such analyses provide limited insight into how local acoustic dynamics within a song relate to short-term physiological changes. Because music evolves over time and autonomic responses may fluctuate within a piece, analyses are needed that link time-varying acoustic features to subsequent or rolling HRV dynamics. Although prior work has examined associations between musical stimuli and HRV, relatively few studies have formulated this relationship as a predictive task for short-term RMSSD changes under song-level generalization. Musical properties such as tempo, rhythmic complexity, and energy have been related to HRV measures [14], but it remains unclear whether literature-guided acoustic feature sets contain predictive information when applied to unseen songs.

Interpretability is also important if predictive models are to inform health-related or adaptive music-listening systems. Explainable AI has been emphasized for transparency and trust in healthcare applications [1], and interpretability methods provide tools for examining feature contributions and model behavior [8]. In the context of acoustic-feature-based prediction, such methods can help identify musical properties associated with predicted HRV increase or decrease, rather than treating model outputs as opaque predictions.

This study investigates whether acoustic features can predict short-term RMSSD dynamics during music listening in a single-listener exploratory dataset. We frame this task as an initial step toward user-state-aware music-based mental health support, evaluate prediction under leave-one-song-out generalization, and apply SHAP-based interpretation to examine which acoustic dynamics contribute to model predictions.

2 Methodology

2.1 Dataset and Prediction Task

This study was conducted as a single-listener exploratory prediction analysis. The dataset consisted of 12 music listening sessions from a single male listener in his twenties. Each session corresponded to one musical piece, as summarized in Table 1. Because the dataset was obtained from a single listener, this study was designed as

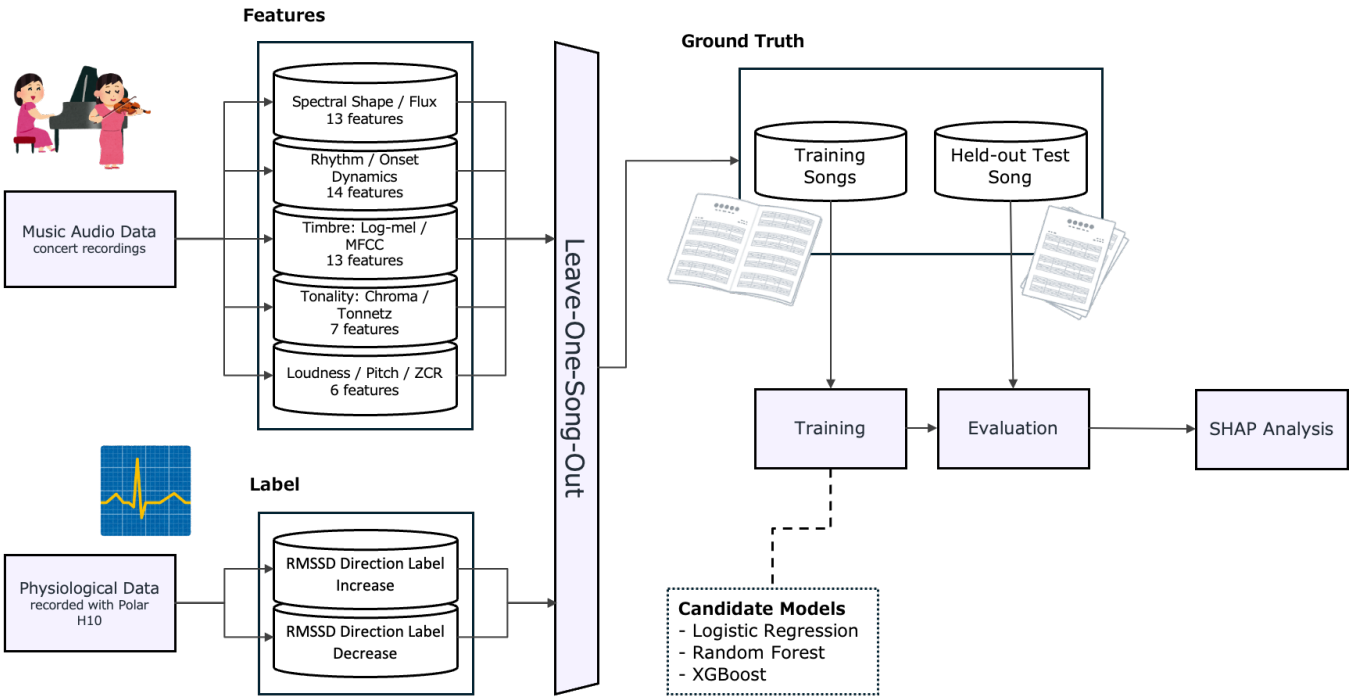

Figure 1: Overview of RMSSD change prediction from music acoustic features.

