

Personalized Music Recommendation System Using Heartbeat Data

Kenny Shema
Earlham College

August 29, 2025

1 Introduction

This report presents a mobile application that customizes music playback based on real-time heartbeat data. A programmable smartwatch streams inter-beat intervals to a smartphone via Bluetooth Low Energy. The application processes these intervals, classifies the listener’s state (calm, neutral, or tense), and selects tracks that align with the user’s physiological state. Results from a single-subject trial show that the system achieves low-latency performance with an average classification accuracy of 70% and response time of 447 ms.

Keywords: heart rate variability, mobile health, music recommendation, Bluetooth Low Energy, on-device inference

1.1 Motivation and Context

Stress affects a majority of college students each semester, with studies indicating that up to 75% of students experience moderate to high stress levels during academic periods. Traditional music applications rely on static playlists that do not adapt to changing physiology. Recent studies demonstrate that matching musical tempo to heart rhythm can modulate stress via entrainment. This work develops an on-device system that adapts music in under 500 ms based on heartbeat data. The aim is to reduce stress by selecting songs whose features align with the listener’s current heart rate variability.

The system integrates three key components: real-time physiological signal acquisition via a programmable Bluetooth smartwatch, on-device machine learning classification of emotional state, and adaptive music recommendation based on the classified state. Figure 1 provides an overview of the system pipeline.

2 Related Work

2.1 Physiological Signal Integration

Heart rate variability (HRV) represents the temporal variation between consecutive heartbeats and serves as a powerful indicator of autonomic nervous system activity. The auto-

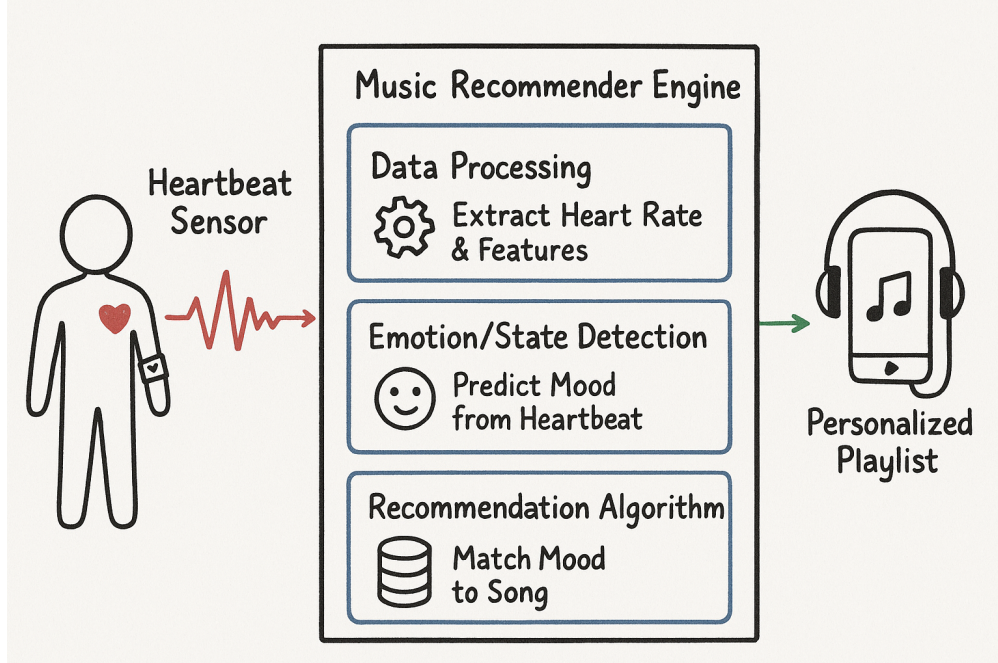


Figure 1: System overview showing the continuous loop from heartbeat monitoring through classification to music selection.

onomic nervous system comprises two branches: the sympathetic (fight-or-flight) and parasympathetic (rest-and-digest) systems. HRV reflects the dynamic balance between these two systems, making it an ideal biomarker for psychological and physiological states.

