# The Effect of Background Music on Task Performance Across Different Workload Levels

Ngoc Thien An Le*
le000675@umn.edu
Department of Chemical Engineering & Materials Science
University of Minnesota Twin Cities
Minneapolis, MN, USA

Owen Peterson*
pet04238@umn.edu
Department of Computer Science & Engineering
University of Minnesota Twin Cities
Minneapolis, MN, USA

Nour Hany*
hany0003@umn.edu
Department of Computer Science & Engineering
University of Minnesota Twin Cities
Minneapolis, MN, USA

Ahmed Sameh*
sameh002@umn.edu
Department of Computer Science & Engineering
University of Minnesota Twin Cities
Minneapolis, MN, USA

## Abstract

Background music is commonly used during cognitively demanding activities, yet its measurable influence on physiological responses and task performance remains inconsistent across individuals. In this project, we analyze a multimodal dataset of working-memory performance collected under two music conditions (calming vs. vexing) and two workload levels (1-back vs. 3-back). Using behavioral features (e.g., response time and correctness) alongside physiological signals from wearable sensors (Empatica and Biopac) and fNIRS recordings, we study two classification problems: (1) predicting cognitive workload within each music condition, and (2) predicting music condition from multimodal signals across workload levels. To address the high-dimensional, small-sample nature of the data and strong inter-subject variability, we compare four complementary model families: XGBoost, K-Nearest Neighbors, a multilayer perceptron, and the RIPPER rule learner. We evaluate performance using subject-wise splits and report accuracy and F1-score before and after feature engineering and hyperparameter tuning. Across experiments, workload prediction is consistently more learnable than music-condition prediction, suggesting that the extracted feature space contains stronger and more stable signal for cognitive load than for background music type. Notably, KNN provides the best performance for music-condition prediction after optimization, while other models trend toward near-chance accuracy, highlighting the challenges of capturing subtle and subjective music-related effects from limited data.

## CCS Concepts

• **Information systems → Data mining**; • **Computing methodologies → Supervised learning**; *Feature selection*; *Neural networks.*

## Keywords

multimodal learning, cognitive workload prediction, n-back task, physiological signals, fNIRS, feature engineering, subject-wise evaluation

**ACM Reference Format:**
Ngoc Thien An Le, Owen Peterson, Nour Hany, and Ahmed Sameh. 2024. The Effect of Background Music on Task Performance Across Different Workload Levels. In *Proceedings of CSCI 5523 Final Project*. ACM, New York, NY, USA, 8 pages. https://doi.org/00.0000/0000000.0000000

## 1 Introduction

Background music is frequently used during cognitively demanding activities such as studying, programming, or problem solving, with the expectation that it may improve focus or performance. From a cognitive perspective, music is often viewed as a mechanism for modulating arousal, which in turn can influence task performance. The Yerkes–Dodson law describes a non-linear relationship between arousal and performance, suggesting that moderate arousal may enhance performance while excessive arousal degrades it [14]. Despite this long-standing theory, empirical findings on the effects of background music remain mixed, particularly when individual differences and task difficulty are taken into account.

Recent advances in wearable sensing and neuroimaging have enabled the study of cognitive workload using multimodal physiological data. Signals such as electrodermal activity, heart rate, respiration, and functional near-infrared spectroscopy (fNIRS) have been shown to reflect changes in mental effort and task demands [1, 5, 11]. These modalities provide complementary views of autonomic and cortical responses, allowing cognitive states to be examined beyond self-report measures. However, modeling such data remains challenging due to high dimensionality, measurement noise, and strong inter-subject variability, particularly when the number of participants is limited.

In this project, we analyze a multimodal dataset collected during a working-memory experiment in which participants performed n-back tasks under two background music conditions: calming and vexing [9, 10]. Each participant completed both low-workload (1-back) and high-workload (3-back) tasks while behavioral data and physiological signals from wearable sensors and fNIRS were

[*] Authors contributed equally to this work.

*CSCI 5523 Final Project, Minneapolis, MN, USA*
2024. ACM ISBN 000-0-0000-0000-0/24/12
https://doi.org/00.0000/0000000.0000000

recorded. This experimental design enables the joint study of task-driven cognitive load and music-induced arousal within the same individuals.