Table 1: Music listening sessions.

ID	Piece	Inst.	Dur.
001	Harenahare	Vn./Pf.	171
002	Let It Go	Vn./Pf.	222
003	Ocean View	Vn./Pf.	161
004	Merry-Go-Round	Vn./Pf.	264
005	The Swan	Vn./Pf.	152
006	Franck Sonata I	Vn./Pf.	746
007	Chopin Waltz Op.18	Pf.	342
008	Méditation	Vn./Pf.	269
009	Nocturne	Vn./Pf.	220
010	Mélodie	Vn./Pf.	217
011	Brahms Sonata I	Vn./Pf.	672
012	Beauty and Beast	Vn./Pf.	166

an exploratory within-listener prediction analysis rather than a population-level validation study. Detailed participant attributes such as health status were unavailable.

Acoustic features extracted from each music segment were used to classify relative RMSSD increase or decrease within each song, framing the task as physiological response modeling during music listening. The main analysis used 60-second RMSSD windows and a 30-second lag, resulting in partially overlapping RMSSD windows. Model performance was evaluated using leave-one-song-out cross-validation to test generalization to unseen songs rather than random window-level discrimination.

The prediction target was constructed from ΔRMSSD. Within each song, samples in the top and bottom 30% of ΔRMSSD were labeled as HRV increase and HRV decrease, respectively, while the middle 40% were excluded. These labels were used to capture relative within-song RMSSD dynamics rather than clinically defined physiological states.

2.2 Signal Processing and Acoustic Feature Extraction

Physiological data were recorded using a Polar H10 heart rate sensor. The resulting RR interval data were reprocessed for RMSSD analysis following standard time-domain HRV measurement principles [5]. RR intervals were converted to milliseconds, and intervals shorter than 300 ms or longer than 2000 ms were excluded as implausible values. RMSSD was computed over 60-second windows, excluding windows with fewer than 40 RR intervals:

$$\mathrm{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RRI_{i+1} - RRI_i)^2}. \tag{1}$$

RMSSD was used because it is commonly interpreted as a time-domain index related to parasympathetic heart rate modulation [12]. Although conventional HRV analysis often uses longer recordings, 60-second RMSSD has been examined in ultra-short-term HRV research [6].

Let RMSSD_t denote RMSSD computed over the 60-second physiological window starting at time t. The short-term change in RMSSD was defined as:

$$\Delta\mathrm{RMSSD}_t = \mathrm{RMSSD}_{t+L} - \mathrm{RMSSD}_t. \tag{2}$$

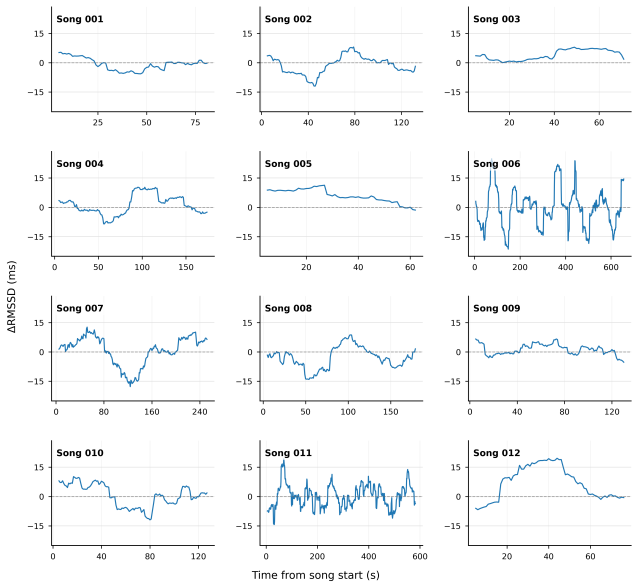

Figure 2: Time series of the RMSSD change target (ΔRMSSD) for each song.