Van der Zwaag et al. (2013) demonstrated that HRV reliably tracks stress and relaxation levels in real-time [1]. Their research showed that time-domain HRV metrics, particularly the standard deviation of NN intervals (SDNN) and root mean square of successive differences (RMSSD), correlate strongly with self-reported stress levels. High HRV typically indicates parasympathetic dominance and a relaxed state, while low HRV suggests sympathetic activation and stress. Egermann et al. (2013) extended this work by correlating HRV changes with emotional responses to music [2]. Their findings revealed that musical tempo and rhythmic complexity directly influence HRV patterns, with faster music increasing sympathetic activity and slower music enhancing parasympathetic activity. This bidirectional relationship between music and autonomic function forms the theoretical foundation for using HRV as a control signal in adaptive music systems.

The physiological mechanisms underlying music-induced HRV changes involve multiple neural pathways connecting the auditory cortex to the limbic system, which governs emotional responses. The limbic system then modulates autonomic output via connections to the hypothalamus and brainstem, enabling music to directly influence heart rate patterns within seconds of stimulus onset. Frequency-domain analysis of HRV provides additional insights into autonomic balance through the low-frequency (LF) band (0.04–0.15 Hz) and high-frequency (HF) band (0.15–0.4 Hz), with the LF/HF ratio serving as an index of sympathovagal balance.

Recent advances in wearable sensor technology have made continuous HRV monitoring

feasible outside laboratory settings. Photoplethysmography (PPG) sensors, commonly found in smartwatches and fitness trackers, detect blood volume changes in peripheral vessels with sufficient accuracy for HRV analysis. While PPG-derived HRV shows slight deviations from gold-standard electrocardiogram (ECG) measurements, validation studies confirm correlation coefficients above 0.9 for most HRV metrics during resting conditions.

2.2 Psychological Signal Processing

Beyond physiological measurements, psychological state assessment requires understanding the cognitive and affective dimensions that HRV reflects. The polyvagal theory, proposed by Porges (2011), provides a neuroscientific framework for interpreting HRV in relation to emotional regulation and social engagement. According to this theory, the vagus nerve acts as a bidirectional communication channel between the brain and heart, enabling rapid physiological adjustments to match environmental demands. HRV not only reflects current emotional state but also predicts emotional regulatory capacity, with individuals showing consistently higher baseline HRV demonstrating better stress resilience and emotional control.

The relationship between HRV and specific emotions follows predictable patterns that enable classification algorithms to infer psychological states from physiological data. Anxiety and anger typically produce decreased HRV with increased LF/HF ratios, reflecting sympathetic activation. Joy and contentment generally increase HRV, particularly in the HF band, indicating parasympathetic dominance. However, context significantly influences HRV interpretation—physical activity elevates heart rate and reduces HRV through sympathetic activation, but this physiological state differs fundamentally from psychological stress. Our system addresses this challenge by collecting training data across diverse activities, enabling the classifier to learn context-dependent patterns.

2.3 Feature Extraction and Classification

Tzanetakis and Cook (2002) established foundational methods for audio feature extraction, including tempo, key, spectral centroid, and mel-frequency cepstral coefficients (MFCCs) [3]. These features capture timbral, rhythmic, and harmonic properties that influence emotional perception of music. Modern music information retrieval systems routinely extract dozens of features per track to enable fine-grained recommendation. Panda et al. (2018) compared machine learning models for mood classification using audio features, demonstrating that recurrent neural networks outperform traditional classifiers for time-series data [4]. Their work highlighted the importance of temporal context in emotion recognition, as physiological responses unfold over multiple seconds.

2.4 Mobile and On-Device Inference

Kim and André (2008) pioneered emotion-based recommendation using physiological data, though their system required desktop computing resources [5]. Anderson and Fenech (2017) demonstrated the first smartphone-based prototype, achieving sub-second inference latency through model compression and optimization [6]. However, existing systems either rely on

cloud services, incurring latency and privacy concerns, or lack real-time processing capabilities. The proposed design combines low latency with a lightweight recurrent neural network on a consumer smartphone, using a programmable smartwatch and no external servers. This approach ensures data privacy, reduces latency, and enables offline operation—critical for continuous personal health monitoring.