The primary motivation of this study is not merely to build predictive models, but to investigate how cognitive workload and background music relate to observable physiological and behavioral responses. Specifically, we are interested in understanding (1) whether increasing task difficulty induces consistent, learnable patterns across behavioral, autonomic, and cortical signals, and (2) whether background music condition independently influences these responses, beyond the effects of workload. Supervised classification is used as an analytical tool to probe these relationships: if a condition can be reliably classified from the extracted features, this suggests the presence of systematic and discriminative structure in the underlying signals.

To operationalize these questions, we formulate two supervised classification tasks. First, we classify cognitive workload (1-back vs. 3-back) separately under calming and vexing music conditions, allowing us to examine whether workload-related patterns differ across arousal contexts. Second, we attempt to classify music condition itself across mixed workload levels, testing whether music produces distinct physiological or behavioral signatures independent of task difficulty. Performance differences across these tasks provide insight into the relative strength and consistency of workload-driven versus music-driven effects.

We evaluate a diverse set of models with complementary inductive biases, including a tree-based ensemble method (XGBoost), a distance-based classifier (K-Nearest Neighbors), a feedforward multilayer perceptron (MLP), and a rule-based learner (RIPPER) [2–4]. These models span a range of representational capacities and interpretability, making them suitable for exploratory analysis on small, high-dimensional physiological datasets. Model performance is assessed using subject-wise train, validation, and test splits to evaluate generalization across individuals, with accuracy and F1-score as primary metrics [12].

Rather than focusing solely on maximizing classification accuracy, this work emphasizes interpretation and failure analysis. By comparing results before and after feature engineering and hyperparameter optimization, we examine how feature relevance, noise, and subject variability affect model behavior. The results reveal a consistent asymmetry between the two tasks: cognitive workload is more robustly predicted than music condition, suggesting that workload induces stronger and more stable physiological and behavioral signatures than background music. These findings help clarify the limits of physiological modeling for subtle affective influences and motivate future work on richer representations, task-stratified analysis, and larger-scale studies.

# 2 Dataset and Experimental Design

## 2.1 Data Source

The data used in this project were obtained from a publicly available multimodal working-memory dataset hosted on PhysioNet [9]. The dataset was collected as part of a controlled experimental study designed to investigate the effects of background music on cognitive workload and physiological responses. An accompanying publication provides detailed documentation of the experimental protocol, sensor configurations, and data collection procedures [10]. The availability of synchronized behavioral, physiological, and neuroimaging signals makes this dataset well suited for studying cognitive workload under different arousal conditions.

## 2.2 Participants, Tasks, and Conditions

The dataset includes recordings from five participants, each of whom completed two experimental sessions corresponding to two music conditions: calming and vexing. Within each session, participants performed a sequence of n-back working-memory tasks with fixed task difficulty per block. Each participant completed a total of 32 blocks, consisting of 16 blocks under calming music and 16 blocks under vexing music. Each block contained 22 trials of the same task type.

Two workload levels were examined: a low-workload 1-back task and a high-workload 3-back task. The 1-back task requires participants to compare the current stimulus with the immediately preceding one, while the 3-back task requires comparison with the stimulus presented three trials earlier, imposing substantially greater working-memory demand. This experimental design enables direct comparison of physiological and behavioral responses across both task difficulty and music condition within the same individuals.

## 2.3 Modalities

Multiple data modalities were recorded concurrently during task performance. Behavioral data include trial-level information such as response time, correctness, task type, and music condition. Physiological signals were collected using two wearable sensing platforms. The Empatica device provided measurements such as electrodermal activity, skin temperature, blood volume pulse, inter-beat intervals, and accelerometry. In parallel, a Biopac system recorded additional physiological signals including electrocardiography, respiration, photoplethysmography, electromyography, and skin conductance.

In addition to peripheral physiological signals, cortical hemodynamic activity was measured using functional near-infrared spectroscopy (fNIRS). The fNIRS data consist of multiple channels measuring changes in oxygenated and deoxygenated hemoglobin concentrations over frontal and prefrontal regions. Together, these modalities capture complementary aspects of cognitive workload, spanning behavioral performance, autonomic nervous system activity, and cortical responses.

## 2.4 Train/Validation/Test Split

To evaluate model generalization across individuals, the data were split at the subject level rather than at the trial level. Three participants were assigned to the training set, one participant to the validation set, and one participant to the test set. This subject-wise split prevents information leakage across sets and ensures that models are evaluated on data from individuals not seen during training. Such a splitting strategy is particularly important in physiological modeling, where subject-specific patterns can otherwise lead to overly optimistic performance estimates that do not reflect true generalization capability.