In the main analysis, $L = 30$ seconds; therefore, RMSSD_t and RMSSD_{t+30} partially overlapped by 30 seconds. Accordingly, the main endpoint was interpreted as a short-term rolling RMSSD change rather than an independent future physiological response. A non-overlapping 60-second lag condition was additionally evaluated as a sensitivity analysis. Acoustic features extracted from the 5-second window ending at time t were paired with ΔRMSSD_t.

Acoustic features were extracted from 22,050-Hz audio using 5-second windows with a 1-second hop. The extracted descriptors covered loudness/intensity, rhythm/onset, spectral, zero-crossing, and pitch/voicing-related features, which are commonly used to characterize musically relevant acoustic variation in music information retrieval and music emotion recognition [10]. For the exploratory prediction analysis, we used a fixed 53-dimensional acoustic feature set selected from the 5-second audio feature table. This set combined interpretable low-level acoustic descriptors, past-looking rolling descriptors, and a limited number of dense timbre/tonal descriptors, including selected log-mel, MFCC, chroma, and tonnetz features. One-step delta terms, physiological target columns, and signal-quality columns were excluded from the model inputs. The retained temporal descriptors primarily used 30-row past-looking summaries such as `std30`, `slope30`, and selected `diff_mean30` terms.

2.3 Modeling, Evaluation, and Interpretation

Random Forest was used as the main classifier because it can model nonlinear relationships and feature interactions in tabular data [2]. Although additional models were examined during exploratory analysis, Random Forest was selected as the primary model for reporting because it achieved the best performance in the main setting while maintaining interpretability through tree-based feature attribution. Missing values were imputed using the training-set median in each fold. The model used 150 trees, maximum depth 5, minimum leaf size 8, square-root feature sampling, balanced subsampling for class weights, and random seed 42. Models were implemented using scikit-learn [11].

Performance was evaluated using leave-one-song-out cross-validation. In each fold, one song was held out for testing and the remaining 11 songs were used for training. This group-wise evaluation avoided random splits of temporally adjacent windows from the same song, which could produce overly optimistic estimates. Accuracy, balanced accuracy, F1-score, ROC-AUC, and PR-AUC were computed from pooled out-of-fold predictions.

To inspect model behavior, we used SHAP-based feature attribution analysis [9]. SHAP assigns feature-level importance values to model predictions, enabling global summaries of which acoustic features most influenced HRV increase or decrease predictions. Mean absolute SHAP values were used to summarize global feature importance.

Table 2: Main predictive performance.

	Accuracy	F1	ROC-AUC	PR-AUC
Value	0.6162	0.6115	0.6567	0.6280

Table 3: Main confusion matrix.

	Predicted decrease	Predicted increase
True decrease	469	278
True increase	295	451

3 Results

3.1 Predictive Performance

After excluding the middle 40% of within-song ΔRMSSD values, the main dataset contained 1,493 labeled samples: 746 HRV-increase and 747 HRV-decrease samples. Under the main 60-second-window, 30-second-lag rolling setting, the Random Forest model achieved an accuracy of 0.6162, F1-score of 0.6115, ROC-AUC of 0.6567, and PR-AUC of 0.6280. These results indicate modest above-nominal-chance performance for classifying relative within-song RMSSD changes.

Song-level performance varied substantially across held-out songs. Balanced accuracy ranged from 0.386 to 0.962 across songs, and the macro-averaged song-level balanced accuracy was 0.619. This suggests that the aggregate performance masked heterogeneity in prediction performance across musical contexts.

Table 3 shows the confusion matrix for the main setting. The model correctly classified 469 of 747 HRV-decrease samples and 451 of 746 HRV-increase samples, with errors distributed across both classes.

We also evaluated a non-overlapping 60-second-lag condition, where RMSSD_t and RMSSD_{t+60} were computed over $[t, t+60]$ and $[t+60, t+120]$, respectively. In this stricter setting, Random Forest performance was near chance level, with an accuracy of 0.4961. This result suggests that the main signal may reflect short-term rolling RMSSD dynamics rather than robust prediction of an independent delayed physiological response.

4 Discussion

4.1 Implications for User-State-Aware Support

This study examined whether local acoustic features contain predictive information about short-term rolling RMSSD changes during music listening. The Random Forest model showed modest above-chance performance under leave-one-song-out evaluation, suggesting that acoustic dynamics may capture limited information about within-listener physiological response patterns.