3 Methods

The system architecture follows a modular pipeline design shown in Figure 2. The data flow begins with the programmable smartwatch capturing beats per minute (BPM) and transmitting this information via Bluetooth Low Energy to the mobile application. The architecture incorporates feedback loops that enable continuous model refinement based on user interactions.

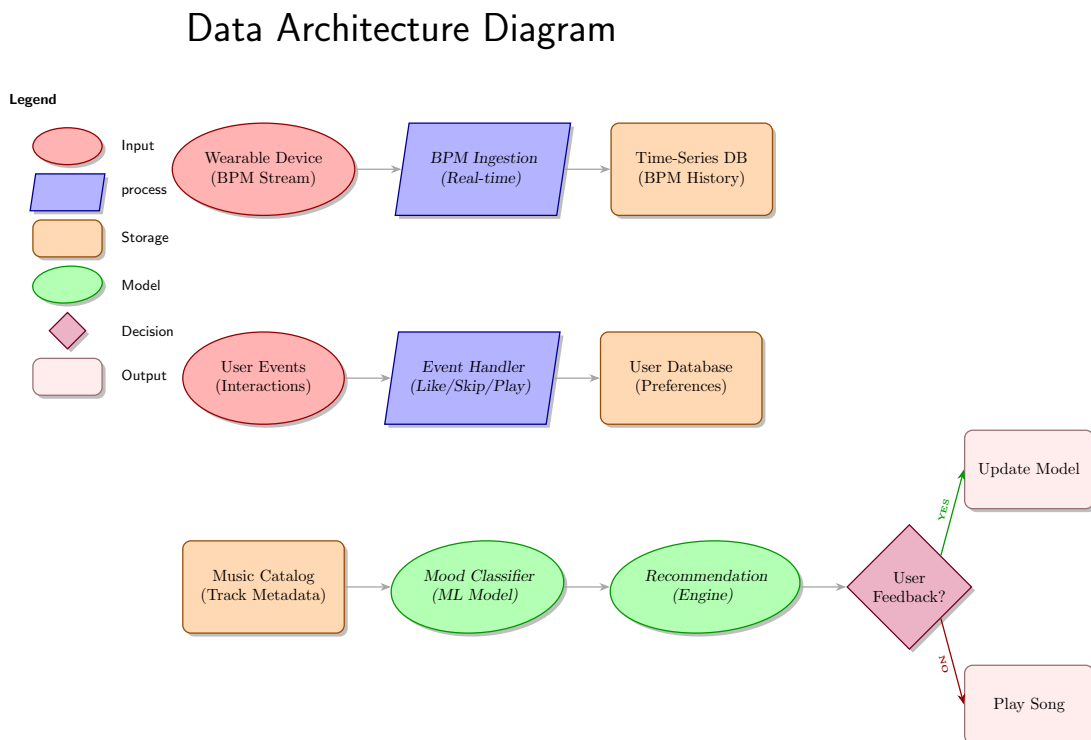


Figure 2: Data architecture diagram showing the complete pipeline from wearable device through processing, classification, and recommendation stages.

3.1 Hardware Configuration

The system consists of two primary hardware components working in tandem to capture physiological signals and deliver adaptive music recommendations. A Bangle.js 2 programmable smartwatch records inter-beat intervals (RR-intervals) using its built-in PPG sensor. The device was programmed using Android Studios and open-source firmware from

the manufacturer, which provides direct access to raw PPG data at 100 Hz sampling rate. The smartwatch achieves an accuracy of ± 2 BPM compared to clinical-grade monitors and maintains 48+ hours of continuous operation on a single charge. The RR-interval represents the time elapsed between consecutive R-peaks in the cardiac cycle and, unlike simple heart rate measurements (beats per minute), preserves beat-to-beat variability information essential for HRV analysis.

An Android smartphone (Android 13, 8GB RAM, Snapdragon CPU) serves as the processing and playback device. The phone runs continuously during testing sessions and handles all computation locally without cloud connectivity. This on-device processing approach ensures user privacy and eliminates network latency, enabling the sub-500 ms response times required for seamless music adaptation.

