Music-Induced Emotion Prediction: A Regression-Based Analysis of Physiological and Acoustic Data

Abstract – Understanding how physiological responses and musical features interact to influence emotional perception is a growing area of research in affective computing and music therapy. Although current research has investigated music emotion recognition (MER) through either audio features or physiological signals, no study has yet represented their joint effect on valence and arousal prediction. The non-linear, complex nature of human emotions limits the accuracy of traditional regression models.

This study examines physiological signals and acoustic features using the HKU956 multimodal dataset. Support Vector Machine **(SVM)** and Random Forest **(RF)** regression models are applied, and performance is evaluated using R^2 and RMSE. Although predictive accuracy remains limited, the findings provide deeper insight into music-induced emotions and physiological effects, contributing to personalized music therapy and adaptive recommendation systems. Future research should focus on more sophisticated feature engineering and deep learning methods to enhance prediction accuracy.

Keywords – Support Vector Machine Regression, Random Forest Regression, Emotion prediction, Music Analysis

I. INTRODUCTION

Music has long been recognized for its strong impact on human emotions, cognition, and mental health. From ancient therapeutic applications to personalized music in the modern era, the effect of music on emotions has been a central research focus in affective computing. Recent advancements in artificial intelligence (AI) and machine learning (ML) have allowed researchers to model the intricate interdependencies among musical features, physiological responses, and emotional experiences, particularly within the valence-arousal model of emotion [8]. Earlier research highlights the importance of timbre, rhythm, and spectral contrast in musical emotional perception [7], while heart rate (HR), electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature (TEMP) serve as physiological indicators of emotional responses [1]. However, limited research has explored how these musical and physiological variables interact to shape perceived emotional experiences.

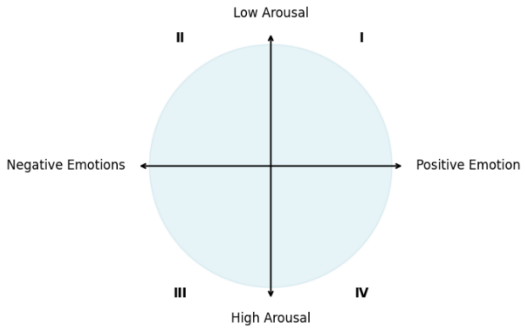
This study addresses this gap by examining the relationship between physiological responses and musical features. We examine the relationship between physiological responses and musical features to model emotional perception using ML techniques. SVM and RF regression models are applied for valence and arousal prediction [3]. Pre-processing techniques such as normalization and feature selection are implemented, and model performance is evaluated using R^2 and RMSE. While reasonable prediction accuracy is achieved, the study offers insights into music-evoked emotional responses, with implications for personalized music therapy and adaptive recommender systems.

Objectives of the Study:

- To examine the connection between physiological signals and musical features in order to estimate emotional reactions.
- To ascertain how valence and arousal are connected with heart rate (HR), electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature (TEMP).
- To evaluate the performance of acoustic features such as MFCCs, spectral contrast, and rhythm features for emotion modeling.
- To utilize supervised learning models for investigating the predictive potential of physiological and audio data.
- To contribute to improving individualized music therapy and adaptive music recommender systems.

II. LITERATURE REVIEW

Music emotion recognition (MER) has been a subject of great interest, especially in music therapy and affective computing. A few studies have investigated the interplay between audio features, physiological signals, and emotional experiences, and their predictive power for valence and arousal [5]. It is crucial to understand this interplay to enhance personalized music recommendation systems and therapies.



Physiological Responses and Emotion Prediction

Physiological signals such as heart rate (HR), electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature (TEMP) have been extensively researched as biomarkers for emotional states [1]. The DEAP dataset, a benchmark for emotion recognition, showed the effectiveness of physiological signals in recognizing arousal and valence levels [1]. Likewise, Kim & André (2006) pointed out that machine learning coupled with physiological signals can enhance emotion recognition accuracy, and it is a strong basis for using these features in MER models [4]. The above findings provide justification for the utilization of multimodal feature extraction in this research; hence, both musical and physiological dimensions are covered.