Beyond the aggregate performance, song-level results showed substantial variability across held-out songs. Balanced accuracy ranged from 0.386 to 0.962 across songs, indicating heterogeneous prediction performance across musical contexts. This heterogeneity suggests that acoustic-feature-based RMSSD modeling may be context-dependent. For future adaptive music-listening support, such variability is important because a system may need to estimate not only the direction of a physiological response, but also the reliability of that prediction for a given song or listening context.

4.2 Interpretable Acoustic Cues

Figure 3 shows the top features ranked by mean absolute SHAP values. The highest-ranked features were related to log-mel variability, spectral bandwidth variability, short-term loudness variability, and onset-count dynamics.

The top-ranked features suggest that the model relied primarily on temporally varying acoustic patterns rather than a single static descriptor. Log-mel variability may reflect local timbral or textural changes through frequency-band energy variation, while spectral bandwidth variability reflects changes in the spread of spectral energy. Short-term LUFS variability indicates that loudness dynamics contributed to prediction, and onset-count slope suggests sensitivity to recent changes in rhythmic or event density. Together, these patterns suggest that the model associated predicted RMSSD changes with evolving timbral, spectral, loudness, and rhythmic cues. However, this interpretation is model-based, and SHAP was not used for causal inference.

4.3 Limitations

This study should be interpreted as an exploratory physiological-response modeling study. The dataset was limited to a single listener and 12 mainly classical or film-related pieces, with limited participant context such as health status, psychological state, stress level, and music familiarity. The labels represented relative within-song ΔRMSSD dynamics rather than clinically meaningful improvement or deterioration.

The main 60-second-window, 30-second-lag setting involved partially overlapping RMSSD windows and temporally correlated adjacent samples, and the non-overlapping 60-second-lag sensitivity analysis did not show comparable performance. Because the study was observational, SHAP-ranked acoustic features should be interpreted as model-based predictors rather than causal factors.

5 Conclusion

This single-listener exploratory study investigated whether acoustic features contain interpretable information about short-term rolling RMSSD dynamics during music listening. A Random Forest model showed modest above-chance performance under leave-one-song-out evaluation, and SHAP-based interpretation suggested that timbral, spectral, loudness, and onset-related dynamics contributed to model predictions.

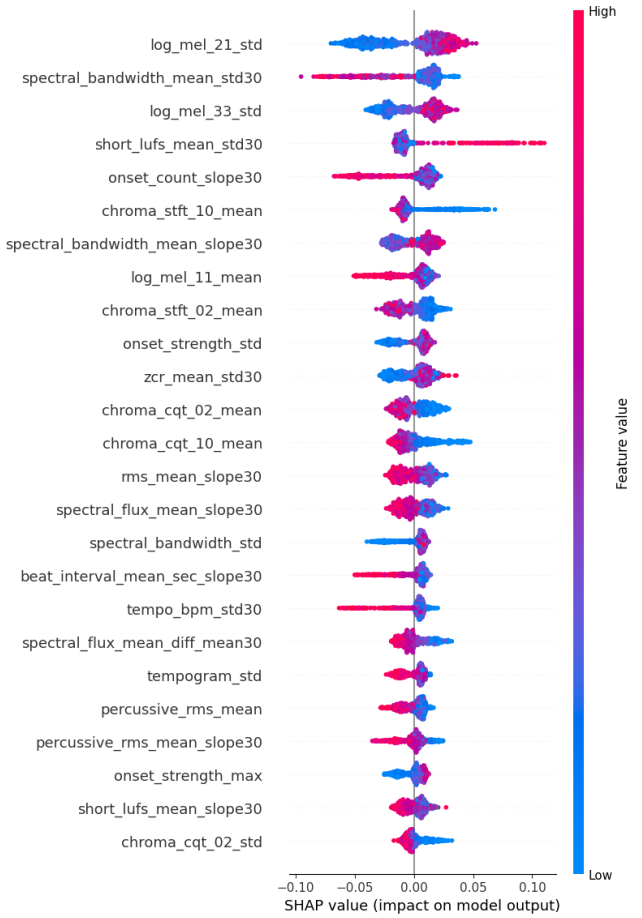

Figure 3: Top features by mean absolute SHAP value.

Because the main target involved partially overlapping RMSSD windows and the non-overlapping sensitivity analysis did not show comparable performance, the findings should not be interpreted as robust prediction of independent future physiological responses. Instead, they provide preliminary evidence that interpretable acoustic-feature-based modeling can characterize short-term physiological response patterns during music listening.

