Modeling Arousal and Valence in Music: A Computational Approach to Affective Audio Analysis

Ashutosh Gupta
B.Tech Undergraduate, IIT Roorkee

July 2025

Abstract

This project represents my first deep dive into the emerging field of affective computing, approached through the lens of music emotion recognition. Specifically, it explores the prediction of two fundamental emotional dimensions — *arousal* (energy/intensity) and *valence* (pleasantness) — from audio signals. These dimensions underpin a range of applications from music recommendation to emotional indexing of multimedia.

The study began with handcrafted feature pipelines and traditional machine learning regressors, followed by a progression into deep learning architectures using log-mel spectrograms. Ultimately, the project converged on a research-backed hybrid strategy, leveraging pretrained audio embeddings (PANNs) for enhanced emotional representation.

Alongside technical development, this journey also served as a foundational research experience, helping bridge the gap between implementation and literature. The report systematically chronicles this iterative process — highlighting not only performance outcomes but also critical reflections, limitations, and future directions. By combining experimental rigor with conceptual learning, this work sets the stage for deeper contributions in the space of music-based affective analysis.

Introduction

Understanding and predicting the emotional impact of music has become an emerging field in music information retrieval and affective computing. Emotions in music are often represented along two continuous axes:

- **Arousal:** the energy or intensity of the music, ranging from calm to excited.
- **Valence:** the pleasantness or positivity, ranging from sad/negative to happy/positive.

These two dimensions offer a simple yet powerful way to model human emotional experience. My goal with this project is to explore how well we can predict these values from audio alone.

Why this problem? As someone intrigued by both music and machine learning, this problem seemed like the perfect entry point. It allows learning important ML/DL concepts while also being grounded in a relatable, expressive domain. More importantly, it offers opportunities to read real research papers, implement practical ideas, and reflect on what works and what doesn't.

Literature-Inspired Approach

To avoid random experimentation, this project was guided by influential studies in music emotion recognition.

Table 1: Key Research Contributions That Guided Modeling Strategy

Paper	**Focus**	**Impact on Project**
Cheuk et al. (2020) [1]	Handcrafted features vs neural embeddings	Motivated the shift away from MFCC-only models toward learned representations
Fong et al. (2022) [2]	Static vs temporal modeling	Encouraged incorporating CNN and GRU layers to model time-evolving musical emotion
Simonetta et al. (2024) [3]	Pretrained representations (PANNs)	Informed the hybrid model architecture using PANNs as embeddings
Soleymani et al. (2015) [4]	DEAM Dataset for Music Emotion	Provided a benchmark dataset with aligned valence/arousal annotations

These papers helped shape our transition: from traditional models to deep learning architectures, and finally to hybrid systems leveraging strong pretrained audio embeddings.

Project Trajectory and Methodological Evolution

This project evolved through a staged, insight-driven workflow that reflects both independent exploration and the gradual incorporation of academic best practices.

Phase 1 – Traditional ML Exploration: The journey began with a fully independent traditional machine learning pipeline. Without initially relying on external benchmarks, I extracted handcrafted audio features using signal processing techniques and modeled arousal and valence using a variety of regressors. This helped develop a grounded understanding of which features and modeling choices offered the most promise.

Phase 2 – Research-Guided Deep Learning Models: After establishing a strong baseline, I turned to recent literature that emphasized the limitations of handcrafted features and the advantages of learning hierarchical representations directly from audio. This led to the design and experimentation of four neural network architectures, ranging from CNNs to GRU-based temporal models. While these models provided greater representational flexibility, they did not yield breakthrough performance gains on their own.

Phase 3 – Hybrid Representation Learning via PANNs: To overcome performance plateaus and align more closely with cutting-edge research practices, I adopted a hybrid approach. Drawing inspiration from works such as Simonetta et al. [3], I used pretrained CNN14 models from the PANNs framework to extract expressive audio embeddings. These embeddings were then used as inputs to lightweight regression networks—successfully merging the robustness of pretrained representations with the interpretability and efficiency of classical modeling strategies.

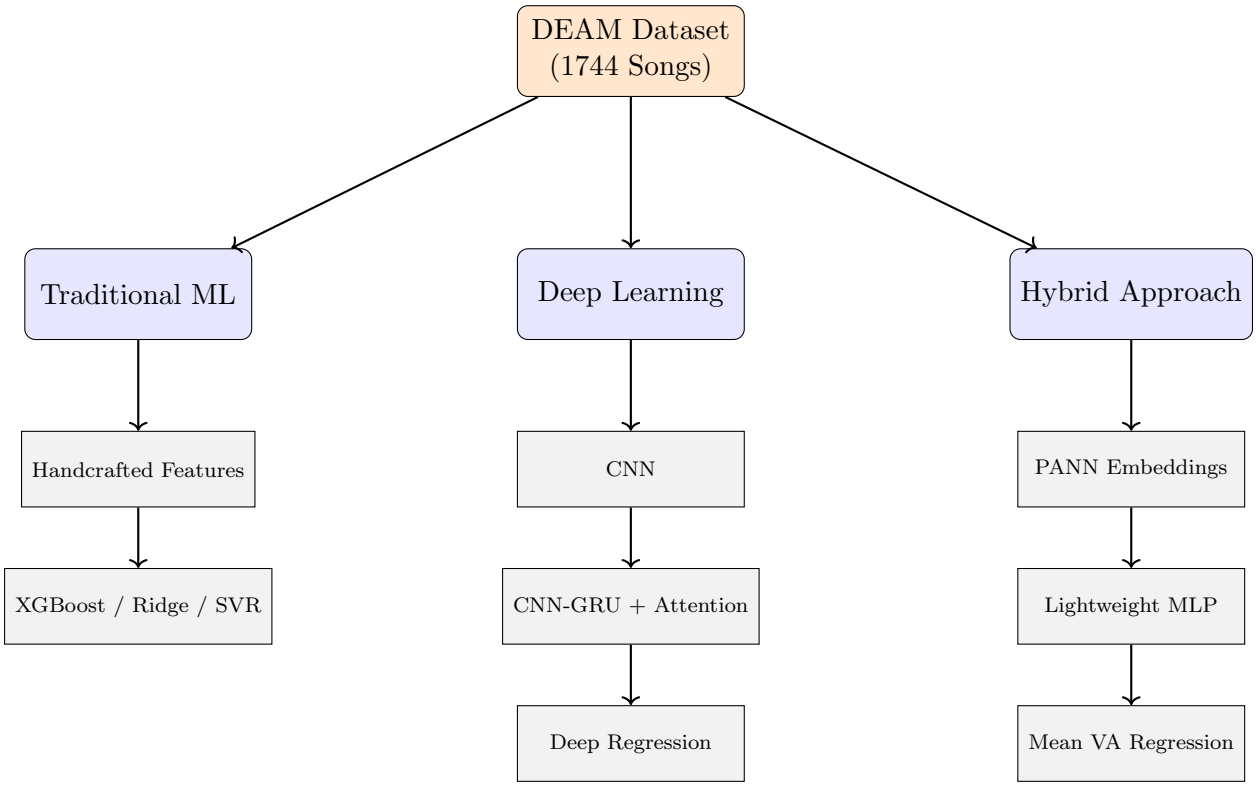


Figure 1: Overall Project Methodology: All approaches stem from the DEAM dataset.

This evolutionary trajectory—from independent experimentation to research-informed design—not only enhanced predictive performance but also deepened my understanding of how emotional cues are encoded in music. Each transition was guided by both empirical observations and theoretical justification, ensuring that every modeling shift was grounded in meaningful rationale.

