Moodify: A Machine Learning Approach of Harnessing Music to Predict and Influence Emotions

Aditya Kumar (1006300)
Josephine Tan (1006128)
Stefanie Tan Hui Zhen (1006073)
Kripashree Suryaprakash (1006507)
Ng Junhao Marcus (1006118)

April 18, 2024

1 Introduction

In today's world of ubiquitous music streaming, the potential for leveraging music to enhance emotional well-being is immense. Moodify is designed to predict the emotions elicited by music and use this insight to curate transformative playlist experiences.

Core Functionality

Moodify integrates emotion prediction models with a dynamic playlist curation system. The application analyzes music tracks to predict the emotional responses they might evoke in listeners. This prediction is powered by robust datasets, including the DEAM (Database of Emotional Musical Experiences), Spotify dataset, and GTZAN dataset, which provide a rich foundation for training our emotion prediction models.

Significance of Emotion-Based Music Recommendation

Emotion-based music recommendation systems are pivotal because they personalize music suggestions based on the listener's emotions, going beyond traditional genre or artist-based recommendations. These systems enhance user experience by aligning song choices with the listener's mood, acknowledging music's power to influence emotions. This personalized approach deepens the listener's connection with music, making the discovery process more intuitive and emotionally resonant. Such advancements in recommendation technology not only improve satisfaction but also unlock new ways for listeners to explore music that complements their emotional state, enriching the overall listening experience.

2 Dataset and Collection

For this project, we are using multiple datasets related to music, specifically focusing on the DEAM Dataset, Spotify API Dataset and

GTZAN Dataset. These datasets are commonly used in music emotion recognition and music recommendation systems.

MediaEval Database for Emotional Analysis in Music (DEAM) Dataset

The DEAM (MediaEval Database for Emotional Analysis in Music) dataset is an extensive collection used for the analysis of emotions in music. It includes over 1800 songs along with their emotional annotations. The dataset is a compilation from the "Emotion in Music" task at the MediaEval benchmarking campaign from 2013 to 2015. It features both 45-second music excerpts and full songs in MP3 format, with annotations made on these excerpts as well as full tracks. The excerpts are chosen from a random starting point in each song, ensuring diversity in the samples. Emotional annotations are made dynamically during the song and averaged to provide a standard measure of emotional content. The annotations cover aspects of arousal and valence on a continuous scale, resampled to 2Hz to standardize across different computer capabilities. The songs are sourced from royalty-free music platforms like the Free Music Archive (FMA), Jamendo, and the MedleyDB dataset. The emotional annotations and song metadata including title, artist, and genre are provided in CSV format. [1]

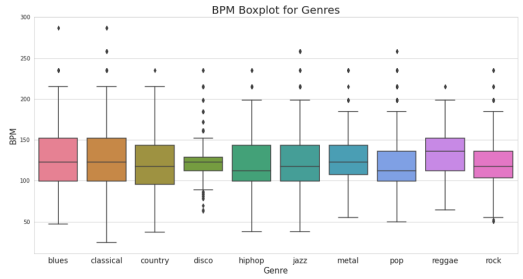
Spotify API Dataset

The Spotify API dataset is a comprehensive collection of Spotify tracks that have been selected manually from Spotify's playlists (essentials or classics for each genre), with each track characterized by various audio features. The dataset has been compiled and cleaned using Spotify's Web API and Python. Presented in CSV format, making it easily accessible and amenable to analysis. It includes features such as track name and song. From the track and song extracted, yt-dlp is used to query for the music and download 45 seconds from each song. [2] This dataset is used to test data to evaluate our models. We classify the genres into a Playlist Distribution as shown in Figure 1.

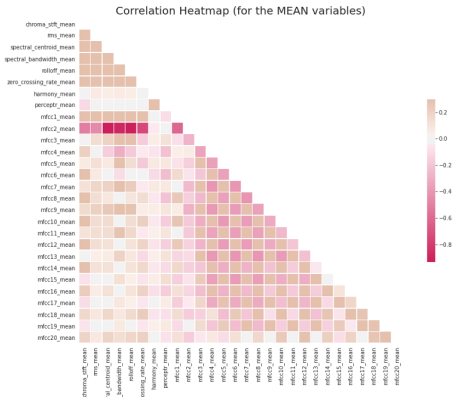
Figure 1: Spotify API Playlist Distribution