These findings motivate future work with larger multi-participant datasets, subjective listener reports, and longer musical pieces or listening sessions. Subjective measures such as perceived emotion, stress, arousal, preference, and familiarity would help relate RMSSD dynamics to experienced mental states. Longer listening sessions would also enable more reliable analysis of temporal physiological responses and delayed effects. Together, these extensions would help clarify the role of interpretable acoustic-feature-based modeling in future user-state-aware music-listening support.

Acknowledgements

This work was supported by the Data Science Education and Training Hub, Institute of Statistical Mathematics, Research Organization of Information and Systems (ROIS). This research was also supported by the Sophia University Special Grant for Academic Research.

References

[1] Subrato Bharati, M Rubaiyat Hossain Mondal, and Prajoy Podder. 2023. A review on explainable artificial intelligence for healthcare: Why, how, and when? *IEEE Transactions on Artificial Intelligence* 5, 4 (2023), 1429–1442.

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[3] Martina De Witte, Ana da Silva Pinho, Geert-Jan Stams, Xavier Moonen, Arjan ER Bos, and Susan Van Hooren. 2022. Music therapy for stress reduction: a systematic review and meta-analysis. *Health psychology review* 16, 1 (2022), 134–159.

[4] Martina De Witte, Anouk Spruit, Susan Van Hooren, Xavier Moonen, and Geert-Jan Stams. 2020. Effects of music interventions on stress-related outcomes: a systematic review and two meta-analyses. *Health psychology review* 14, 2 (2020), 294–324.

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

[6] Michael R Esco and Andrew A Flatt. 2014. Ultra-short-term heart rate variability indexes at rest and post-exercise in athletes: evaluating the agreement with accepted recommendations. *Journal of sports science & medicine* 13, 3 (2014), 535.

[7] Blake Anthony Hickey, Taryn Chalmers, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. 2021. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors* 21, 10 (2021), 3461.

[8] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.

[9] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[10] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. 2020. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing* 14, 1 (2020), 68–88.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[12] Fred Shaffer and Jay P Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* 5 (2017), 290215.

[13] Qishou Tang, Zhaohui Huang, Huan Zhou, and Peijie Ye. 2020. Effects of music therapy on depression: A meta-analysis of randomized controlled trials. *PloS one* 15, 11 (2020), e0240862.

[14] Hui-Min Wang and Sheng-Chieh Huang. 2014. Musical rhythms affect heart rate variability: Algorithm and models. *Advances in Electrical Engineering* 2014, 1 (2014), 851796.

Appendix

Full List of Acoustic Features

Table 4 lists the 53 acoustic features used in the main analysis. Features are grouped by broader acoustic family. The descriptor `raw` denotes features computed directly from the 5-second analysis window. Temporal descriptors were computed within each song using a past-looking 30-row rolling window, corresponding to approximately 30 seconds given the 1-second hop. Specifically, `std30` denotes the rolling standard deviation of a feature, `slope30` denotes the least-squares linear slope across the rolling window, and `diff_mean30` denotes the current feature value minus its rolling mean. These rolling summaries ended at the current analysis window and did not use future audio frames.

Table 4: Full list of the 53 acoustic features used in the model.