Phase 1: Traditional Machine Learning Approach

Before diving into deep learning, I began this project by independently exploring traditional machine learning techniques to assess how well handcrafted features could predict emotional attributes in music. This initial phase provided foundational insights and served as a baseline for later approaches.

Dataset and Aggregation

The **DEAM dataset** was used for all experiments in this section. It contains **1744 full-length songs**, each annotated with valence and arousal labels sampled at regular

intervals. For traditional ML, I computed **song-level features** and used the **mean valence and arousal** for each song as the target values.

Feature Extraction

Using `librosa` and `pydub`, I extracted a rich set of handcrafted audio features. Each feature was summarized using **mean and standard deviation** over the full song, leading to a compact yet informative representation. The full feature set included:

- **MFCCs (Mel-frequency cepstral coefficients):** 20 coefficients × (mean + std) = 40 + 40 = 80 features
- **Chroma Features:** 12 pitch classes × (mean + std) = 12 + 12 = 24 features
- **Mel Spectrogram Stats:** 128 bands × (mean + std) = 128 + 128 = 256 features
- **Spectral Contrast:** 7 bands × (mean + std) = 7 + 7 = 14 features
- **Tonnetz Features:** 6 dimensions × (mean + std) = 6 + 6 = 12 features

This resulted in a total of **386 handcrafted features** per song.

Exploratory Data Analysis

Valence vs Arousal Relationship

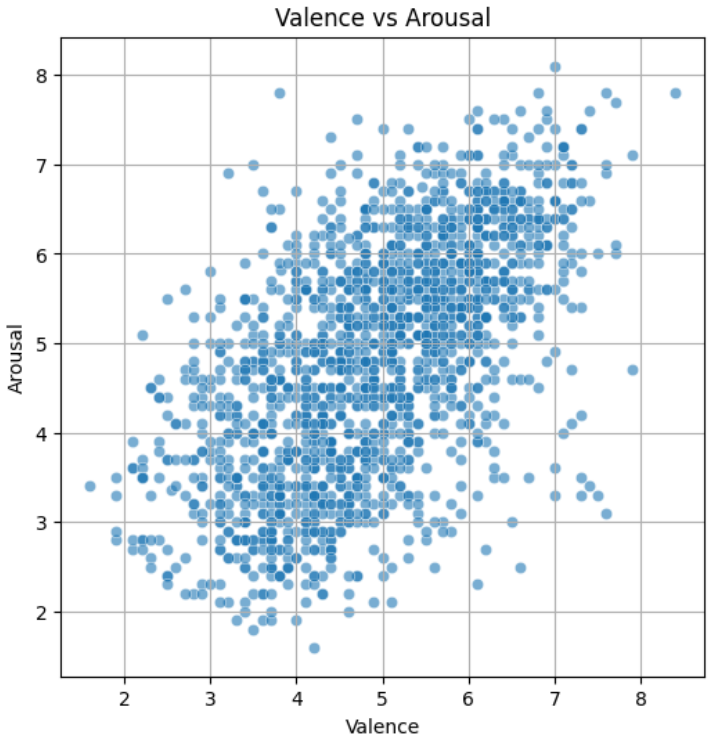
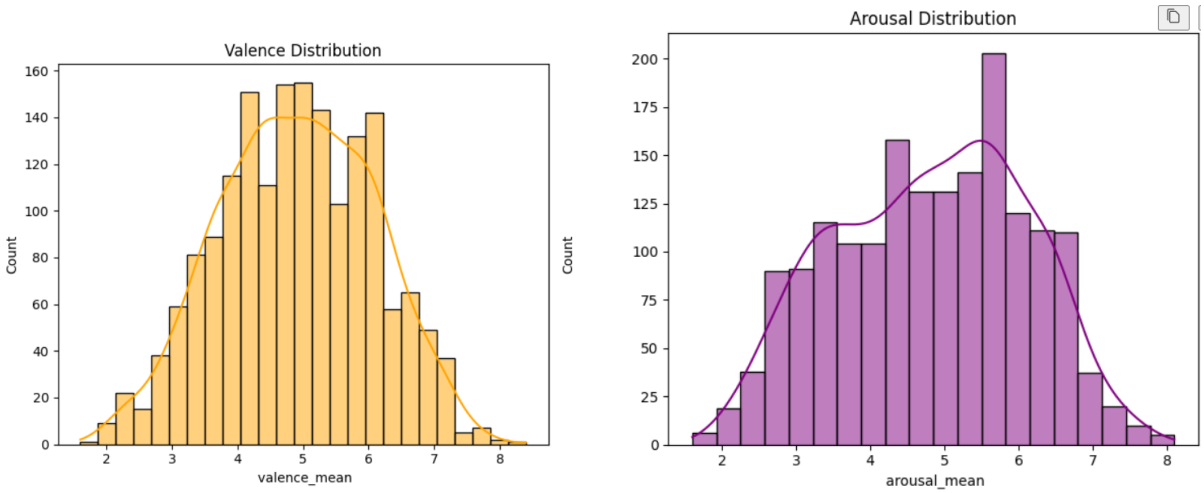


Figure 2: Scatter plot showing the relationship between valence and arousal

- Figure 2 shows the relationship between mean valence and arousal scores for 1744 songs in the DEAM dataset.

- There is a clear upward trend, indicating a moderate positive correlation — as valence increases, arousal tends to increase as well.
- Happier or more pleasant-sounding songs often carry higher energy, while sadder or more negative tracks tend to be calmer.
- This correlation implies that valence and arousal share mutual information, even though they are not perfectly colinear.
- From a modeling standpoint, this supports the use of multi-output regression models that learn shared patterns, rather than treating valence and arousal as fully independent targets.

Distribution of Valence and Arousal



(a) Valence Distribution

(b) Arousal Distribution

Figure 3: Histogram and kernel density plots of valence and arousal across songs

The distributions of valence and arousal are shown in Figures 3a and 3b, respectively. Both variables are continuous and follow approximately unimodal, bell-shaped distributions centered around neutral emotional states.

Valence is nearly symmetric and centered around 5 on a 1–9 scale, indicating a balanced representation of positive and negative emotional content. Arousal, on the other hand, shows a slight skew toward lower values, suggesting a greater presence of calm or low-energy tracks in the dataset.

These distributions are important from a learning perspective. A well-balanced target distribution reduces model bias and helps ensure generalization. However, slight skewness in arousal may result in underestimation of high-arousal predictions if not accounted for, making evaluation metrics like RMSE and MAE essential for performance monitoring.

Feature Selection

Given the high dimensionality of the extracted features, I applied both statistical and model-based feature selection techniques to isolate the most relevant predictors.

5.3.1 PCA Experimentation and its Limitations

As a dimensionality reduction attempt, I also applied Principal Component Analysis (PCA) on the handcrafted features to see if reducing the dimensionality would improve generalization. However, this approach resulted in noticeably lower performance across all evaluation metrics.

One probable reason is that PCA is an unsupervised method that projects data along axes of maximum variance without considering the target variable. This means the resulting components might capture irrelevant acoustic variance (e.g., volume shifts, background noise) rather than emotionally salient information like timbral changes or harmonic content. In contrast, methods like SHAP and Pearson, which are label-aware (either explicitly or via supervised modeling), were able to preserve the emotional signal more effectively.

5.3.2 Pearson Correlation Analysis

To identify features most linearly associated with arousal and valence, I used the **Pearson correlation coefficient**.