Musical Feature Engineering and Emotional Perception

Emotional responses to music are founded on timbre, pitch, rhythm, loudness, and harmonic characteristics [2]. Panda et al. (2018) proposed new musical texture and expression approaches to enhance MER performance by using higher-level acoustic features beyond the conventional MFCC-based approach [2]. Multiple studies [2,3,4] confirm that integrating spectral and rhythmic features enhances MER, supporting their inclusion in this study. These results corroborate the utilization of spectral contrast, MFCC, and chroma features, which in this study were discovered to be highly correlated with the valence and arousal dimensions.

Machine Learning Models for Music Emotion Prediction

Traditional supervised learning algorithms like Support Vector Machines (SVM) and Random Forest (RF) have been extensively used in MER tasks [3]. The capacity of these models to process complicated, high-dimensional data renders them ideal for investigating the interaction between physiology and music. Lai et al. (2023) discussed the use of supervised models in personalized music therapy and stressed their importance in adjusting in real-time [5]. Moreover, valence/arousal regression model research confirms the efficacy of ordered feature representation in machine learning model training [3]. Based on these observations, this research utilizes SVM and RF regression models to investigate the predictability of emotional responses from musical and physiological features.

Justification for Project Approach

This research builds upon previous work by integrating physiological signals and acoustic feature extraction within a machine learning framework. Unlike previous work that focused on either physiological or musical features, this paper explores their interaction in predicting emotion. Prior work has also given more emphasis to deep learning methods, whereas this research employs supervised regression models to provide interpretability and computational tractability. The strategy of the project is consistent with current research and expands upon it by explaining how multimodal data helps improve emotional perception, ultimately promoting personalized music therapy and affective computing applications.

By combining results from previous research, the present work contributes to enhancing emotion prediction models and practical applications in real-world music therapy and adaptive recommendation systems.

III. DATA MANAGEMENT

Dataset Acquisition and Description

This study uses the HKU956 multimodal dataset, a large and meticulously curated dataset that was expressly created for investigating the interconnections between musical features, physiological responses, and emotional states [9]. The physiological measurements comprise heart rate (HR), electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature (TEMP), all known biomarkers of emotional states [1]. The database contains audio samples and song metadata gathered systematically and saved in song metadata files. The structured database makes it possible to perform an in-depth analysis of physiology-emotion relationships and their alignment with musical features.

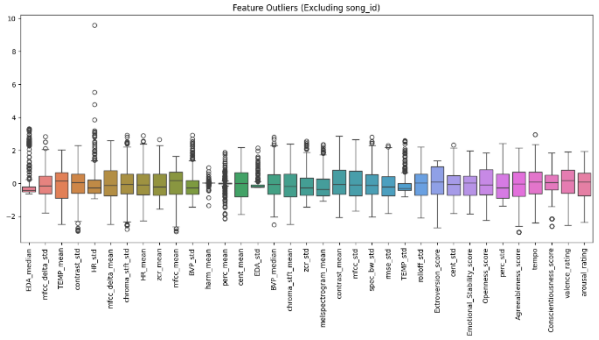
Musical and Physiological Feature Extraction

Feature extraction operations were separately designed for music and physiological data to acquire emotionally relevant features. Music features were acquired through the Librosa library, which is a well-known audio processing library in music information retrieval (MIR) applications [2]. Acoustic features like Mel-frequency cepstral coefficients (MFCCs), chroma features, spectral contrast, tempo, spectral centroid, spectral roll-off, zero-crossing rate (ZCR), and rhythm-based features were extracted from all audio files. These acoustic characteristics are firmly grounded in their role in feeling perception and music analysis, providing detailed descriptions of each song's acoustic characteristics.