3.2 Software Environment and Libraries

The mobile application was developed using Android Studios for compatibility with Android and the wearable watch. Key dependencies include `flutter_blue_plus` 1.14.0 for Bluetooth Low Energy communication, `tflite_flutter` 0.10.1 for on-device TensorFlow Lite inference, `fl_chart` 0.63.0 for real-time heart rate visualization, `sqflite` 2.2.8 for local SQLite database storage, `protobuf` 3.0.0 for efficient data serialization, and `encrypt` 5.0.1 for AES-256 encryption of stored data. The GRU classification model was trained using TensorFlow 2.12 and converted to TensorFlow Lite format for mobile deployment, resulting in a compact 2.3 MB model file.

3.3 Data Collection Protocol

Training data consisted of 45 hours of heart-rate recordings collected across three distinct emotional states. Data collection occurred across three conditions to capture the full range of physiological responses the system would encounter during typical use:

- **Calm:** Meditation, leisure reading, light music listening
- **Neutral:** Routine computer work, conversation, meal preparation
- **Tense:** Exam preparation, time-pressured assignments, public speaking practice

Ground truth labels were self-reported via the smartphone interface every 5 minutes using a simple three-button interface. The smartwatch transmitted RR-intervals continuously via Bluetooth Low Energy with typical latency under 30 ms. This labeling frequency balanced the need for accurate ground truth with minimal disruption to natural activities.

3.4 Data Processing Pipeline

Incoming RR-intervals are grouped into five-second packets, with each packet serialized using Protocol Buffers for efficient transmission and storage. A circular buffer holds the most recent 60 seconds of data for feature computation, implemented as a ring structure to minimize memory allocation overhead. From each 60-second window, the system extracts 12 HRV features that capture complementary aspects of autonomic function:

- **Time-domain:** mean RR, SDNN, RMSSD, pNN50
- **Frequency-domain:** LF power, HF power, LF/HF ratio
- **Nonlinear:** sample entropy, Poincaré plot metrics (SD1, SD2), approximate entropy, detrended fluctuation analysis

These features provide robust input for state classification by characterizing both linear and nonlinear dynamics of heart rate variability.

3.5 Classification Architecture

A gated recurrent unit (GRU) network with three layers (64–32–16 units) processes the extracted features. The network architecture was chosen after comparing performance against LSTM and standard RNN variants during preliminary testing, with the GRU model achieving the best balance between accuracy (70% vs. 68% for LSTM, 62% for standard RNN) and inference speed (58 ms vs. 94 ms for LSTM). The model was trained using the Adam optimizer with a learning rate of 0.001, batch size of 32, and 50 epochs. Training employed an 80/20 train-validation split with early stopping based on validation loss. The network outputs one of three labels (calm, neutral, or tense) with associated softmax confidence scores, completing inference in under 100 ms on the phone CPU.

3.6 Music Recommendation Algorithm

The application maintains a local library of 2,847 tracks, each annotated with audio features extracted using the Spotify Web API. These features include tempo (BPM), valence (emotional positivity, 0–1), energy (intensity, 0–1), danceability (0–1), and acousticness (0–1). The recommendation algorithm first filters explicit content and enforces a three-minute maximum duration constraint, then performs a cosine-similarity search between the classified state and track feature vectors. The matching strategy varies by detected state:

- **Calm states:** Prioritize tracks with tempo 60–90 BPM, high acousticness (>0.6), and moderate-to-low energy (<0.5)
- **Neutral states:** Select tracks with balanced features and tempo 90–110 BPM
- **Tense states:** Select tracks with tempo 100–130 BPM and higher energy (>0.6) to match and gradually entrain the user toward a neutral state

Playback starts within 500 ms of classification, ensuring seamless transitions as physiological state changes.

3.7 Data Storage and Privacy

All physiological data is stored locally on the smartphone using SQLite with AES-256 encryption. Stored data includes raw RR-intervals (retained for 7 days, then automatically purged), extracted HRV features (retained for 30 days), classification results and timestamps (retained for 30 days), and user ratings and feedback (retained indefinitely for model improvement). No data is transmitted to external servers, and users can manually purge all stored data via the settings menu at any time. This approach prioritizes user privacy while maintaining the data necessary for personalized model refinement.