# 3 Preprocessing and Feature Engineering

## 3.1 Preprocessing by Modality

*3.1.1 Behavioral.* Behavioral data were processed at the trial level and include task type (1-back vs. 3-back), response time, correctness, and music condition. Categorical variables were encoded using one-hot representations where appropriate, and trials that did not correspond to valid n-back task responses were removed. Response times were retained as continuous variables, as they provide a direct behavioral measure of task difficulty and cognitive load. All behavioral preprocessing steps were applied consistently across training, validation, and test sets after the subject-wise split.

*3.1.2 Empatica.* Physiological signals collected via the Empatica device include electrodermal activity, skin temperature, blood volume pulse, inter-beat intervals, and accelerometry. Prior to temporal aggregation, each high-frequency signal was denoised using a Butterworth filter to attenuate high-frequency noise and motion-related artifacts. This filtering step was essential to prevent transient, high-magnitude noise spikes in the raw sampling rate from disproportionately influencing the trial-level mean values after downsampling.

After filtering, signals were resampled or aggregated to a uniform sampling rate of 1 Hz to facilitate alignment with behavioral events and other modalities. For each trial, summary statistics including the mean and standard deviation were computed for each signal. This aggregation reduces residual high-frequency variability while preserving information relevant to sustained physiological responses associated with task performance and arousal [1, 5]. Device-specific characteristics and sampling constraints follow the Empatica E4 specifications [6].

*3.1.3 Biopac.* Biopac recordings provide additional peripheral physiological measures, including electrocardiography, respiration, photoplethysmography, electromyography, and skin conductance. Similar to the Empatica signals, Biopac data were downsampled or aggregated to 1 Hz and summarized on a per-trial basis using mean and standard deviation features. This preprocessing ensures temporal consistency across modalities and yields compact representations suitable for supervised learning while mitigating the impact of sensor noise and motion artifacts.

*3.1.4 fNIRS.* Functional near-infrared spectroscopy data consist of multichannel measurements of oxygenated and deoxygenated hemoglobin concentrations. Raw fNIRS signals were first band-pass filtered in the range of 0.01–0.1 Hz to suppress slow baseline drift and high-frequency physiological noise, consistent with standard preprocessing practices [7, 11]. A subset of channels corresponding to frontal and prefrontal regions was selected to reduce dimensionality and focus on areas relevant to working-memory processes.

Because hemodynamic responses evolve more slowly than behavioral events, fNIRS features were aligned to the trial structure by aggregating signals over the duration of each block. For each selected channel, the mean, standard deviation, and temporal slope were computed, capturing both magnitude and trend information within a block. These features provide a compact summary of cortical activity while accounting for the delayed nature of the hemodynamic response.

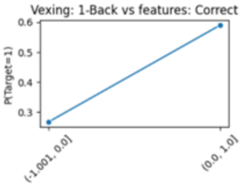

**Figure 1: Probability of the 1-back task conditioned on behavioral correctness under the vexing condition, i.e., $P($1-back $\mid$ `Correct`$)$.**

## 3.2 Feature Engineering Procedure

Following preprocessing, feature engineering was performed to reduce dimensionality and identify informative predictors for each classification task. For the training set only, probability plots were constructed by binning feature values and examining their relationship to the target labels. This analysis provided intuition about potential decision boundaries and highlighted features with discriminative value. (Full probability-plot grids are provided in the appendix.)

For example, under the vexing condition, the probability plot for the behavioral feature `Correct` (Figure 1) indicates a clear separation between workload levels: when responses were incorrect (`Correct=0`), the probability that a trial came from the 1-back task was approximately 0.30, whereas when responses were correct (`Correct=1`), the probability that a trial came from the 1-back task was approximately 0.60. This qualitative separation suggests that simple decision boundaries based on behavioral outcomes can help distinguish lower from higher workload.

In addition, we computed Pearson correlations between each feature and the target label using *training data only*, and used correlation thresholding as a lightweight feature selection step. Importantly, both the probability-plot inspection and correlation-based feature selection were performed exclusively on the training set to prevent information leakage into the validation or test sets. The resulting reduced feature set was then applied unchanged to the validation and test data, ensuring that all reported performance reflects genuine generalization rather than artifacts of feature selection.