GTZAN Dataset

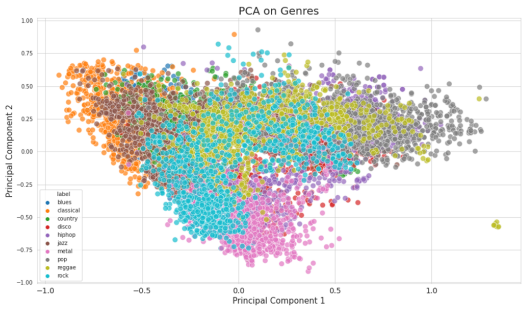
The GTZAN dataset is a widely used public dataset in machine listening research for music genre recognition (MGR). It was collected between 2000-2001 from various sources including personal CDs, radio, and microphone recordings to represent a variety of recording conditions. There are 10 genres with 100 audio files each, all with a length of 30 seconds. This dataset is often referred to as the "MNIST of sounds" for its popularity and usefulness in the field. [3] This dataset enables analysis and exploration of musical characteristics across different genres, facilitating genre prediction based on audio features. This dataset is preferable as we can utilize the audio features to build a recommendation system that suggests tracks based on user emotional preferences based on similarities in musical characteristics. We presented these models in data visualization in Figure 2.



(a) BPM Boxplot for Music Genres from GTZAN Dataset



(b) Correlation Heatmap (for the MEAN variables) from GTZAN Dataset



(c) PCA on Music Genres from GTZAN Dataset

Figure 2: Data Visualization for GTZAN Dataset

3 Data Pre-Processing

Data pre-processing is a crucial step to ensure the quality and consistency of the datasets before they are used for analysis or modelling. For the DEAM and Spotify datasets, the following pre-processing steps were applied:

1. **Cleaning and Formatting** : The datasets were cleaned to remove any missing or inconsistent data. This included handling missing values, standardizing the format of text fields, and ensuring numerical features were in the correct format.

2. **Feature Extraction**: For the DEAM dataset, audio features were extracted from the music tracks using audio analysis libraries. This included features such as tempo, key, mode, and various spectral features. For the lyrics, natural language processing techniques were used to extract features such as word count, sentiment, and thematic elements.

3. **Normalization**: Numerical features were normalized to ensure they were on a similar scale. This is important for models that are sensitive to the scale of input features, such as neural networks.

4. **Data Augmentation**: To enhance the diversity and size of the datasets, data augmentation techniques were applied. For the audio features, this included methods such as time stretching and pitch shifting. For the lyrics, techniques such as synonym replacement and back translation were used.

5. **Encoding**: Categorical features, such as mood labels in the DEAM dataset, were encoded using one-hot encoding or label encoding to convert them into a numerical format suitable for machine learning models.

6. **Splitting**: The datasets were split into training, validation, and test sets to evalu-

ate the performance of models and prevent overfitting.

4 Algorithm and Model

Methodology Overview

We began by comparing four different types of models (Random Forest Regressor, Support Vector Regression, Feed Forward Neural Network and 1D Convolutional Neural Network) to identify which one performs best for our needs.

Performance Metrics

The effectiveness of each model was evaluated using three key metrics:

- MSE (Mean Squared Error): This measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.
- RMSE (Root Mean Squared Error): This is the square root of the average of squared differences between prediction and actual observation. It offers a measure of the quality of the estimator.
- R^2 R-squared: This statistic provides an indication of goodness of fit and therefore a measure of how well-unseen samples are likely to be predicted by the model, relative to the mean of the observed data.

The models compared were:

Random Forest Regressor

A Random Forest Regressor is used to predict a continuous-valued output. A random forest is composed of multiple decision trees, each trained on a random subset of the training data and features. This process is known as bootstrapping. Each decision tree in the forest makes its own prediction for a given input. The final prediction of the random forest regressor is typically the average of all the individual tree predictions.[4] The Random Forest Regressor Model ran through the performance metrics R^2, Mean Square Error (MSE) and Root Mean Squared Error (RMSE) and is represented in Table 1.

Performance Metrics	Evaluate	Score
R^2 Score	Valence	0.3913
	Arousal	0.4901
MSE	Valence	0.6723
	Arousal	0.3913
RMSE	Valence	0.8199
	Arousal	0.9586

Table 1: Random Forest Performance Metrics

Support Vector Regression (SVR)

Support Vector Regression (SVR) aims to find a hyperplane (or a set of hyperplanes in higher-dimensional spaces) that best fits the training data in a way that the deviation from the actual target values is minimized.[4] The SVR Model ran through the performance metrics R^2, Mean Square Error (MSE) and Root Mean Squared Error (RMSE) and is represented in Table 2.