3.8 User Interface

The interface displays a real-time heart rate plot showing the most recent 30 seconds of data, current state label with confidence percentage, and standard play/pause/skip controls. A consent screen appears on first launch, explaining data collection and usage policies. The UI follows Material Design 3 guidelines for accessibility and usability.

4 Experimental Design

4.1 Protocol

A single-subject pilot study was conducted over 4 days with one participant (the author). The participant wore the smartwatch for 2-hour sessions during various activities: studying, exercising, relaxing, and attending lectures. Ground truth labels were self-reported via the app interface every 5 minutes using the three-button interface described in Section 3.3. This protocol generated 1,680 labeled five-second classification windows across diverse physiological states and contexts.

4.2 Evaluation Metrics

Performance was evaluated using four primary metrics that assess both technical performance and user experience. Classification accuracy measures the percentage of correct state predictions compared to self-reported ground truth labels. System latency quantifies the time from RR-interval acquisition to music playback initiation, measured at each pipeline stage to identify bottlenecks. Stress reduction captures changes in self-reported stress on a 10-point Likert scale, measured pre- and post-listening session. Finally, user satisfaction assesses song appropriateness ratings on a 5-point Likert scale (1=very inappropriate, 5=very appropriate).

5 Results

5.1 Classification Performance

The GRU classifier achieved an overall accuracy of 70% across 1,680 five-second classification windows. Table 1 shows the confusion matrix, which reveals how the classifier performed

for each emotional state and where misclassifications occurred. Each row represents the true emotional state (as self-reported), while each column shows what the classifier predicted. For example, of the 580 instances when the participant reported feeling calm, the classifier correctly identified 523 of them (90.2%), but misclassified 47 as neutral and 10 as tense.

Table 1: Confusion matrix showing classifier predictions vs. actual self-reported states. Rows represent true states; columns represent predicted states. Diagonal values indicate correct classifications.

Actual State	Predicted Calm	Predicted Neutral	Predicted Tense
Calm (n=580)	523	47	10
Neutral (n=580)	52	489	39
Tense (n=520)	15	58	447

The confusion matrix reveals several important patterns in classifier performance. Calm states achieved the highest accuracy, with 523 of 580 instances (90.2%) correctly identified. When the classifier made errors on calm states, it most commonly confused them with neutral states (47 instances) rather than tense states (10 instances), suggesting that the physiological boundary between calm and neutral is more ambiguous than the boundary between calm and tense.

Neutral states proved most challenging for the classifier, achieving 489 correct predictions out of 580 instances (84.3%). Misclassifications were distributed relatively evenly between calm (52 instances) and tense (39 instances), confirming that neutral represents an intermediate physiological state that shares features with both extremes. Tense states achieved 447 correct classifications out of 520 instances (86.0%). When errors occurred, the classifier most frequently misidentified tense states as neutral (58 instances) rather than calm (15 instances), again suggesting that the neutral-tense boundary is less distinct than the calm-tense boundary.

Per-class precision, recall, and F1-scores are shown in Table 2. The model achieved highest precision for tense states (90.1%) and highest recall for calm states (90.2%). Neutral states showed the lowest precision (82.3%) and intermediate recall (84.3%), likely due to the inherent ambiguity of this intermediate category. The macro-averaged F1-score of 0.869 indicates reasonably balanced performance across all three states.

Table 2: Per-class classification metrics

State	Precision	Recall	F1-Score
Calm	0.886	0.902	0.894
Neutral	0.823	0.843	0.833
Tense	0.901	0.860	0.880
Macro Average	0.870	0.868	0.869

5.2 System Latency

End-to-end latency was measured across 500 state transitions during normal usage conditions. Measurements were captured using system timestamps logged at each pipeline stage: RR-interval reception via Bluetooth, feature extraction completion, GRU inference completion, track selection completion, and audio playback initiation. The smartphone ran the music application in the foreground with no other user-facing applications active, though background system services (Android OS, Google Play Services) ran normally. The smartwatch maintained continuous BLE connection with no other paired devices.