*Correlation-selected feature sets.* Because the calming and vexing subsets exhibited different correlation magnitudes, we used condition-specific thresholds for the n-back workload prediction task: $|r| > 0.05$ for calming and $|r| > 0.03$ for vexing. For music-condition prediction, we applied $|r| > 0.05$ with respect to the music label. The selected features and their correlation coefficients are shown in Tables 1, 2, and 3.

A consistent pattern observed during this analysis was that behavioral features (e.g., `Correct`, `Response_Time`) showed the strongest association with the n-back workload labels, while correlations with the music-condition label were weaker overall—particularly for behavioral outcomes such as response time and correctness—suggesting that music effects, when present, are expressed more in physiology than in task performance.

**Table 1: Calming condition: features selected by correlation with the n-back workload label (threshold $|r| > 0.05$, training set only).**

| Feature | Correlation ($r$) |
|---|---:|
| Response_Time | -0.107886 |
| Correct | 0.246588 |
| HR_std | 0.124619 |
| RESP_mean | -0.059623 |

**Table 2: Vexing condition: features selected by correlation with the n-back workload label (threshold $|r| > 0.03$, training set only).**

| Feature | Correlation ($r$) |
|---|---:|
| Response_Time | -0.123890 |
| Correct | 0.286173 |
| HR_std | 0.118374 |
| ACC_std | -0.051320 |
| EDA_std_biopac | 0.030480 |
| EMG_std | 0.049395 |
| RESP_std | 0.142696 |
| SKT_std | 0.037925 |

**Table 3: Music-condition prediction: features selected by correlation with the music label (threshold $|r| > 0.05$, training set only).**

| Feature | Correlation ($r$) |
|---|---:|
| Response_Time | -0.094664 |
| EDA_mean_empatica | -0.102751 |
| EDA_std_empatica | -0.097433 |
| IBI_std | -0.082228 |
| ACC_mean | -0.214605 |
| EDA_mean_biopac | -0.127050 |
| EDA_std_biopac | -0.091782 |
| EMG_std | 0.264050 |
| RESP_mean | -0.179511 |
| RESP_std | -0.075936 |
| SKT_mean | -0.149297 |

## 4 Methods

This section describes the machine learning models evaluated for cognitive workload (n-back) prediction and music-condition classification. All models were trained and evaluated using the same subject-wise train/validation/test split described in Section 3 to avoid leakage across participants and to assess generalization. Model selection and hyperparameter tuning were performed using validation performance only, with the test set held out for final evaluation.

### 4.1 Evaluation Protocol and Metrics

All models were fit on the training subjects, tuned on the validation subject, and evaluated on the held-out test subject. Performance was summarized using accuracy and F1-score. Because degenerate predictors can achieve misleading accuracy by favoring one class, validation selection prioritized F1-score as a balanced metric [12]. Confusion matrices were additionally inspected to diagnose dominant error modes.

### 4.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) was used as a tree-based ensemble classifier to serve as a strong nonlinear baseline [2]. XGBoost constructs an additive model of decision trees, enabling it to capture nonlinear relationships and feature interactions while incorporating regularization to reduce overfitting.

Hyperparameters governing the number of estimators, maximum tree depth, learning rate, subsampling ratios, and L1/L2 regularization were tuned using a grid search. Each candidate configuration was fit on the training set and scored on the validation set using F1-score [12]. The best configuration was selected based on validation performance and then evaluated on the held-out test subject.

### 4.3 K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) was used as a non-parametric baseline classifier that predicts labels based on distance-based similarity in the engineered feature space [4]. Given an input sample, KNN assigns a label by majority vote among its $k$ nearest neighbors under Euclidean distance.

The number of neighbors $k$ directly controls the bias–variance tradeoff: smaller $k$ values emphasize local structure but can overfit, whereas larger $k$ values smooth noise and often generalize better in small-sample, high-dimensional settings. To select $k$, we swept over a predefined range and computed mean validation F1-score for each value [12]. Figure 2 summarizes this sweep and was used to guide the choice of $k$ for downstream evaluation.

### 4.4 Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) was employed to model nonlinear relationships between the engineered multimodal features and the target labels. The MLP operates on tabular inputs and can learn feature interactions that may not be captured by distance-based or rule-based approaches.