This metric captures the degree of linear relationship between a feature x and target y, calculated as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- $r_{xy} \in [-1, 1]$: values closer to ± 1 indicate stronger linear dependence
- $\bar{x}$, $\bar{y}$ denote feature and target means

For each audio-derived feature, the Pearson coefficient was computed against the valence and arousal targets separately. I then selected the **top 15 features (by absolute correlation)** for both dimensions.

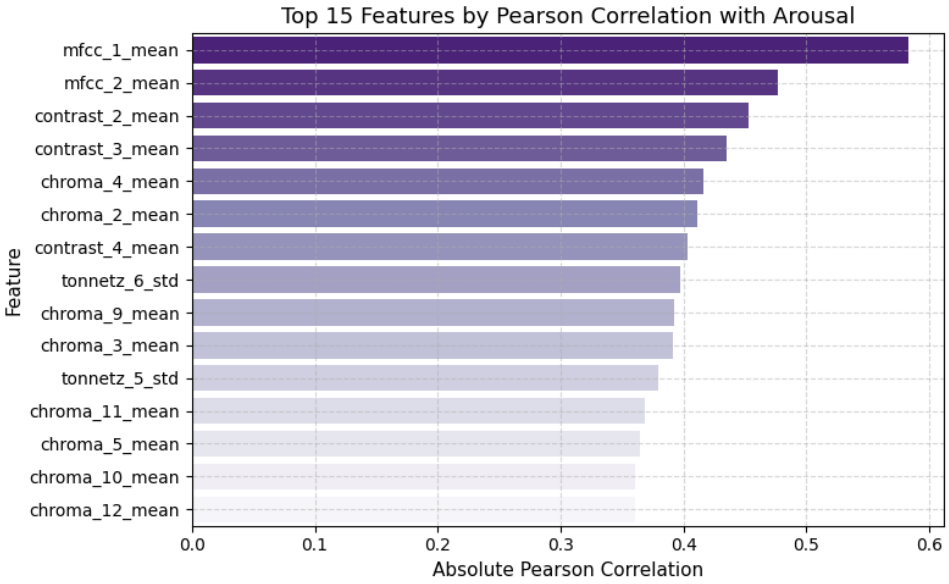


Figure 4: Top 15 Features by Pearson Correlation with Arousal

Arousal Correlation As seen in Figure 4, arousal strongly correlates with:

- `mfcc_1_mean`, `mfcc_2_mean` — low-order MFCCs indicating broad energy changes.
- `contrast_2_mean`, `contrast_3_mean` — spectral contrast may capture dynamics influencing perceived excitement.
- `chroma_4_mean`, `chroma_2_mean` — indicating some tonality-based emotional clues.
- `tonnetz_6_std` — tonal dissonance/instability appears linked to higher arousal.

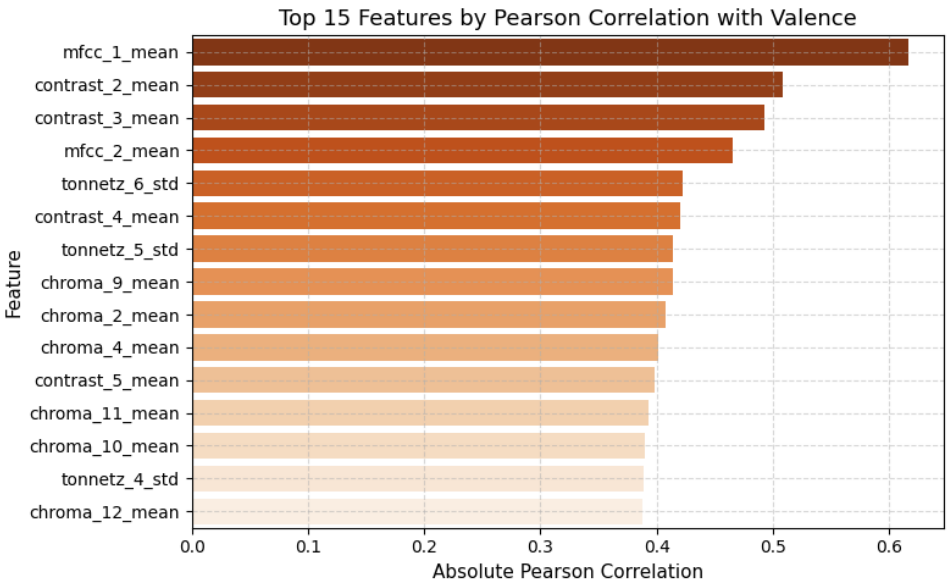


Figure 5: Top 15 Features by Pearson Correlation with Valence

Valence Correlation As shown in Figure 5, valence was most associated with:

- `mfcc_1_mean` again — representing stable spectral structure linked to perceived positivity.
- `contrast_2_mean` to `contrast_4_mean` — high-frequency contrast appears tied to brighter, happier music.
- `tonnetz_6_std`, `tonnetz_5_std` — tonal stability again plays a key role.
- `chroma_11_mean`, `chroma_10_mean` — suggesting pitch class usage patterns can influence valence perception.

These insights were essential for the initial feature selection, as they helped reduce the high-dimensional feature set while retaining the most relevant emotional predictors.

5.3.3 SHAP-Based Feature Importance

To capture complex non-linear relationships between audio features and emotional dimensions, I trained an XGBoost model and employed SHAP (SHapley Additive exPlanations) values to quantify global feature importance. Unlike traditional correlation-based methods such as Pearson, SHAP values account for feature interactions and provide consistent, interpretable attributions for model predictions.

Figure 6 and Figure 7 present SHAP summary plots for valence and arousal predictions, respectively. Each point represents an individual SHAP value, with its color indicating the corresponding feature value (red = high, blue = low).