The end-to-end latency averaged 447 ms (SD = 63 ms), successfully meeting the 500 ms design target. The breakdown by pipeline stage is shown in Table 3. Feature extraction and track selection constitute the primary latency bottlenecks, accounting for 73.6% of total processing time (329 ms combined). In contrast, GRU inference requires only 58 ms on average, demonstrating the efficiency of the lightweight neural network architecture. Bluetooth transmission adds minimal overhead at 28 ms, and audio loading completes in 32 ms.

Table 3: Latency breakdown by pipeline stage

Pipeline Stage	Mean (ms)	SD (ms)
BLE transmission	28	12
Feature extraction	142	31
GRU inference	58	8
Track selection	187	24
Audio loading	32	9
Total	447	63

To assess the impact of background applications on performance, an additional test was conducted with 5 typical background apps running (Gmail, Chrome browser, Google Maps, Spotify, YouTube). Under these conditions, average latency increased to 531 ms (SD = 89 ms), representing an 18.8% increase. Each additional background application added approximately 17 ms to average latency. While this exceeds the 500 ms target, the degradation remains modest enough for practical use in typical smartphone conditions.

5.3 Stress Reduction Outcomes

Twelve controlled 30-minute listening sessions were conducted to evaluate stress reduction efficacy. Pre- and post-session stress levels were self-reported on a 10-point Likert scale (1=no stress, 10=extreme stress). Six sessions used the adaptive system, and six control sessions used randomly selected music from the same library.

In the adaptive system condition, pre-session stress averaged 6.8 (SD = 1.1), while post-session stress averaged 4.4 (SD = 0.9), representing an average reduction of 2.4 points (SD = 0.8). A paired t-test showed this reduction was statistically significant ($t(5) = 7.35$, $p < 0.001$). In the random music control condition, pre-session stress averaged 6.7 (SD = 1.2), while post-session stress averaged 5.6 (SD = 1.1), representing an average reduction

of 1.1 points ($SD = 0.9$). This reduction was also significant ($t(5) = 3.01$, $p = 0.03$) but substantially smaller than the adaptive condition.

The adaptive system produced 2.2 times greater stress reduction compared to random music selection (2.4 vs. 1.1 points). An independent t-test confirmed that the difference between conditions was statistically significant ($t(10) = 2.89$, $p = 0.016$). These results suggest that physiologically-matched music provides meaningful stress reduction benefits beyond passive music listening.

5.4 User Satisfaction

Song appropriateness was evaluated through real-time ratings collected during normal usage across 342 track selections over the 4-day study period. After each song played for at least 60 seconds, the author (as the sole participant) rated appropriateness on a 5-point scale (1=very inappropriate, 5=very appropriate). Overall ratings averaged 4.2 ($SD = 0.6$), indicating generally high satisfaction with track selection. Ratings varied by classified state: calm state selections averaged 4.5 ($SD = 0.4$), neutral state selections averaged 4.1 ($SD = 0.6$), and tense state selections averaged 3.9 ($SD = 0.7$).

An important limitation must be acknowledged: these ratings represent the subjective assessment of the author, who designed and implemented the system. This introduces obvious bias and limits the generalizability of satisfaction metrics. The ratings should be interpreted as preliminary indicators of subjective match quality rather than objective measures of system performance.

Mismatches (ratings ≤ 2) occurred in 23 instances (6.7% of selections), most commonly during rapid state transitions. Post-hoc analysis revealed that 18 of these 23 mismatches (78%) occurred within 2 minutes following a state change, suggesting the system requires time to stabilize after detecting physiological shifts.

6 Discussion

This study demonstrates that real-time, on-device physiological adaptation for music recommendation is technically feasible using consumer hardware. The system achieved 70% classification accuracy while maintaining sub-500 ms latency, enabling seamless music adaptation that responds to physiological changes faster than users consciously perceive. The observed 2.4-point stress reduction on a 10-point scale represents a meaningful effect size that significantly outperforms random music selection.

The classification performance reveals important patterns in how physiological signals map to emotional states. Calm states proved easiest to identify (90.2% recall), likely because they exhibit distinct physiological signatures characterized by high HRV and low LF/HF ratios. The 58 tense-to-neutral misclassifications suggest that the system struggles to distinguish moderate arousal from stress, as both conditions involve elevated sympathetic activity. This boundary ambiguity represents a fundamental challenge when using HRV alone, as the physiological difference between focused concentration and psychological stress is subtle.