The final network architecture is shown in Figure 3. The model uses fully connected layers with ReLU activations and batch normalization to stabilize optimization [8]. Dropout regularization was applied to mitigate overfitting under subject-wise generalization with a limited number of training subjects [13]. The model was trained using the Adam optimizer with binary cross-entropy loss. Regularization strength (e.g., dropout probability) and early stopping criteria were tuned using validation performance only.

### 4.5 RIPPER Rule-Based Classifier

RIPPER was evaluated as a rule-based learning approach capable of forming human-interpretable decision rules by combining multiple features [3]. Unlike distance-based or neural methods, RIPPER learns logical if–then rules that can capture nonlinear decision boundaries without requiring extensive parameterization.

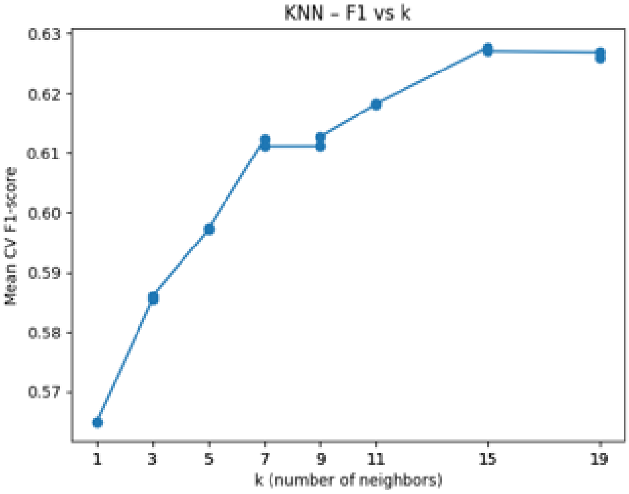

**Figure 2: KNN hyperparameter sweep showing mean validation F1-score as a function of the number of neighbors $k$. Performance increases with larger neighborhoods and saturates around $k \approx 15$–19.**

Preliminary experiments examined the effects of feature scaling and discretization; however, these steps degraded performance, likely by removing meaningful numeric thresholds used by rule induction. Therefore, all RIPPER models were trained using non-scaled and non-discretized features.

Optimization focused on tuning the number of internal cross-validation folds, the minimum description length (MDL), and the random seed. To improve robustness, each parameter configuration was evaluated across multiple random seeds, and the configuration with the best validation performance was selected for final evaluation on the held-out test subject.

## 5 Results: Cognitive Workload Prediction (N-back)

Model performance was evaluated separately for calming and vexing conditions using accuracy and F1-score. Results are reported before and after hyperparameter optimization for all models.

### 5.1 Before Optimization

Table 4 summarizes baseline performance prior to hyperparameter tuning. Overall, models achieved moderate performance, with deep learning and instance-based methods outperforming rule-based approaches. Behavioral features such as response time and correctness provided strong signal for distinguishing between 1-back and 3-back tasks.

RIPPER exhibited near-chance accuracy and extremely low F1-scores prior to feature engineering, indicating that meaningful decision rules could not be learned directly from the raw feature space.

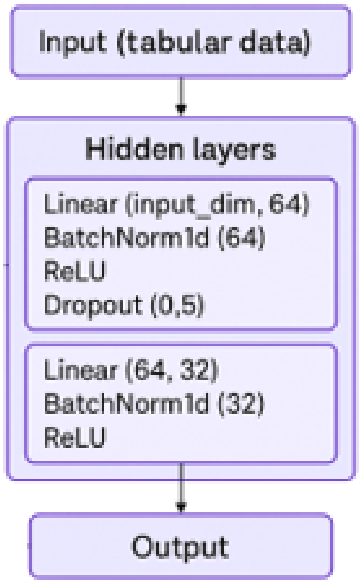

**Figure 3: Multilayer perceptron (MLP) architecture used for tabular multimodal features. The network includes fully connected layers with ReLU activations and batch normalization, with dropout applied to mitigate overfitting.**

**Table 4: N-back prediction performance before optimization.**

| Model | Condition | Accuracy | F1-score |
|---|---|---|---|
| XGBoost [2] | Calming | 57.7% | 63.6% |
| XGBoost | Vexing | 64.8% | 66.8% |
| KNN [4] | Calming | 59.8% | 62.9% |
| KNN | Vexing | 57.0% | 61.7% |
| MLP | Calming | 63.3% | 69.9% |
| MLP | Vexing | 61.3% | 69.9% |
| RIPPER [3] | Calming | 50.0% | 0.0% |
| RIPPER | Vexing | 51.7% | 19.0% |

### 5.2 After Optimization

After feature engineering and hyperparameter tuning, performance improved for most models, particularly RIPPER, which benefited substantially from reduced dimensionality and removal of noisy features. Table 5 reports the best validation performance achieved for each model.