Physiological signals were extracted through the computation of statistical values such as mean, median, and standard deviation on all physiological signals (HR, EDA, BVP, TEMP), representing corresponding changes in the emotional states of listeners [4]. Based on these measurements, the dataset retains intricate physiological dynamics aligned with emotion changes occurring during music listening.

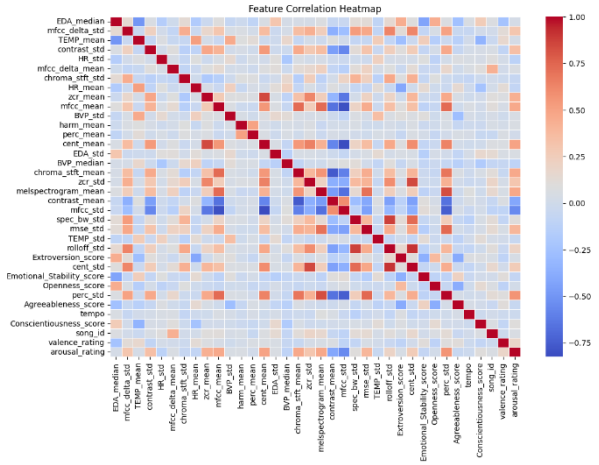
Data Pre-processing and Cleaning

Data pre-processing was necessary to enhance data quality and consistency before analysis. Physiological data were rigorously quality-checked, including participant-wise median imputation for the treatment of missing values to preserve data integrity and continuity [1]. Physiological outliers were identified and processed through Z-score analysis to reduce potential biases due to erroneous measurements [4]. Audio metadata were analysed to eliminate duplicates, preserving the uniqueness and authenticity of musical entries in the database [9]. Additionally, all feature sets were normalized using uniform scaling techniques, making the data ready for successful integration into the subsequent modeling processes [2].



Feature Selection and Dimensionality Reduction

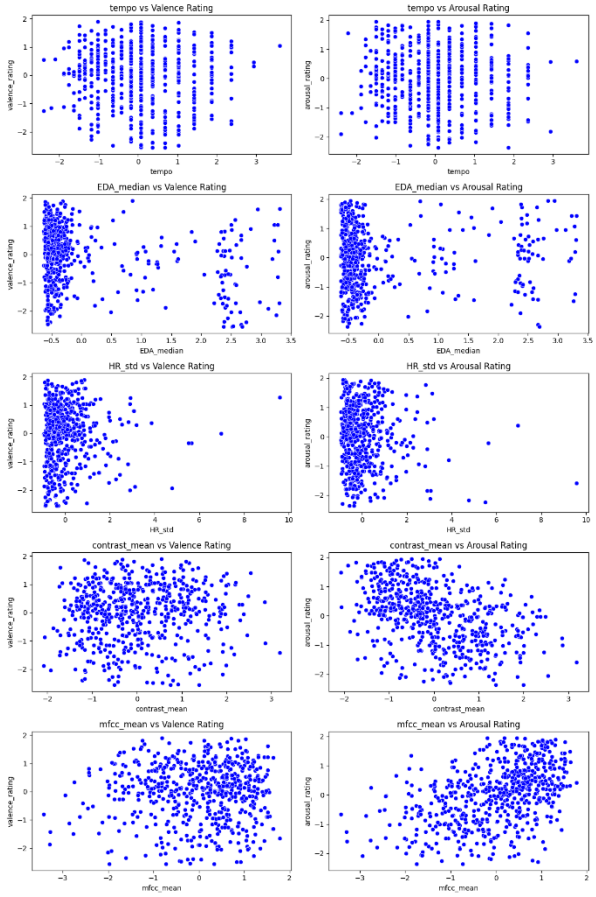
For better model interpretability and predictive performance, feature selection methods were applied to identify the most relevant features. Random Forest ranking, correlation analysis, and recursive feature elimination (RFE) were used to retain the most relevant predictors while eliminating redundant features. [5].