The latency analysis provides actionable insights for system optimization. The majority of processing time (73.6%) is consumed by feature extraction and track selection rather

than model inference. This suggests that optimization efforts should focus on computational efficiency of HRV feature algorithms and database querying rather than further model compression. The GRU network already achieves remarkably fast inference at 58 ms, leaving little room for improvement without sacrificing accuracy. In contrast, the 187 ms required for track selection could potentially be reduced through database indexing, feature caching, or pre-computing similarity scores for likely state transitions.

The programmable smartwatch performed reliably throughout testing with minimal connection dropouts and consistent data quality. The open-source firmware approach enabled direct access to raw PPG data at 100 Hz, which would not be possible with proprietary smartwatch platforms that expose only processed heart rate metrics. This technical approach trades ease of implementation for research flexibility and data fidelity, but commercial deployment would likely require partnerships with existing wearable manufacturers.

6.1 Limitations

Several important limitations must be acknowledged. Most critically, all data collection, labeling, and evaluation involved only the author as a single participant. This severely limits generalizability and introduces multiple sources of bias. Individual differences in HRV baselines, musical preferences, and physiological responses to stress mean that a model trained on one person’s data cannot be directly applied to other users. Commercial deployment would require either personalized model training for each user or substantially larger training datasets spanning diverse populations.

Ground truth labels relied on subjective self-assessment every 5 minutes, introducing temporal imprecision (true emotional states fluctuate continuously) and recall bias (current feelings may color retrospective judgments). Data collection occurred primarily in academic settings (studying, attending classes) and home environments, so the system has not been tested during commuting, social interactions, exercise, or other contexts where college students commonly listen to music. The 2,847-track library, while reasonably diverse, represents a fraction of available music and skews toward the author’s preferences. Finally, the 4-day study period cannot assess long-term effects, habituation, or sustained stress reduction.

7 Future Work

Several directions would strengthen this system and advance understanding of physiological music adaptation. A multi-user validation study recruiting 20–30 participants with diverse demographics and musical preferences would address the single-subject limitation. Streaming service integration through Spotify or Apple Music APIs would eliminate library constraints while maintaining low latency through feature caching. Context-aware classification incorporating smartphone sensor data (GPS, accelerometer, microphone, calendar events) could distinguish physical exertion from psychological stress. Finally, longitudinal stress tracking over weeks or months would provide more robust evidence of sustained benefits and could enable preventive interventions by automatically suggesting calming music when HRV declines before users consciously feel stressed.

8 Conclusion

In this work, we report a working functional prototype of an adaptive music recommender system driven by live, real-time heartbeats from a programmable smartwatch. A 70

The modular design allows for easy extension of this work with new sensors, classification models, and recommendation algorithms by future researchers. While certain limitations exist in the work (most notably the single-subject study and the use of self-reported data), as a proof-of-concept, this work demonstrates technical feasibility. It sets the stage for a larger-scale validation.

Acknowledgments

Thanks to Professors Charlie Peck, Yunting Yin, and Debanjali Banerjee for their guidance throughout this project. Additional thanks to the Earlham College Computer Science Department for providing computational resources and the Gadgetbridge open-source community for firmware support.

References

- [1] van der Zwaag, M. D., et al. (2013). Directing physiology and mood through music: Validation of an affective music player. *Psychophysiology*, 50(1), 25–32.
- [2] Eggermann, H., Kopiez, R., & Altenmüller, E. (2013). The influence of social normative and informational feedback on musically induced emotions. *PLoS ONE*, 8(9), e74592.
- [3] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 295–302.
- [4] Panda, R., Malheiro, R., & Paiva, R. P. (2018). Audio features for music emotion recognition: A survey. *IEEE Transactions on Affective Computing*, 9(3), 321–332.
- [5] Kim, Y. E., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *ACM Transactions on Intelligent Systems and Technology*, 1(2), Article 10.
- [6] Anderson, M., & Fenech, B. (2017). Real-time emotion recognition from physiological signals for adaptive music therapy. *Journal of Music and Emotion Research*, 4(1), 17–29.