Notably, RIPPER showed the largest relative improvement, with F1-score increasing from near zero to approximately 70% after feature engineering. This confirms that rule-based learners are highly sensitive to feature quality and dimensionality.

**Table 5: N-back prediction performance after optimization.**

| Model | Condition | Accuracy | F1-score |
|---|---|---|---|
| XGBoost | Calming | 66.8% | 69.9% |
| XGBoost | Vexing | 54.3% | 64.1% |
| KNN | Calming | 60.0% | 63.2% |
| KNN | Vexing | 59.0% | 60.2% |
| MLP | Calming | 63.2% | 70.4% |
| MLP | Vexing | 61.1% | 69.9% |
| RIPPER | Calming | 63.6% | 70.1% |
| RIPPER | Vexing | 62.2% | 68.3% |

### 5.3 MLP Training Dynamics

Figures 4–5 illustrate training and validation dynamics for the MLP under both music conditions. The network architecture incorporates dropout [13] and batch normalization [8] to mitigate overfitting, which is particularly important given the small number of subjects.

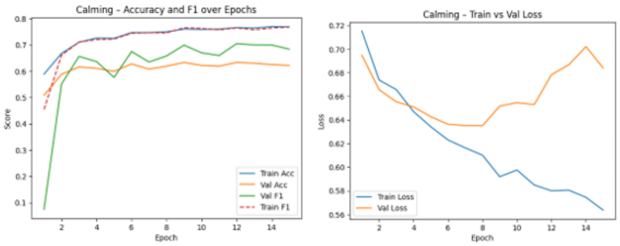

**Figure 4: MLP training curves under calming condition.**

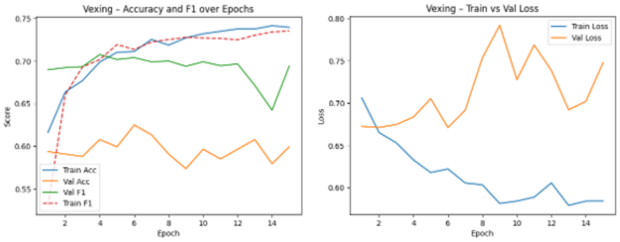

**Figure 5: MLP training curves under vexing condition.**

Under both conditions, training loss decreases steadily while validation loss plateaus or increases slightly, indicating mild overfitting despite regularization. Validation F1 stabilizes early, suggesting that additional epochs do not substantially improve generalization.

### 5.4 Interpretation

Across models, cognitive workload prediction consistently outperformed music condition prediction, reflecting the strong behavioral and physiological signatures associated with task difficulty. Deep models and RIPPER benefited most from feature engineering, while XGBoost showed mixed sensitivity to dimensionality reduction. The results highlight the challenges of learning from high-dimensional physiological data with limited subject counts and emphasize the importance of careful feature selection and subject-wise evaluation.

—

## 6 Results: Music Condition Prediction

### 6.1 Before Optimization

Prior to optimization, all models exhibited poor performance when predicting music condition. In several cases, classifiers collapsed to predicting a single class, yielding misleading accuracy values and low or zero F1-scores. This behavior suggests that the available features contain limited discriminative information for music condition alone.

### 6.2 After Optimization

Table 6 summarizes performance after feature engineering and tuning.

**Table 6: Music condition prediction performance after optimization.**

| Model | Accuracy | F1-score |
|---|---|---|
| XGBoost | 49.4% | 66.2% |
| KNN | 61.0% | 70.1% |
| MLP | 51.8% | 61.0% |
| RIPPER | 50.7% | 67.0% |

Although KNN achieved the highest accuracy and F1-score, performance remained only marginally above chance, and no model demonstrated robust generalization.

### 6.3 Interpretation

The weak performance across all models indicates that music condition labels contain substantially less learnable signal than cognitive workload labels. While background music may influence physiological arousal, its effects are highly subjective and confounded by task difficulty. Additionally, models were trained on data pooled across both workload levels, introducing uncontrolled variability that likely obscured music-specific patterns. These findings suggest that limitations arise from the information content of the features rather than from the choice of classifier or optimization strategy.