Feed Forward Neural Network

Two neural network configurations were created using the libraries: TensorFlow and Keras. A feed-forward neural network (NN) is shown in Figure 3. The 1-1-1 NN layer layout (Input, Hidden, and Output) architecture was to see if a simple NN architecture would be sufficient enough to predict valence and arousal through

Performance Metrics	Evaluate	Score
R^2 Score	Valence	0.3219
	Arousal	0.4551
MSE	Valence	0.7489
	Arousal	0.9821
RMSE	Valence	0.8654
	Arousal	0.9909

Table 2: SVR Performance Metrics

the use of the different audio features obtained. Otherwise, we would turn to the use of a more complex feed-forward structure.

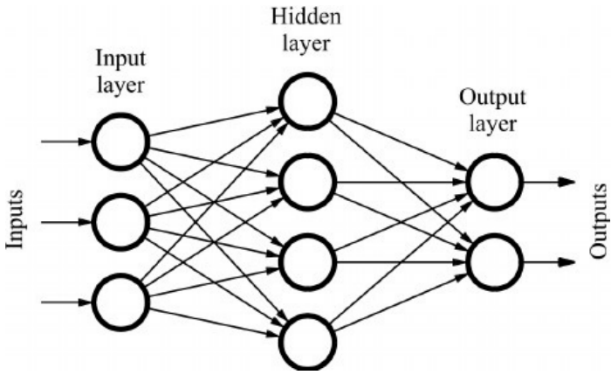


Figure 3: A Simple Feed Forward Neural Network

The result from the simple feed-forward network (Table 3) that had one hidden layer was not successful compared to the complex network. The neural network architecture developed is experimental and needs refinement. A different type of neural network may prove to be more efficient than a feed-forward network for this kind of problem. A recurrent neural network works well with speech recognition and would be a better option in proceeding with emotion prediction of audio.

LR = 0.005, epochs = 100	
Test MSE	0.7874
Test RMSE	0.8874
Test R^2	0.3061

Table 3: Simple Feed Forward Network Model Scores

1D Convolutional Neural Network

The 1D Convolutional Neural Network (Conv1D) model showed the most promising results at the initial stage.

To further enhance the performance of the

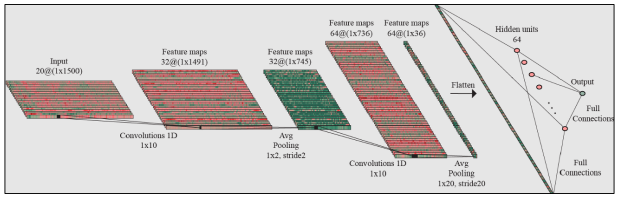


Figure 4: Conv1D Model Architecture

Conv1D model, we implemented the following strategies:

- Feature Addition: We included two additional features to provide the model with more information and potentially improve its learning capability.

- Data Augmentation: We expanded the dataset by augmenting it, which involves artificially increasing the diversity of data available for training without actually collecting new data.

- Attention Mechanism: We integrated an attention mechanism into the model, which helps the model focus on the most relevant parts of the data, potentially improving its accuracy.

The results from the enhanced performance of the Conv1D Model are seen in Table 4 where there is an increase in Test R^2 scores and a decrease in Test Loss scores.

Original Conv1D Model Scores		
Test Loss	Valence	0.78
	Arosual	1.00
Test R^2	Valence	0.32
	Arosual	0.43
Added Features, Attention and Augmentation		
Test Loss	Valence	0.29
	Arosual	0.71
Test R^2	Valence	0.75
	Arosual	0.78

Table 4: Conv1D Model Scores

Comparing Models

We compare the exploration of different model architectures. Each of these models represents a different approach to capturing and processing information from audio sequences to predict valence and arousal scores. The goal was to find the most suitable model for this task based on performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) values.

MSCNN (MultiScaleConvolutional Neural Network)
MSCNN is a variant of a convolutional neural network (CNN) that includes a MultiScaleConv layer. After the initial convolutional and pooling layers, a MultiScaleConv layer is introduced. This layer performs convolutions with multiple kernel sizes simultaneously (e.g., 3, 5, 7), capturing features at different scales before concatenating the outputs. The performance metrics can be seen in Table 5 (MSCNN).

TCN (Temporal Convolutional Network)
TCN is a type of neural network architecture designed for processing temporal sequences. This version of the model introduces a TCN architecture using TemporalBlock modules after the initial convolutional layers. Each block has dilated convolutions, allowing the network to have a larger receptive field and capture temporal dependencies over longer sequences. The performance metrics can be seen in Table 5 (TCN).