#	Feature	Desc.	Family
1	spectral_bandwidth_mean_std30	std30	spectral
2	zcr_mean_std30	std30	temporal
3	onset_strength_std	raw	rhythm
4	spectral_bandwidth_mean_slope30	slope30	spectral
5	short_lufs_mean_slope30	slope30	loudness
6	rms_mean_slope30	slope30	loudness
7	tempo_bpm_std30	std30	rhythm
8	onset_count_slope30	slope30	rhythm
9	percussive_rms_mean_slope30	slope30	rhythm
10	spectral_bandwidth_std	raw	spectral
11	spectral_centroid_mean_std30	std30	spectral
12	short_lufs_mean_std30	std30	loudness
13	spectral_contrast_01_mean	raw	spectral
14	spectral_bandwidth_mean	raw	spectral
15	onset_strength_mean_std30	std30	rhythm
16	spectral_flux_mean_diff_mean30	diff_mean30	spectral
17	yin_f0_hz_mean_slope30	slope30	pitch
18	onset_strength_mean_slope30	slope30	rhythm
19	spectral_entropy_mean	raw	spectral
20	spectral_rolloff_85_mean_slope30	slope30	spectral
21	tempogram_std	raw	rhythm
22	spectral_contrast_04_mean	raw	spectral
23	yin_f0_hz_mean	raw	pitch
24	percussive_rms_mean_std30	std30	rhythm
25	onset_strength_max	raw	rhythm
26	spectral_flatness_p10	raw	spectral
27	spectral_rolloff_95_p90	raw	spectral
28	onset_count_diff_mean30	diff_mean30	rhythm
29	percussive_rms_mean	raw	rhythm
30	beat_interval_mean_sec_slope30	slope30	rhythm
31	onset_interval_mean_sec	raw	rhythm
32	spectral_flux_mean_slope30	slope30	spectral
33	log_mel_21_std	raw	timbre
34	chroma_cqt_02_std	raw	tonal
35	tonnetz_03_mean	raw	tonal
36	chroma_cqt_02_mean	raw	tonal
37	log_mel_33_std	raw	timbre
38	chroma_stft_02_mean	raw	tonal
39	chroma_cqt_10_mean	raw	tonal
40	delta_delta_mfcc_02_std	raw	timbre
41	delta_mfcc_02_std	raw	timbre
42	chroma_stft_10_mean	raw	tonal
43	log_mel_26_std	raw	timbre
44	log_mel_11_mean	raw	timbre
45	chroma_stft_09_mean	raw	tonal
46	log_mel_02_std	raw	timbre
47	onset_strength_p90	raw	rhythm
48	log_mel_30_std	raw	timbre
49	mfcc_09_std	raw	timbre
50	delta_mfcc_13_std	raw	timbre
51	mfcc_12_mean	raw	timbre
52	delta_mfcc_10_std	raw	timbre
53	mfcc_17_mean	raw	timbre

Per-Song Leave-One-Song-Out Performance

Table 5 reports the per-song performance of the primary random forest model under the leave-one-song-out evaluation. Each row corresponds to one held-out song. The results show substantial variability across held-out songs, suggesting that the predictability of short-term RMSSD changes depended strongly on song-specific acoustic and physiological response patterns. F1 scores were computed using the default decision threshold, whereas ROC-AUC and PR-AUC summarize ranking performance over decision thresholds. In addition to the per-song results, the table reports the macro average across songs and the pooled performance across all held-out predictions.

Non-Overlapping 60-Second-Lag Sensitivity Analysis

To examine whether model performance depended on temporal overlap between the current and future HRV windows, we conducted an additional sensitivity analysis using a 60-second future lag. In the primary setting, RMSSD was computed over 60-second windows and the future window started 30 seconds after the current anchor, resulting in a 30-second overlap between the current and

Table 5: Per-song leave-one-song-out performance of the primary random forest model.

Held-out song	*N*	Acc.	Bal. Acc.	F1	ROC AUC	PR AUC
001	46	.783	.783	.800	.894	.941
002	78	.962	.962	.961	.968	.980
003	40	.750	.750	.773	.755	.694
004	103	.709	.709	.706	.653	.721
005	36	.389	.389	.421	.361	.477
006	392	.635	.635	.581	.681	.626
007	150	.560	.560	.522	.602	.690
008	106	.566	.566	.641	.544	.521
009	76	.645	.645	.658	.663	.561
010	74	.473	.473	.621	.523	.557
011	348	.569	.569	.569	.644	.618
012	44	.386	.386	.000	.605	.519
Macro mean	1493	.619	.619	.604	.658	.659
Pooled	1493	.616	.616	.612	.657	.628

Table 6: Sensitivity analysis comparing the primary overlapping setting and the non-overlapping 60-second-lag setting for the random forest model.

Setting	Overlap sec.	*N*	Acc.	Bal. Acc.	F1	ROC AUC	PR AUC
Primary 30s lag	30	1493	.616	.616	.612	.657	.628
Non-overlap 60s lag	0	1276	.496	.496	.516	.484	.481

future HRV windows. In the sensitivity setting, the future window started 60 seconds after the current anchor, yielding no overlap between the current and future HRV windows.

Table 6 compares the primary overlapping setting with the non-overlapping 60-second-lag setting. Performance decreased to near-chance levels in the non-overlapping setting, suggesting that the primary result should be interpreted as evidence for short-timescale temporal association rather than robust prediction of independent future physiological responses.