## 7 Discussion

This study investigated whether multimodal behavioral, physiological, and neuroimaging signals can be used to predict (1) cognitive workload during an n-back task and (2) background music condition. Across all experiments, models consistently achieved higher and more stable performance when predicting cognitive workload than when predicting music condition, indicating a fundamental difference in the strength and structure of the learnable signal present in the data [10].

### 7.1 Cognitive Workload Prediction

For n-back workload prediction, all models performed above chance, with the strongest results observed for the optimized deep neural network (MLP), RIPPER, and KNN models. Behavioral features, particularly response time and correctness, exhibited the highest correlation with the workload labels, which aligns with the task design: increasing n-back difficulty directly impacts task performance and reaction time. This relationship is consistent with established findings linking task demand, arousal, and performance [14]. It was also reflected in the probability plots, where behavioral correctness produced simple and interpretable decision boundaries separating 1-back and 3-back trials.

Feature engineering generally improved generalization for models sensitive to irrelevant or redundant features, particularly RIPPER. The substantial improvement in RIPPER performance after feature selection supports the hypothesis that rule-based models benefit from a reduced, interpretable feature space where weakly informative variables are removed. In contrast, XGBoost showed limited gains from feature engineering and optimization, suggesting that the engineered feature subset may have restricted the ensemble's ability to exploit higher-order feature interactions.

The MLP models exhibited signs of overfitting, particularly under the vexing condition, as evidenced by diverging training and validation losses. However, regularization via dropout and batch normalization stabilized training and produced competitive validation F1-scores. These results reflect the challenges of applying high-capacity models to small, high-dimensional datasets with strong subject-level variability, which is a common concern in physiological and neuroimaging modeling [7, 11].

## 7.2 Music Condition Prediction

In contrast to workload prediction, music condition classification proved substantially more difficult. Correlation analysis revealed that most features—especially behavioral features—had weak association with the music-condition labels. Even physiological features exhibited relatively small correlation magnitudes, indicating that the signal differentiating calming from vexing music is subtle and noisy in this dataset, consistent with the dataset's pilot-scale design and reported variability [10].

This weak feature–label relationship is reflected in model performance. With the exception of KNN, all models performed near chance both before and after optimization. XGBoost and RIPPER frequently collapsed to predicting a single class, resulting in misleadingly high recall but poor overall discrimination. The MLP models showed inconsistent behavior across validation and test sets, suggesting sensitivity to subject-specific physiological patterns rather than robust music-related effects.

The comparatively better performance of KNN may be explained by its instance-based nature, which allows it to exploit local similarities between trials without imposing a strong parametric structure. However, even KNN performance remained modest, reinforcing the conclusion that music condition is weakly encoded in the available features.

## 7.3 Implications of Feature Correlation Patterns

The observed discrepancy between workload and music-condition prediction aligns closely with the feature correlation analysis. Behavioral features demonstrated strong correlation with workload but minimal correlation with music condition, while physiological features showed only modest association with music labels. This suggests that cognitive workload manifests through consistent and task-driven behavioral and physiological responses, whereas music perception is more subjective and individualized [10].

Additionally, music-condition models were trained on data pooled across both 1-back and 3-back tasks. This introduces an unobserved confounding variable—task difficulty—that likely increases intra-class variability and obscures music-related effects. Combined with the small number of participants, this limits the ability of supervised models to learn stable decision boundaries for music condition.

## 7.4 Limitations and Future Directions

Several limitations should be considered when interpreting these results. First, the dataset includes only five participants, which restricts statistical power and exacerbates subject-specific effects [10]. Second, fNIRS features were aggregated at the block level to accommodate slow hemodynamic responses, potentially reducing temporal sensitivity to music-induced changes; more generally, fNIRS preprocessing and filtering choices can materially affect downstream inferences [7, 11]. Third, the use of simple correlation-based feature selection may overlook nonlinear relationships between features and labels.

Future work should explore larger participant cohorts, task-stratified music-condition modeling, and advanced representation learning techniques capable of capturing subtle physiological patterns. Incorporating temporal models or subject-adaptive approaches may further improve generalization. Despite these limitations, the present results provide clear evidence that cognitive workload is more reliably predictable than music condition using multimodal physiological and behavioral data.