Soft Attention (SOFT)
Soft attention is a mechanism that allows neural networks to focus on specific parts of the input data. This version of AudioNet introduces an Attention module before the fully connected layers. The attention mechanism assigns importance weights to different parts of the feature map output by the previous layers. The performance metrics can be seen in Table 5 (SOFT).

Bidirectional (BI) LSTM with SOFT ATTENTION
Bidirectional LSTM is a type of recurrent neural network (RNN) that processes inputs in both forward and backward directions. In this model, a BI LSTM replaces the convolutional layers of the original AudioNet to capture dependencies in both directions of the audio sequence. Additionally, a self-attention mechanism computes importance weights across the sequence outputs of the Bi-LSTM. The performance metrics can be seen in Table 5 (BI LSTM).

After evaluating the performance of the models, we selected the Soft Attention model as the best performer. This model introduces an attention

		MSCNN	TCN	SOFT	BI LSTM
Valence	MSE	0.53	0.49	0.37	0.70
	RMSE	0.72	0.70	0.60	0.83
	R^2	0.49	0.53	0.63	0.33
Arousal	MSE	0.72	0.64	0.63	0.85
	RMSE	0.84	0.80	0.79	0.92
	R^2	0.58	0.63	0.63	0.50

Table 5: Comparing Models Performance Metrics

mechanism that computes a single weight for each feature in the feature map, which is then applied to the feature map before passing it to the fully connected layers. This mechanism allows the model to focus on the most relevant parts of the input, which proved beneficial for predicting valence and arousal scores in our dataset.

In conclusion, the Soft Attention model in AudioNet demonstrated superior performance compared to other explored models for predicting valence and arousal scores of audio files. This is likely due to its ability to focus on relevant features in the input data, which is crucial for predicting emotional responses to music. The attention mechanism in this model played a crucial role in improving prediction accuracy by focusing on relevant features in the input data.

5 Evaluation Methodology

Grid search is a hyperparameter optimisation technique used to find the best combination of hyperparameters for a machine learning model. In this context, grid search involves systematically searching through a predefined grid of hyperparameter values to identify the combination that minimises the MSE values and thus maximises the accuracy of the model in predicting the valence and arousal of the audio extracts. The hyperparameters are the various parameters related to the machine learning model used for emotion analysis. In this case, the hyperparameters considered were, learning rates(LR), number of epochs, weight decays and number of hidden units. The values of valence and arousal from the attention mechanism R^2 are 0.63 and 0.63 respectively (seen in Table 5 under SOFT attention). We use this score on the grid search parameters shown (Table 6) and get the valence and arousal Test R^2 values 0.75 and 0.78.

	Valence	Arousal
LR	0.0001	0.001
Epochs	150	150
Decay	0.01	0.005
Hidden	128	64
MSE	0.25	0.37
RMSE	0.50	0.60
Test R^2	0.75	0.78

Table 6: Hyperparameter Tuning: Testing Configuration

6 Results and Discussion

The results of performing the grid search indicated a significant reduction of MSE values obtained from hyperparameter tuning. MSE values were compared between those obtained before and after data augmentation to observe the improved accuracy of the model. Valence MSE value reduced to 0.25 from an initial value of 0.37 and Arousal MSE value reduced to 0.37 compared to the initial value of 0.63. These values indicated an improvement in the accuracy of the model analysed through hyperparameters.

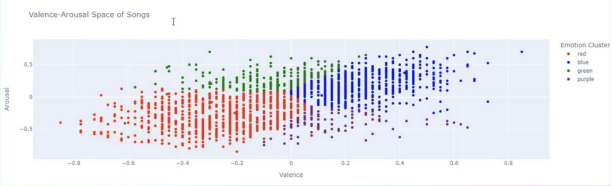


Figure 5: Scatter Plot for Emotion Mapping Clusters

Emotions identified from the audio extracts in the dataset were mapped and grouped into four different colours - purple, green, red and blue through K means clustering. (Figure 5) Data augmentation techniques such as pitch shifting were performed on the dataset to improve the distribution of data points in each colour group. Emotion mapping using K means clustering resulted in a 0.75 accuracy value for emotion colour classification.