Feature Engineering and Dataset Merging

The next step involved the rigorous integration of elicited musical and physiological features into a unified multimodal dataset. Derived features of tempo variability, variations in spectral contrasts, and certain EDA level changes were computed to

denote temporal dynamics and enhance emotional predictability accuracy [3]. Strategically derived features were intended to enhance the dataset's emotional interpretability. Secondly, the music and physiological features were paired with corresponding matching ratings for emotions and song labels to construct a consistent, congruent feature set appropriate for solid multimodal analysis.



Interaction Features

Interaction features were included to address the dynamic interaction between musical and physiological features. Since emotional responses to music may be controlled by the interaction among tempo, spectral contrast, and physiological arousal markers (e.g., EDA and HRV), interaction terms such as tempo × EDA and spectral contrast × HRV variability were conceived. These interaction features were designed to extract the cumulative effects of music and physiological variables on affect experience, taking into account nonlinear interdependencies that might be disregarded by single features.

Through these meticulous data management processes, including large-scale exploratory analysis, this study ensures a well-structured, high-quality multimodal dataset in anticipation of rigorous analytical processes. Meticulous feature integration and tuning enable a robust investigation of the emotional impact of music, leading to advances in personalized music therapy and adaptive affect recognition systems.

IV. METHODOLOGY

This section discusses the machine learning approach applied to studying the relationship among physiological responses, musical features, and emotion. Support Vector Machine (SVM) regression and Random Forest (RF) regression are applied in predicting valence and arousal ratings in the study.

The nature of human emotional responses is complex, regression models were preferred over classification models to maintain the continuous nature of the emotional axes [5]. The input features included MFCCs, spectral contrast, tempo, chroma features, and physiological statistics (HR, EDA, BVP, TEMP), all of which were normalized prior to model training [2].

Support Vector Machine (SVM) Regression

SVM regression was chosen for its ability to model non-linear relationships, using an RBF kernel for optimization [4]. The model attempts to find a function f(x) that best fits the data by minimizing the error within a defined tolerance margin [4]. The function is represented as:

$$f(x) = w^T x + b$$

where ‘w’ is the weight vector, ‘x’ represents the feature set, and b is the bias term. To improve performance, a Radial Basis Function **(RBF)** kernel was used, which allows SVM to model complex relationships by transforming the feature space.

The epsilon-insensitive loss function used in SVM regression ensures that small deviations from actual values do not contribute to the loss, defined as:

$$L(y, f(x)) = \max(0, |y - f(x)| - \varepsilon)$$

where y is the actual valence or arousal score, and ϵ defines a threshold below which errors are ignored [4].

Random Forest (RF) Regression

Random Forest regression was employed due to its capability of handling feature interactions and reducing variance through ensemble learning [2]. The model learns several decision trees during training and averages their predictions, mitigating overfitting and improving generalizability. Each tree is trained on a subset of training examples and features, enhancing robustness to noise.

Mathematically, Random Forest regression predicts an output by averaging predictions from individual decision trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

where $\mathbf{T_i(x)}$ is the **prediction from each tree**, and **N** is the **number of trees in the forest**.

Feature importance was analysed using the Mean Decrease in Impurity (Gini Importance) method, allowing for the identification of the most influential features in determining valence and arousal. This analysis aided feature selection, promoting model efficiency.

Training Process and Model Evaluation

The dataset was split into **70% training, 15% validation, and 15% testing**, ensuring a balanced approach to training while allowing model performance evaluation on unseen data. **Hyperparameter tuning** was performed using **grid search cross-validation** to optimize the parameters of both models:

- **SVM Parameters**: Regularization parameter (C), kernel coefficient (gamma), and epsilon margin.
- **RF Parameters**: Number of trees, maximum depth, and minimum samples per split.

The models were evaluated using **Root Mean Squared Error (RMSE)** and **coefficient of determination (R^2)**, which quantify predictive accuracy and variance explanation, respectively:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where yiy_i are actual values, y^i\hat{y}_i are predicted values, and y¯\bar{y} is the mean of the actual values.