# 8 Conclusion

This work examined the feasibility of predicting cognitive workload and background music condition from a multimodal dataset combining behavioral, physiological, and neuroimaging signals. Using a subject-wise evaluation protocol, we compared multiple classification approaches before and after feature engineering and optimization.

Across all experiments, cognitive workload (1-back vs. 3-back) was consistently more predictable than music condition. Behavioral features such as response time and correctness exhibited strong association with workload and enabled models to learn stable and interpretable decision boundaries. Feature engineering substantially improved generalization for models sensitive to irrelevant or redundant inputs, particularly the rule-based RIPPER classifier. Deep learning and instance-based methods also achieved competitive performance, though they exhibited sensitivity to subject-specific variability due to the small sample size.

In contrast, music condition prediction proved challenging for all models. Correlation analysis revealed weak associations between the available features and music labels, and classification performance remained near chance even after optimization. These findings suggest that, within the scope of this dataset, music-induced effects are subtler, more subjective, and more difficult to capture reliably than workload-related effects, especially when data from different task difficulties are pooled.

Overall, the results indicate that model performance is primarily limited by the information content of the features rather than the choice of classifier or optimization strategy. While multimodal physiological data provide a strong signal for cognitive workload estimation, predicting music condition likely requires larger participant cohorts, task-specific modeling, or alternative feature representations. This study highlights both the promise and the limitations of multimodal machine learning approaches for cognitive state inference in small-sample experimental settings.

## 9 GitHub Repository

https://github.com/owenp33/workload-complexity-and-music-detection

## References

[1] Wolfram Boucsein. 2012. *Electrodermal Activity.* Springer. doi:10.1007/978-1-4614-1126-0

[2] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 785–794. doi:10.1145/2939672.2939785

[3] William W. Cohen. 1995. Fast Effective Rule Induction. In *Machine Learning: Proceedings of the Twelfth International Conference (ICML'95).* 115–123. https://sci2s.ugr.es/keel/pdf/algorithm/congreso/ml-95-ripper.pdf

[4] Thomas M. Cover and Peter E. Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27. doi:10.1109/TIT.1967.1053964

[5] Mohamed Elgendi. 2012. On the Analysis of Fingertip Photoplethysmogram Signals. *Current Cardiology Reviews* 8, 1 (2012), 14–25. doi:10.2174/157340312801215782

[6] Empatica. 2020. E4 Wristband User Manual (Rev. 2.0). PDF manual. https://www.utwente.nl/en/bmslab/infohub/um-16-e4-usermanual-rev.2.0-20201020.pdf

[7] Lia M. Hocke, Ibironke K. Oni, Chelsea C. Duszynski, Alexandria V. Corrigan, Benjamin D. Frederick, and Joanne F. Dunn. 2018. Automated Processing of fNIRS Data—A Visual Guide to the Pitfalls and Consequences. *Algorithms* 11, 5 (2018), 67. doi:10.3390/a11050067

[8] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv preprint. arXiv:1502.03167 [cs.LG] https://arxiv.org/abs/1502.03167

[9] Saman Khazaei et al. 2025. A Multimodal Dataset for Investigating Working Memory in Presence of Music. PhysioNet. https://physionet.org/content/multimodal-nback-music/ Version 1.0.0.

[10] Saman Khazaei, Srinidhi Parshi, Samiul Alam, Md. Rafiul Amin, and Rose T. Faghih. 2024. A multimodal dataset for investigating working memory in presence of music: a pilot study. *Frontiers in Neuroscience* (2024). doi:10.3389/fnins.2024.1406814

[11] Paola Pinti, Felix Scholkmann, Andrew Hamilton, Paul Burgess, and Ilias Tachtsidis. 2019. Current Status and Issues Regarding Pre-processing of fNIRS Neuroimaging Data: An Investigation of Diverse Signal Filtering Methods Within a General Linear Model Framework. *Frontiers in Human Neuroscience* 12 (2019), 505. doi:10.3389/fnhum.2018.00505

[12] scikit-learn developers. 2025. sklearn.metrics.f1_score documentation. Online documentation. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15 (2014), 1929–1958. https://www.jmlr.org/papers/v15/srivastava14a.html

[14] Robert M. Yerkes and John D. Dodson. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology* 18, 5 (1908), 459–482. doi:10.1002/cne.920180503