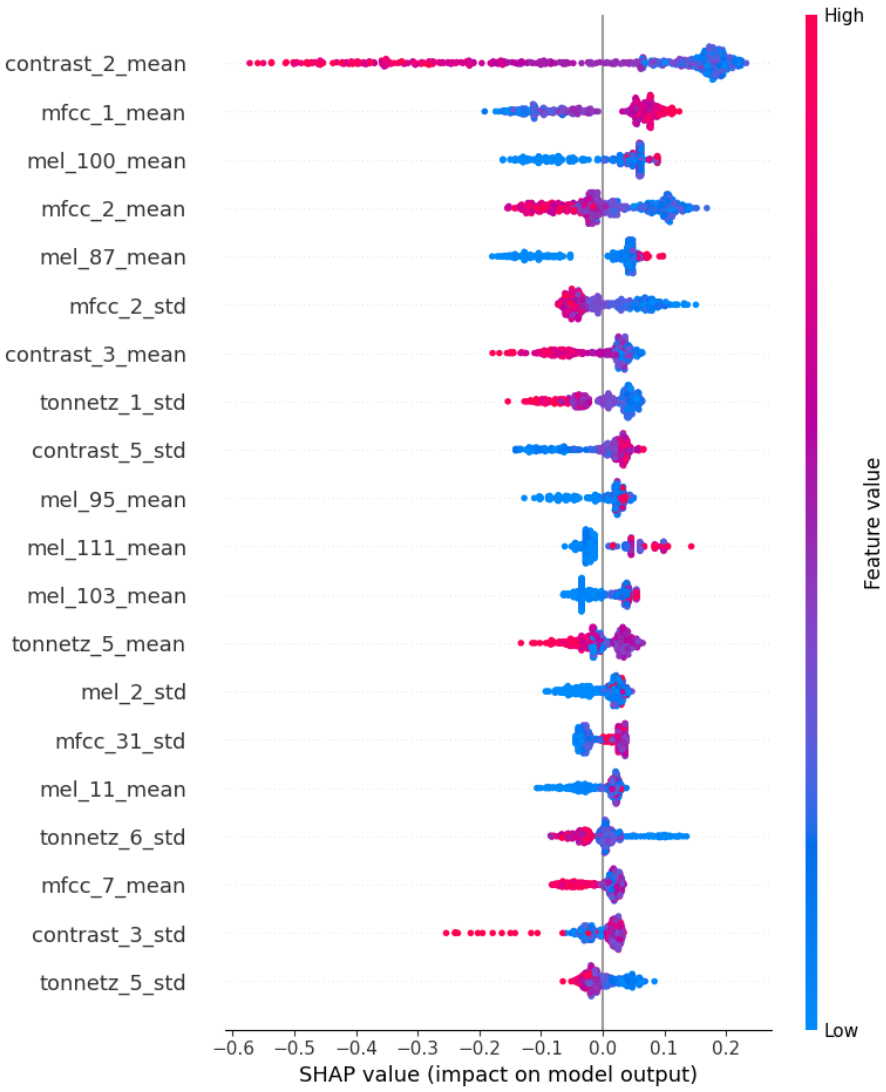


Figure 6: SHAP summary plot for Valence prediction using XGBoost.

For **valence**, the most influential features were `contrast_2_mean`, `mfcc_1_mean`, and `mel_100_mean`. Notably, `contrast_2_mean` exhibited a strong negative SHAP value when high, implying that greater spectral contrast may be associated with lower emotional positivity.

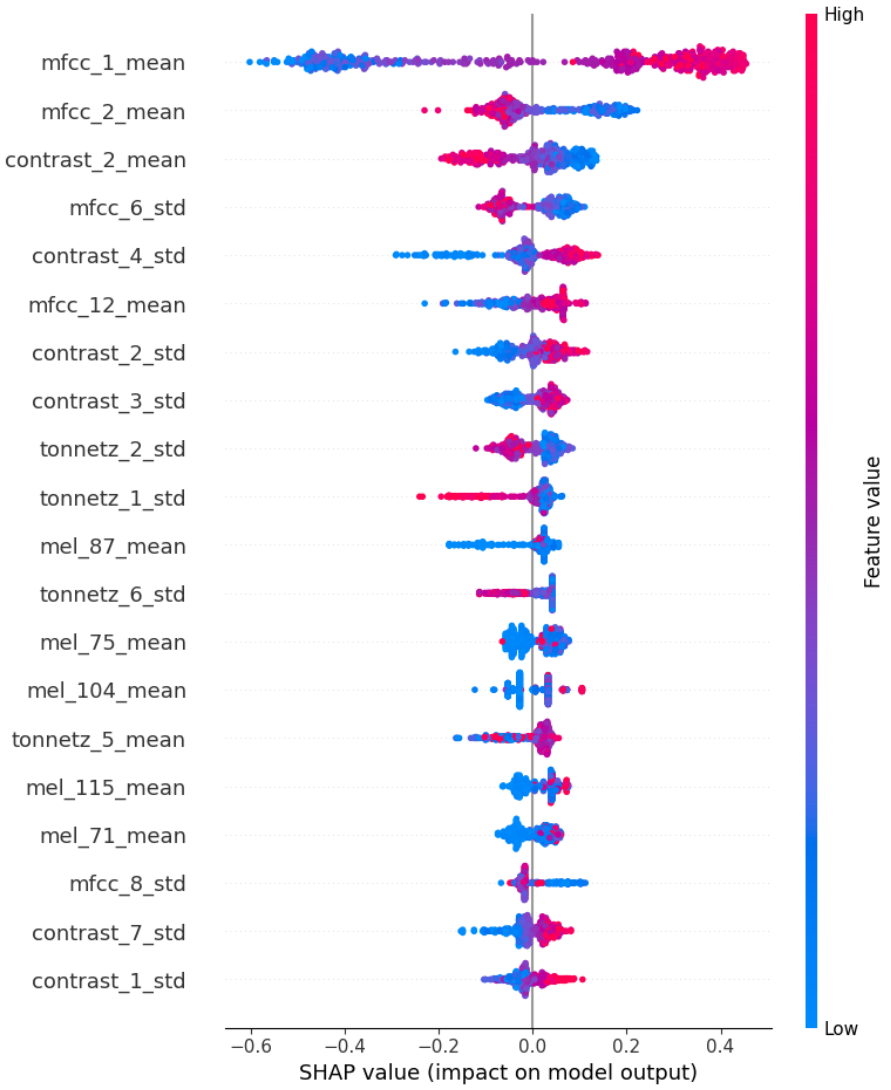


Figure 7: SHAP summary plot for Arousal prediction using XGBoost.

In the case of **arousal**, the most impactful predictors included `mfcc_1_mean`, `mfcc_2_mean`, and `contrast_2_mean`. Higher values of `mfcc_1_mean` correlated with increased arousal scores, suggesting that sharper spectral envelopes may convey heightened energetic states.

Overall, the SHAP visualizations reinforce the hypothesis that valence and arousal are influenced by distinct yet overlapping sets of spectral and cepstral features. These insights validate the model's capacity to detect nuanced patterns between audio characteristics and emotional responses.

Feature Combination Experiments

I evaluated multiple combinations of Pearson and SHAP-selected features. The table below summarizes their predictive performance (measured using R^2, RMSE, and MAE):

Table 2: Performance of Feature Combinations (Valence / Arousal)

Feature Set	R^2	**MAE**	**RMSE**
Full Dataframe	0.554 / 0.5267	0.623 / 0.694	0.767 / 0.874
Top 15 Pearson	0.504 / 0.499	0.650 / 0.722	0.808 / 0.899
Union (SHAP + Pearson)	0.505 / 0.462	0.657 / 0.748	0.808 / 0.931
Intersection Only	0.466 / 0.411	0.682 / 0.781	0.839 / 0.975

Modeling and Hyperparameter Tuning

Using `GridSearchCV` with 5-fold cross-validation, I tested the following regression models:

- Ridge Regression
- Lasso Regression
- Support Vector Regressor (SVR)
- Random Forest Regressor
- Gradient Boosting Regressor
- **XGBoost Regressor**

XGBoost consistently outperformed others on both valence and arousal prediction. Best scores (from full dataframe feature set):

- **Valence:** $R^2 = 0.554$, RMSE = 0.767, MAE = 0.623
- **Arousal:** $R^2 = 0.526$, RMSE = 0.874, MAE = 0.694

Conclusion of This Phase

The traditional machine learning phase provided a solid foundation for understanding the relationships between handcrafted audio features and emotional attributes in music. Using the DEAM dataset and a comprehensive set of time-aggregated descriptors (e.g., MFCCs, Chroma, Tonnetz), I was able to build a structured modeling pipeline with rigorous feature selection techniques like Pearson correlation and SHAP. This not only improved model performance, but also yielded interpretable insights into which spectral or tonal elements contribute to perceived valence and arousal.

However, despite careful feature engineering and selection, the best-performing model — XGBoost — yielded R^2 scores in the 0.50–0.55 range for both targets. This performance plateau is consistent with findings in prior literature [1], [3], which highlight that such handcrafted features often fail to capture the temporal, hierarchical, and semantic richness inherent in musical audio.