Justification of Model Selection

RF regression and SVM were selected based on their capability to handle high-dimensional feature spaces and their strong performance in previous emotion recognition studies [5]. SVM is particularly effective at representing complex patterns through kernel transformations, while RF offers interpretability and robustness against overfitting by aggregating multiple decision trees. Their complementary nature ensures a balanced strategy for predicting emotional responses from multimodal inputs.

By leveraging these methodologies, this study ensures a robust predictive framework for exploring the intricate interactions between **physiological responses, musical features, and emotional states**, contributing to advancements in **music emotion recognition and personalized therapy applications**.

V. RESULTS AND ANALYSIS

This section presents the findings of model training, hyperparameter tuning, and interaction feature evaluation, followed by a detailed discussion of the results, feature relationships, and comparison with existing literature.

The models were trained using a **70-15-15 split** for training, validation, and testing. The following table summarizes the initial performance metrics of Support Vector Machine (SVM) and Random Forest (RF) regression:

Model	Target	MSE	R^2 Score
Random Forest	Valence	0.8202	0.1246
Random Forest	Arousal	0.6578	0.36
SVM	Valence	0.8169	0.1281
SVM	Arousal	0.6356	0.3815

- **SVM regression** achieved the **best R^2 for Arousal (0.3815)**, suggesting a **stronger ability to predict high-energy emotional responses**.
- **Random Forest regression** slightly outperformed **SVM for Valence prediction** (**R^2 = 0.1246 vs. 0.1281 for SVM**), though both models **struggled to achieve high accuracy for Valence** [1].

Hyperparameter Tuning

To improve performance, **grid search cross-validation** was applied:

- **Optimized Model Parameters**:

```
RF Parameters:
  - Valence:
- max_depth = None
- min_samples_split = 2
- n_estimators = 100
  - Arousal:
- max_depth = 10
- min_samples_split = 2
- n_estimators = 150

SVM Parameters:
  - Valence:
- C = 1
- gamma = 'scale'
- kernel = 'rbf'
  - Arousal:
- C = 1
- gamma = 'scale'
- kernel = 'rbf'
```

Post-optimization, the models yielded minor improvements, but the overall R^2 values remained moderate, indicating challenges in capturing variance in valence prediction.

Model	Target	MSE	R^2 Score
Optimized RF	Valence	0.8112	0.1342
Optimized RF	Arousal	0.6431	0.3742
Optimized SVM	Valence	0.8169	0.1281
Optimized SVM	Arousal	0.6356	0.3815

Evaluating Interaction Features and Model Performance

Since both pairings of audio features and physiological responses drive emotional ratings, interaction terms were added to capture more intricate dependencies between musical and physiological variables. The interaction terms allowed for explanations of how musical dynamics (e.g., tempo, spectral contrast) impact physiological signals (e.g., heart rate variability, EDA levels), which in turn influence valence and arousal ratings. The best-performing models with interaction terms are presented below.

Model	Target	MSE	R^2 Score
RF with Interactions	Valence	0.8087	0.1369
RF with Interactions	Arousal	0.6444	0.373
SVM with Interactions	Valence	0.9055	0.0335
SVM with Interactions	Arousal	0.7865	0.2347

- Including interaction features likely improved RF performance in valence (R^2 = 0.1369) but worsened SVM performance.
- Although physiological features were highly correlated with arousal, their correlations with valence were weaker.
- This aligns with the literature, which states that valence is more contextual and subjective in nature, whereas arousal is more strongly linked to physiological responses [1].

Discussion of Model Performance and Feature Relationships

1. Feature Importance Analysis

Top features influencing Arousal prediction:

- Electrodermal Activity (EDA Median)
- Heart Rate Variability (HR Std)
- Spectral Contrast (Musical Feature)
- Chroma Mean (Tonality Indicator)

Tempo and loudness had moderate effects on Arousal but were weak predictors of Valence. This supports findings that physiological measures correlate more with arousal than valence [3].