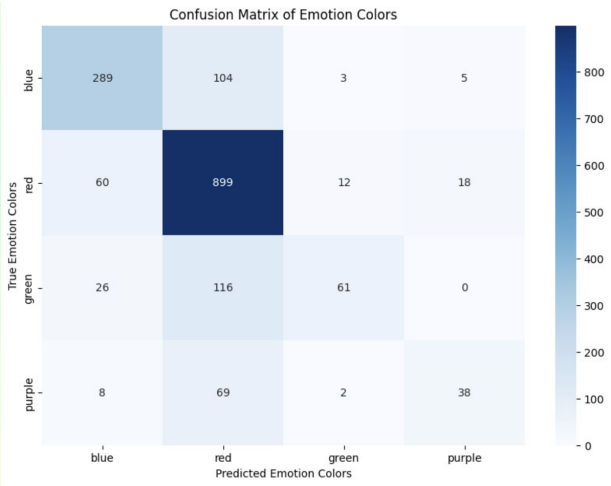


Figure 6: Confusion Matrix representing Accuracy of Emotion Colour Mapping

The confusion matrix (Figure 6) plotted to evaluate the performance of the emotion colour classification model also provided insight into how well the model can predict the correct colour to group the emotions for each input. In addition to the high accuracy score obtained for emotion colour mapping, the confusion matrix also provided an F1 score of 0.60 and a precision score of 0.72, indicating an overall enhancement of model accuracy due to precise emotion classification and effective dataset enhancement techniques.

7 Moodify's Recommender System

In the recommender system, we aimed to develop a system that allows users to input an MP3 file and receive predictions for the song's valence, arousal score, and genre. Additionally, the system enables users to explore similar songs in terms of valence and arousal scores using a chosen shape for projection.

Methodology

We initially collected a dataset of Spotify playlist songs and generated valence and arousal scores for evaluation. Using this dataset, we trained a machine learning model to predict valence, arousal scores, and genre based on song inputs.

User Interaction

Users can input an MP3 file, and our model predicts the valence, arousal scores, and genre of the song. Users can choose a shape (line, parabola, triangle, circle) that represents their musical preference. Based on the anchor point of the input song and the chosen shape, we find the closest 9 other points on the valence-arousal graph for music exploration.

Results

Our model achieves high accuracy in predicting valence and arousal scores, providing users with valuable insights into the emotional content of

the music. This model accurately predicts the genre of the input song, allowing users to discover new music within their preferred genre. Users can explore similar songs based on valence and arousal scores, enhancing their music discovery experience.

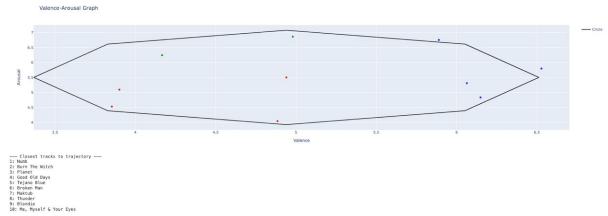


Figure 7: Valence-Arousal Graph for song: Numb - Linkin Park

For example, if a user inputs song: Numb by Linkin Park, the trajectory valence and arousal score of 4.94 and 5.50 respectively. The emotion detected shows it is red and the predicted genre of the song is rock. (Figure 7)

Conclusion

The developed system effectively predicts valence, arousal scores, and genre of music based on user inputs. Additionally, the music exploration feature provides users with a unique way to discover new music that aligns with their musical preferences. Overall, the system offers a comprehensive music analysis and exploration experience for users.

References

[1] Mohammad Soleymani, Anna Aljanaki, and Yi-Hsuan Yang, *MediaEval Database for Emotional Analysis in Music (DEAM)*, Swiss Center for Affective Sciences, University of Geneva, Switzerland; Academia Sinica, Taiwan, 2018.

[2] Github: YT-DLP - A feature-rich command-line audio/video downloader, `https://github.com/yt-dlp/yt-dlp`

[3] GTZAN Dataset - Music Genre Classification, `https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classificatio`

[4] J M Brotzer, E R Mosqueda, and K Gorro, *Predicting emotion in music through audio pattern analysis*, `https://iopscience.iop.org/article/10.1088/1757-899X/482/1/012021/pdf` University of San Carlos, Talamban, Cebu City, Cebu, Philippines Department of Computer and Information Sciences, University of San Carlos, Talamban, Cebu City, Cebu, Philippines, 2019.

[5] K R Tan, M L Villarino, C Maderazo, *Automatic music mood recognition using Russell's twodimensional valence-arousal space from audio and lyrical data as classified using SVM and Naïve Bayes*, `https://iopscience.iop.org/article/10.1088/1757-899X/482/1/012021/pdf` Department of Computer and Information Sciences - University of San Carlos – Talamban Campus, Cebu City, Philippines , 2019.