Moreover, the static nature of traditional features — summarizing an entire song by simple statistics — ignores the sequential structure that defines emotional evolution in music. These limitations motivated a paradigm shift towards deep learning architectures that can learn hierarchical representations directly from spectrograms or raw waveforms, and capture time-varying emotional dynamics.

In the next phase, I embrace this shift by implementing a set of deep learning models inspired by recent research, gradually increasing architectural complexity and temporal awareness. This transition marks a deliberate movement from feature-driven design to representation learning, with the goal of better aligning model capacity with the emotional richness of musical content.

Phase 2: Deep Learning Progression

Why Move Beyond Traditional Models?

Traditional machine learning approaches offered a solid baseline, but their reliance on static, song-level summaries limits their ability to capture the **temporal nature** of music. Emotion in music unfolds over time — shaped by rhythm, dynamics, and progression — and aggregating features over an entire song discards this crucial structure.

Recent research supports this limitation. Cheuk et al. [1] demonstrated that hand-crafted features fail to model time-evolving emotional cues, especially for arousal. Fong et al. [2] compared static vs. temporal approaches and found that models incorporating **sequence modeling** (like CNN-GRU) better captured emotional trajectories, especially when using frame-level input.

Motivated by these insights and the performance ceiling observed in Phase 1 ($R^2 \approx 0.55$), I transitioned to deep learning architectures designed to model **spatiotemporal patterns**. This included:

- **CNNs** for local temporal-spectral learning,
- **CNN-GRU** to model longer-term temporal dependencies,
- **Attention-enhanced variants** to learn weighted emotional influence over time.

This shift aimed to address the limitations of static modeling and better align with the research-backed direction of modern music emotion recognition.

Data Preparation: Frame-Level Spectrograms

To shift from static, song-level modeling to dynamic, frame-level deep learning, I processed the DEAM dataset into 3-second audio chunks. Each chunk was converted into a fixed-size **log-mel spectrogram** using the `librosa` library. These spectrograms serve as 2D time-frequency representations, suitable for CNN-based architectures.

Key preprocessing steps:

- **Sample rate:** 22,050 Hz
- **Chunk duration:** 3 seconds (66,150 samples)
- **Mel bands:** 128
- **Fixed shape:** 128×128
- **Normalization:** Power spectrogram converted to dB scale

Each of the 1744 full-length songs was segmented, and one spectrogram per chunk was saved as a `.npy` file. This enabled a frame-wise training pipeline, aligning spectrograms with their corresponding time-sampled valence and arousal annotations.

Spectrogram Dataset Construction

To enable training of convolutional models, I converted each song in the DEAM dataset into a log-mel spectrogram representation using 3-second audio chunks. Each spectrogram was then paired with a valence or arousal target label, forming the basis for supervised learning.

6.3.1 Target Label Aggregation

For this phase, I used song-level **mean valence and mean arousal values** as target labels, extracted from the official DEAM CSV file. The label CSV was preprocessed as follows:

- Filtered to retain only relevant columns: `song_id`, `valence_mean`, and `arousal_mean`.
- Converted the `song_id` field to string format for consistent filename matching.
- Stored all labels in a Python dictionary for efficient indexing during training.

6.3.2 Dataset Implementation

I implemented a custom PyTorch `Dataset` class to load precomputed log-mel spectrograms and their associated labels on the fly. Each item in the dataset returns:

- A `(1, 128, 128)` log-mel spectrogram tensor
- A single float label tensor for either valence or arousal

```
class LogMelEmotionDataset(Dataset):
  def __init__(self, npy_dir, label_map, target='valence_mean'):
    # Loading filenames and targets
  def __getitem__(self, idx):
    # Loading .npy spectrogram as (1, 128, 128)
    # Return:  (mel_tensor, label_tensor)
```

Figure 8: Simplified structure of the spectrogram-based PyTorch dataset

This modular setup supports training of CNN-based architectures on log-mel inputs, with flexibility to switch between valence and arousal targets.

Model 1: CNN-Based Regression on Log-Mel Spectrograms

To establish a deep learning baseline, I implemented a convolutional neural network (CNN) designed to regress either valence or arousal values from log-mel spectrograms. This model treats emotion prediction as a supervised regression task using song-level audio summaries.

6.4.1 Architecture Overview

The architecture follows a standard 2D CNN structure:

- **Input:** `(1, 128, 128)` log-mel spectrogram
- **Three convolutional blocks**, each consisting of:
  - `Conv2D` layer with increasing filter sizes (16, 32, 64)
  - `BatchNorm2D` for improved convergence
  - `ReLU` activation
  - `MaxPooling2D` to downsample the spatial resolution
- **Flattening layer** to convert spatial features into a dense vector
- **Fully-connected layers:**
  - `fc1`: $64 \times 16 \times 16 \rightarrow 128$, with `Dropout(0.3)`
  - `fc2`: $128 \rightarrow 1$, producing the final regression output

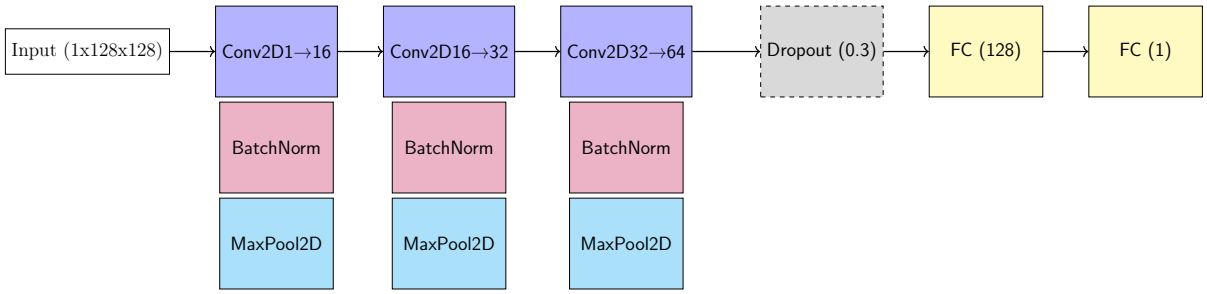


Figure 9: CNN architecture for emotion regression (valence/arousal) based on 3 convolutional blocks.

6.4.2 Forward Pass Logic

The input passes through three convolutional blocks that extract hierarchical features at increasing levels of abstraction. Max pooling reduces spatial size and helps with translation invariance. After the convolutional stack, the feature map is flattened and passed through a dropout-regularized dense layer before outputting a single continuous prediction.

The model is defined using PyTorch as follows:

```
class EmotionCNN(nn.Module):
  def __init__(self):
    Conv → BN → ReLU → Pool (×3)
    FC(16384 → 128) → ReLU → Dropout → FC(128 → 1)

  def forward(self, x):
    x = self.conv_block(x) → flatten → fc1 → fc2 → prediction
```

Figure 10: Simplified structure of the EmotionCNN architecture

6.4.3 Why CNN for This Task?

CNNs are particularly effective for analyzing 2D structured data like spectrograms, where local time-frequency patterns can be detected via convolutional kernels. This architecture can learn:

- Temporal energy shifts (vertical patterns)
- Spectral envelope variation (horizontal bands)
- Emotional indicators like timbral brightness or harmonic richness

6.4.4 Loss and Target

Each training instance corresponds to a `(mel, valence)` or `(mel, arousal)` pair. The model is trained using mean squared error (MSE) loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the predicted score and y_i is the ground truth.