2. Model Struggles and Constraints

- Low R^2 values for valence suggest that subjective impressions cannot be well-fitted using acoustic or physiological features alone.
- Arousal is more reliable, as it possesses stronger physiological roots, supported by prior research involving autonomic nervous system activation and music-induced arousal.
- Including interaction features somewhat improved performance but did not dramatically enhance predictability, highlighting the subtlety and complexity of emotional responses.

Comparison with Existing Research

- Prior research using deep learning models (e.g., RNNs, CNNs) has reported higher accuracies, suggesting that traditional regression models may be insufficient in

capturing the non-linear nature of music-emotion mappings [2].

- The findings confirm earlier research indicating that physiological signals, particularly EDA and HR, are more predictive of arousal than valence [1].
- This aligns with the hypothesis that valence is more influenced by cognitive and personal factors, making it harder to generalize in machine learning models [3].

VI. CONCLUSION

The study confirms that physiological cues contribute significantly to the prediction of arousal and supports their use in real-time emotion detection, whereas valence remains difficult to predict due to subject variability. The fusion of music and physiological features provided a more elaborate description of emotional reactions but also highlighted the need for more advanced modeling strategies to capture valence-based nuances.

Future research should focus on multimodal deep learning and real-time adaptation to enhance prediction accuracy, while exploring personalized models to improve emotion-aware applications. Additionally, future research should focus on real-time feedback mechanisms that allow for continuous model tuning based on user experience, making the system more adaptive and practical for personalized music therapy treatment.

VII. REFERENCES

1. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). **DEAP: A database for emotion analysis using physiological signals**. *IEEE Transactions on Affective Computing*, 3(1), 18-31. https://ieeexplore.ieee.org/document/5871728

2. Panda, R., Malheiro, R., & Paiva, R. P. (2018). **Novel audio features for music emotion recognition**. *IEEE Transactions on Affective Computing*. https://ieeexplore.ieee.org/document/8327886

3. Coutinho, E., & Cangelosi, A. (2011). **Musical emotion recognition from audio: A review of audio features**. *Journal of New Music Research*. https://pubmed.ncbi.nlm.nih.gov/21859207/

4. Kim, J., & André, E. (2006). **Emotion recognition using physiological and speech signals in short-term observation**. In *International Workshop on Perception and Interactive Technologies for Speech-Based Systems* (pp. 53-64). Springer. https://link.springer.com/chapter/10.1007/11768029_6

5. Lai, N. Y. Y., Philiastides, M. G., Kawsar, F., & Deligianni, F. (2023). **Towards personalised music-therapy: A neurocomputational modelling perspective**. *arXiv preprint*, arXiv:2305.14364. https://arxiv.org/abs/2305.14364

6. Kwon, C.-Y., Kim, H., & Kim, S.-H. (2024). **The modernization of Oriental music therapy: Five-element music therapy combined with artificial intelligence**. *Healthcare (Basel)*, 12(3), 411. https://www.mdpi.com/2227-9032/12/3/411

7. Nalini, N. J., & Palanivel, S. (2016). **Music emotion recognition: The combined evidence of MFCC and residual phase**. *Egyptian Informatics Journal*, 17(1), 1-10. https://www.sciencedirect.com/science/article/pii/S1110866515000419?via%3Dihub

8. Hu, X., Li, F., & Liu, R. (2022). **Detecting Music-Induced Emotion Based on Acoustic Analysis and Physiological Sensing: A Multimodal Approach**. *Applied Sciences*, 12(18), 9354. https://www.mdpi.com/2076-3417/12/18/9354

9. Hu, X., Li, F., & Liu, R. (2022). **HKU956: A Multimodal Dataset for Analyzing Listeners' Emotion and Physiological Responses Induced by Music**. Available under CC BY-NC 4.0 License. https://www.mdpi.com/2076-3417/12/18/9354