Table 7 reports per-song results for the primary random forest model under the non-overlapping 60-second-lag setting. As in the primary leave-one-song-out analysis, performance varied substantially across held-out songs. F1 scores were computed using the default decision threshold, whereas ROC-AUC and PR-AUC summarize ranking performance over decision thresholds.

Class Balance and Sample Counts

Table 8 summarizes the number of candidate rows, labeled rows, and class counts for each song in the primary 30-second-lag dataset. Labels were assigned within each song using the bottom 30% and top 30% of ΔRMSSD values as decrease and increase classes, respectively. The middle 40% of samples were treated as neutral and excluded from model training and evaluation. This within-song thresholding procedure produced nearly balanced positive and negative classes for all songs.

Table 7: Per-song performance in the non-overlapping lag.

Held-out song	*N*	Acc.	Bal. Acc.	F1	ROC AUC	PR AUC
001	28	.679	.679	.571	.755	.827
002	60	.767	.767	.788	.863	.815
003	22	.182	.182	.100	.116	.351
004	84	.345	.345	.444	.280	.421
005	18	.389	.389	.353	.519	.662
006	374	.519	.519	.477	.512	.493
007	132	.417	.417	.025	.338	.401
008	88	.545	.545	.667	.474	.507
009	58	.603	.603	.676	.409	.422
010	56	.321	.321	.367	.337	.514
011	330	.512	.512	.612	.499	.513
012	26	.346	.346	.000	.509	.501

Table 8: Per-song class balance and sample counts.

Song	Cand.	Model	Neg.	Pos.	Neutral
001	77	46	23	23	31
002	128	78	39	39	50
003	67	40	20	20	27
004	170	103	52	51	67
005	58	36	18	18	22
006	652	392	196	196	260
007	248	150	75	75	98
008	175	106	53	53	69
009	126	76	38	38	50
010	123	74	37	37	49
011	578	348	174	174	230
012	72	44	22	22	28
Total	2474	1493	747	746	981

Note. "Cand." denotes candidate rows after temporal alignment and quality filtering. "Model" denotes rows used for model training and evaluation after excluding neutral samples.

Model Comparison and Hyperparameter Settings

Table 9 summarizes model-level performance under the primary 30-second-lag leave-one-song-out setting. The untuned random forest achieved the highest accuracy, balanced accuracy, and ROC-AUC among the fixed-threshold models, and was therefore used as the primary model in the main analysis. Although nested hyperparameter tuning produced comparable performance for XGBoost, it did not improve over the primary untuned random forest. Logistic regression performed substantially worse, suggesting that nonlinear models were better suited to the present feature representation.

Table 10 lists the preprocessing pipelines and hyperparameter settings used for the primary model comparison. For all models, missing values were imputed using the training-set median within each fold. Standardization was applied only to logistic regression. For random forest and XGBoost, tree-based models were trained without feature standardization.

Table 9: Primary 30-second-lag model comparison.

Model	*N*	Acc.	Bal. Acc.	F1	ROC AUC	PR AUC
Logistic regression	1493	.541	.540	.523	.564	.559
Random forest	1493	.616	.616	.612	.657	.628
XGBoost	1493	.607	.607	.601	.646	.624
Random forest, tuned	1493	.604	.604	.597	.642	.612
XGBoost, tuned	1493	.615	.615	.610	.653	.629

Note. Results use the estimator default threshold of 0.5. Tuned models were selected using inner group cross-validation within each outer leave-one-song-out fold.

Table 10: Preprocessing pipelines and fixed hyperparameter settings for the primary model comparison.

Model	Pipeline	Main settings
Logistic regression	Median imputation → standardization → logistic regression	`class_weight=balanced`, `max_iter=5000`, `random_state=42`
Random forest	Median imputation → random forest	`n_estimators=150`, `max_depth=5`, `min_samples_leaf=8`, `max_features=sqrt`, `class_weight=balanced_subsample`, `random_state=42`
XGBoost	Median imputation → XGBoost classifier	`n_estimators=120`, `max_depth=2`, `learning_rate=0.05`, `min_child_weight=5`, `subsample=0.8`, `colsample_bytree=0.8`, `reg_lambda=5.0`, `objective=binary:logistic`, `eval_metric=logloss`, `random_state=42`