6.4.5 Training Results and Evaluation

The CNN model was trained for 15 epochs with early stopping enabled, targeting either `valence` or `arousal` prediction. The model was evaluated using standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

Valence Prediction Results are summarized in Table 3, showing relatively consistent improvements in RMSE and R^2 through early epochs:

Table 3: CNN Regression Results per Epoch (Valence Prediction)

Epoch	**Loss (MSE)**	R^2	**RMSE**	**MAE**
1	28.0040	0.1267	1.0648	0.8382
2	2.0038	0.2624	0.9786	0.7845
3	1.8051	0.3816	0.8960	0.7123
4	1.7211	-0.0779	1.1830	0.9610
5	1.5826	0.4350	0.8565	0.6781
6	1.7031	0.3648	0.9081	0.7297
7	1.6268	0.3768	0.8995	0.7215
8	1.5580	0.4279	0.8619	0.6896
9	1.5125	0.2281	1.0011	0.8001
10	1.4352	-0.1816	1.2386	1.0219
11	1.4960	0.4138	0.8724	0.6996
12	1.3375	0.4315	0.8592	0.6852
13	1.2732	0.3752	0.9007	0.7215
14	1.2063	0.3759	0.9002	0.7163
15	1.2322	0.3102	0.9464	0.7638

Arousal Prediction Results are shown in Table 4. Performance showed early gains but remained more volatile, indicating a harder regression task.

Table 4: CNN Regression Results per Epoch (Arousal Prediction)

Epoch	Loss (MSE)	R^2	RMSE	MAE
1	40.7200	-0.2169	1.3902	1.1476
2	2.1395	0.3786	0.9935	0.7884
3	1.9700	0.1577	1.1566	0.9114
4	1.7370	0.3773	0.9945	0.8055
5	1.7221	0.4467	0.9375	0.7546
6	1.7936	0.3681	1.0018	0.8116
7	1.8717	0.4160	0.9630	0.7642
8	1.6552	0.3621	1.0066	0.7868
9	1.7185	0.4373	0.9454	0.7504
10	1.6822	0.1637	1.1525	0.9293
11	1.4591	0.4170	0.9622	0.7672
12	1.3878	0.4391	0.9439	0.7466
13	1.3379	0.4352	0.9472	0.7556
14	1.3945	**0.4830**	**0.9061**	**0.7220**
15	1.2387	0.3127	1.0448	0.8379

Insight: While the valence model showed relatively consistent improvements in early epochs with a best R^2 of 0.4350 (epoch 5), arousal prediction was more challenging and erratic. The best arousal R^2 of 0.4830 was observed at epoch 14, though with signs of instability elsewhere. These results reflect that emotional energy (arousal) may be harder to infer from static mel-spectrograms alone compared to emotional positivity (valence), potentially requiring more temporal modeling to capture rapid dynamic changes.

Model 2: SE-Attention CNN with Temporal Convolution

To address the limitations of the basic CNN in modeling temporal evolution and channel-wise relevance in spectrograms, I designed a second architecture incorporating **Squeeze-and-Excitation (SE)** blocks and a **1D convolution** over the flattened time-frequency axis.

6.5.1 Architecture Overview

The network introduces two important modifications over the baseline CNN:

- **Squeeze-and-Excitation (SE) Blocks:** These apply channel-wise attention by first compressing the spatial dimensions via global average pooling and then expanding them back through a two-layer feedforward network. This mechanism allows the network to recalibrate feature maps and prioritize emotionally salient channels.
- **Temporal Convolution:** After spatial convolution and pooling, the output tensor is reshaped and passed through a `Conv1D` layer to explicitly model temporal dependencies.

The rest of the network follows a similar structure: fully connected layers after flattening, dropout regularization, and a final regression output.

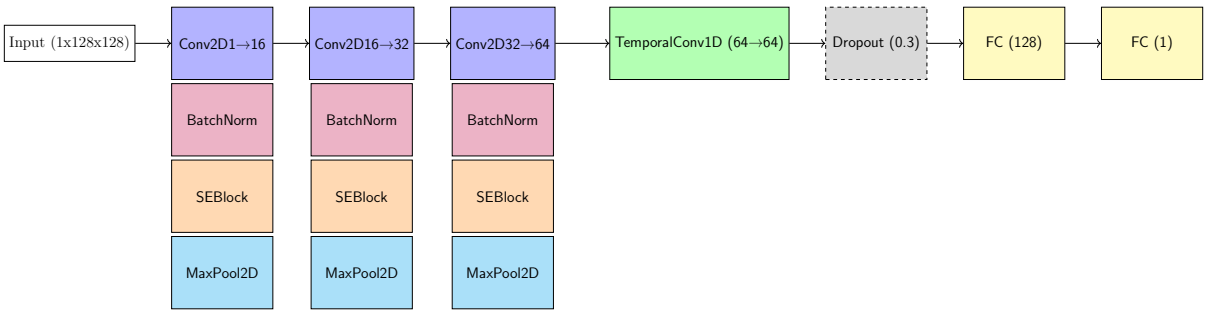


Figure 11: Proposed CNN model with SE blocks and temporal Conv1D for emotion regression.

6.5.2 Forward Pass Logic

The model processes the input spectrogram through three convolutional blocks with attention, followed by a temporal convolution and fully connected layers. Each block consists of:

- `Conv2D` → `BatchNorm` → `ReLU` → `MaxPooling2D`
- `SEBlock`: Global pooling and gated excitation

Then, the tensor is reshaped to `(Batch, Channels, Time × Freq)` and passed through a `Conv1D` temporal block. This output is flattened and passed through a dense layer to generate the final valence or arousal score.

6.5.3 PyTorch Skeleton

```
class EmotionCNN_SE_Temporal(nn.Module):
  ConvBlock1 (1→16) + SE → Pool
  ConvBlock2 (16→32) + SE → Pool
  ConvBlock3 (32→64) + SE → Pool
  Reshape:  (B, C, Time × Freq) → Conv1D (Temporal)
  FC (flattened → 128) → Dropout → FC (1)
```

Figure 12: Simplified architecture logic for SE + Temporal CNN

6.5.4 Rationale for Design Enhancements

This model addresses two critical challenges in audio-based emotion regression:

1. **Channel Redundancy**: SE blocks help suppress irrelevant channels and enhance useful ones — beneficial when many spectrogram bands may be emotionally neutral.
2. **Temporal Flattening Problem**: Unlike the basic CNN, this model uses temporal convolution to retain a degree of time-aware processing, which is more biologically and psychologically aligned with how we perceive evolving emotion in music.

6.5.5 Loss and Target

As in the previous model, we frame this as a regression task. Each training instance is a spectrogram-label pair, and the model minimizes Mean Squared Error:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the predicted valence/arousal and y_i is the ground truth from the DEAM dataset.

6.5.6 Training Results and Performance

The SE-enhanced temporal CNN was trained using Mean Squared Error (MSE) loss for up to 50 epochs with early stopping enabled. The model was trained independently for both `valence` and `arousal` targets, and evaluated using standard regression metrics: R^2, RMSE, and MAE.

Valence Results: The best validation performance was observed at **epoch 14**, achieving:

- **R^2:** 0.4642
- **RMSE:** 0.8340
- **MAE:** 0.6688

Table 5: Model 2 Valence Results (Selected Epochs)

Epoch	R^2	RMSE	MAE
1	0.1267	1.0648	0.8382
4	0.4419	0.8512	0.6852
6	0.4516	0.8438	0.6710
10	0.4361	0.8556	0.6840
14	**0.4642**	**0.8340**	**0.6688**
18	0.4067	0.8777	0.7066
21	0.4377	0.8545	0.6962
24	0.4054	0.8786	0.7091

Arousal Results: The best arousal performance occurred at **epoch 19**, reaching:

- **R^2:** 0.5273
- **RMSE:** 0.8665
- **MAE:** 0.6834

Table 6: Model 2 Arousal Results (Selected Epochs)

Epoch	R^2	RMSE	MAE
2	0.4507	0.9340	0.7475
5	0.4703	0.9172	0.7433
8	0.4987	0.8923	0.7143
12	0.4807	0.9082	0.7223
15	0.4843	0.9050	0.7167
18	0.5239	0.8696	0.7002
19	**0.5273**	**0.8665**	**0.6834**
24	0.5066	0.8852	0.7184
29	0.5004	0.8908	0.7180

Insight: Compared to the plain CNN, this SE-based model demonstrated a consistent improvement in both valence and arousal prediction. Its architecture — combining spectral and temporal abstraction with channel-wise attention — enabled more robust feature learning. Arousal predictions benefited particularly well, surpassing all previous models with a peak R^2 of 0.5273. Notably, the smoother loss progression and convergence behavior also reflected reduced overfitting, affirming the utility of adding domain-aligned inductive biases like temporal filters and SE-blocks in this phase of the pipeline.

Model 3: SE-Enhanced CNN + GRU Temporal Modeling

In this third architecture, I explored combining convolutional feature extraction with a GRU (Gated Recurrent Unit) to better capture temporal dependencies in music. This hybrid CNN-GRU model draws inspiration from music emotion recognition studies that emphasize sequential modeling of timbre and rhythm over time.

6.6.1 Architecture Overview

The model consists of:

- **Three CNN blocks:** Each composed of:
  - `Conv2D` → `BatchNorm2D` → `ReLU` → `MaxPool2D`
  - `SEBlock`: Squeeze-and-Excitation for channel-wise attention
- **Temporal Modeling:** After convolutional processing, the output is reshaped into a time-series and passed to a single-layer GRU.
- **Fully-Connected Layers:**
  - `fc1: GRU output → 128`, followed by `Dropout(0.3)`
  - `fc2: 128 → 1` for final valence/arousal prediction

The GRU layer enables the model to understand how feature patterns evolve over time—something not possible in purely CNN-based architectures.

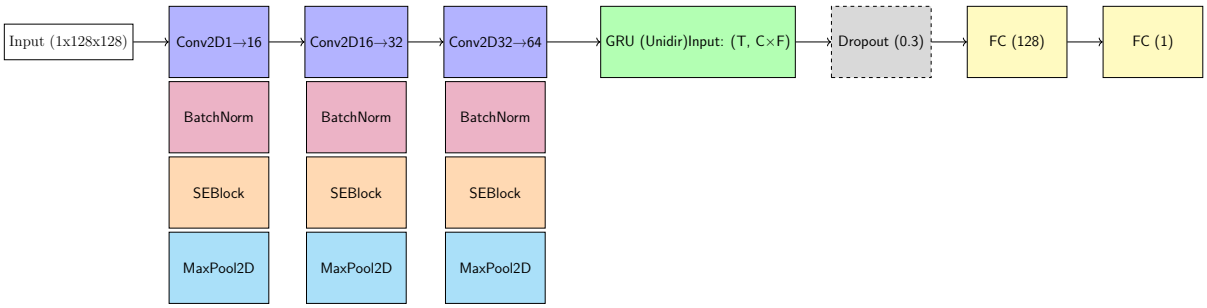


Figure 13: Proposed CNN-GRU model with SE blocks for temporal emotion regression.

6.6.2 PyTorch Implementation Summary

Key Logic:

- Spectrogram shape after convolution: (B, C, T, F)
- Reshaped to sequence: $(B, T, C \times F)$
- GRU encodes time-dependent structure, with only the final hidden state used.

```
class EmotionCNN_GRU(nn.Module):
  conv1 → SE → pool → conv2 → SE → pool → conv3 → SE → pool
  reshape:  (B, C, T, F) → (B, T, C×F)
  GRU (input_size=C×F, hidden=64)
  FC: GRU_out → 128 → Dropout → 1
```

Figure 14: Simplified PyTorch structure for the EmotionCNN_GRU architecture

6.6.3 Motivation for CNN-GRU Hybridization

While CNNs capture local spatial features across time-frequency space, GRUs are well-suited for learning long-range temporal dependencies in sequential data. By integrating both, the model can:

- Leverage low- and mid-level spectral patterns (via CNN)
- Model their dynamic progression (via GRU)
- Assign attention weights through SE blocks to prioritize relevant channels

6.6.4 Loss and Objective

Like the previous models, this network minimizes MSE loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

The prediction target is again either `valence_mean` or `arousal_mean`, depending on training objective.

6.6.5 Training Performance

The CNN-GRU hybrid model was trained for up to 50 epochs with early stopping based on validation R^2. Training was conducted separately for valence and arousal prediction using Mean Squared Error (MSE) loss. The model's performance was evaluated using R^2, RMSE, and MAE metrics.

Valence Results: The best performance occurred at **Epoch 7**:

- $\mathbf{R^2 = 0.4599}$
- RMSE = 0.8374
- MAE = 0.6627

Table 7: Model 3 Valence Results (Selected Epochs)

Epoch	R^2	RMSE	MAE
1	0.2591	0.9808	0.8074
3	0.3824	0.8954	0.7060
5	0.3649	0.9081	0.7332
7	**0.4599**	**0.8374**	**0.6627**
10	0.4428	0.8506	0.6739
14	0.4591	0.8380	0.6769
17	0.4315	0.8592	0.6799

Arousal Results: The best arousal performance was observed at **Epoch 16**:

- $\mathbf{R^2 = 0.5119}$
- RMSE = 0.8805
- MAE = 0.6853

Table 8: Model 3 Arousal Results (Selected Epochs)

Epoch	R^2	RMSE	MAE
1	0.0024	1.2588	1.0697
6	0.4215	0.9585	0.7494
7	0.4888	0.9011	0.7058
10	0.4533	0.9319	0.7348
14	0.4113	0.9669	0.7667
16	**0.5119**	**0.8805**	**0.6853**
23	0.4802	0.9086	0.7191
26	0.4983	0.8926	0.7181

Insight: Compared to earlier models, the addition of a GRU module introduced sequential awareness, which enhanced the network's ability to process long-term emotional transitions within the audio. This translated into slightly stronger R^2 scores for both valence and arousal, along with lower error values. The improvement, though not drastic, supports the benefit of incorporating temporal recurrence into spectro-temporal emotion modeling.

Model 4: CNN-GRU with Attention Mechanism

This model extends the CNN-GRU architecture by integrating an **attention mechanism** on top of the GRU outputs. Inspired by recent sequence learning literature, attention allows the model to focus selectively on important temporal regions within the song's spectrogram, rather than relying solely on the final hidden state of the GRU.

6.7.1 Architecture Overview

The model comprises the following major components:

- **Input:** `(1, 128, 128)` log-mel spectrograms.
- **CNN Feature Extractor:**
  - 3 convolutional blocks: `Conv2D → BatchNorm → ReLU → MaxPool → SEBlock`
  - Output tensor shape: `(B, C=64, T=16, F=16)` for further temporal modeling
- **GRU Temporal Model:**
  - 1-layer unidirectional GRU
  - Input reshaped to `(B, T=16, C×F)` = `(B, 16, 1024)`
  - Output: full GRU sequence `(B, 16, H=64)`
- **Attention Block:**
  - A trainable attention layer computes weights over each timestep
  - Weighted average of GRU outputs provides a focused representation
- **Fully-connected Layers:**
  - `fc1: 64 → 128`, followed by `ReLU + Dropout`
  - `fc2: 128 → 1`, final regression output

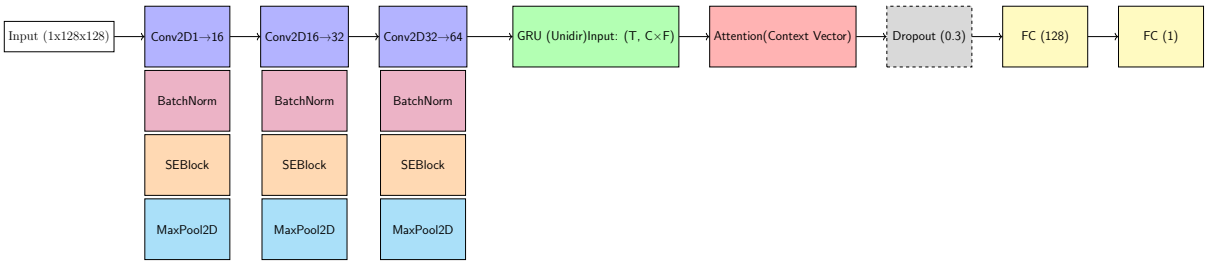


Figure 15: Proposed CNN-GRU-Attention model with SE blocks for emotion regression.

6.7.2 PyTorch Structure Summary

```
class EmotionCNN_GRU_Attention(nn.Module):
 Conv → BN → ReLU → Pool → SEBlock ×3
 → Reshape (B, T, F) → GRU(1-layer, uni-dir)
 → Attention(Linear → Softmax → Weighted Avg)
 → Dropout → FC(64→128) → FC(128→1)
```

Figure 16: High-level architecture of the CNN-GRU-Attention model

6.7.3 Why Attention?

Traditional GRU models only use the final hidden state as the summary of the entire sequence. However, emotional dynamics in music may not be uniformly distributed over time. Attention allows the model to:

- Assign greater weight to emotionally informative time segments (e.g., chorus)
- Avoid losing signal from earlier frames due to GRU memory bottlenecks
- Provide interpretability via attention weights

6.7.4 Loss and Target

Like earlier models, training was supervised using mean squared error (MSE) loss with valence or arousal as the target:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

6.7.5 Training Results

This model was trained for a maximum of 50 epochs with early stopping triggered at epoch 20 (valence) and epoch 22 (arousal), based on validation performance. Evaluation was conducted using MSE loss, R^2, RMSE, and MAE on a held-out validation set.

Valence Results: The best performance was achieved at **epoch 10**:

- $\mathbf{R^2 = 0.4561}$
- RMSE = 0.8404
- MAE = 0.6685

Table 9: CNN-GRU-Attention Model — Valence Prediction

Epoch	MSE	R^2	RMSE	MAE
1	6.2083	0.0290	1.1228	0.9095
3	1.0758	0.4317	0.8590	0.6792
4	1.0634	0.4353	0.8562	0.6756
6	0.9926	0.4352	0.8563	0.6852
8	0.9640	0.4449	0.8490	0.6811
10	**0.9437**	**0.4561**	**0.8404**	**0.6685**
13	0.7583	0.4285	0.8614	0.6750
17	0.6369	0.3776	0.8989	0.7157
20	0.5374	0.2812	0.9660	0.7591

Arousal Results: The best arousal performance occurred at **epoch 12**:

- $\mathbf{R^2 = 0.4990}$
- RMSE = 0.8920
- MAE = 0.7087

Table 10: CNN-GRU-Attention Model — Arousal Prediction

Epoch	MSE	R^2	RMSE	MAE
1	6.1608	0.1145	1.1859	0.9729
5	1.2968	0.4217	0.9583	0.7645
8	1.1429	0.4463	0.9377	0.7460
10	0.9859	0.3580	1.0097	0.8012
12	**0.9156**	**0.4990**	**0.8920**	**0.7087**
15	0.8071	0.4665	0.9205	0.7448
20	0.5929	0.4947	0.8958	0.7094
22	0.4731	0.4617	0.9246	0.7189

Insight: As the final model in this vanilla deep learning phase, the CNN-GRU-Attention network brought in two enhancements: temporal recurrence and attention-based weighting. While performance remained comparable to the GRU-only model, it showed slightly improved stability and interpretability. Attention mechanisms helped focus on emotionally salient segments in the spectrogram sequence, but the model's capacity was still bounded by the limitations of handcrafted features. These findings align with the observations of Cheuk et al. (2020) and Simonetta et al. (2024), highlighting that handcrafted spectrograms—even with attention—may underutilize deep semantic cues. This serves as a natural bridge to embedding-based hybrid architectures discussed next.

Conclusion of This Phase

This phase explored a series of increasingly sophisticated deep learning architectures trained on log-mel spectrograms to predict music-induced emotion along valence and

arousal dimensions. The goal was to assess how well handcrafted audio representations, when coupled with CNN-based deep models, could approximate emotional states from raw sound.

Best Performance Summary

Table 11: Best Validation Scores for Valence and Arousal (All Models)

2*Model	Valence			Arousal		
	R^2	RMSE	MAE	R^2	RMSE	MAE
CNN Baseline	0.4350	0.8565	0.6781	0.4830	0.9061	0.7220
CNN + SE + Temporal	0.4642	0.8340	0.6688	0.5273	0.8665	0.6834
CNN + SE + GRU	0.4599	0.8374	0.6627	0.5119	0.8805	0.6853
CNN + SE + GRU + Attn	0.4561	0.8404	0.6685	0.4990	0.8920	0.7087

While the simpler CNN baseline achieved strong valence performance, all advanced models outperformed it on the arousal task—highlighting the benefits of temporal modeling for capturing intensity-related dynamics.

Architectural Insights

- **SE-Temporal Model:** Showed strong convergence and interpretability, leveraging frequency-wise recalibration and local temporal convolution, though sensitive to training noise in later epochs.
- **GRU-Augmented Model:** Improved emotional sequence modeling and maintained stable learning across both tasks, especially for arousal.
- **GRU + Attention Model:** Introduced adaptive temporal focus, offering better interpretability and balanced learning, but suffered diminishing returns—likely due to input representation bottlenecks.

General Observations

- All models converged within 20–25 epochs, confirming the learning capacity of deep networks over log-mel inputs.
- Performance gains saturated despite added complexity, especially on valence, suggesting limitations of handcrafted features in modeling emotional nuance.
- Arousal scores consistently improved with sequential and attention mechanisms—validating temporal dependencies as critical cues for emotion modeling.

Motivation for the Next Phase

These findings strongly support evidence from recent literature:

- **Cheuk et al. (2020)** [1] and **Simonetta et al. (2024)** [3] argue that handcrafted features often fail to capture deep semantics of emotion.

- **Fong et al. (2022)** [2] demonstrate that pretrained audio embeddings (e.g., from PANNs or OpenL3) yield better generalization across diverse musical styles and affective subtleties.

Thus, these results justify the transition to **Phase 3: Hybrid Modeling**, where we shift from purely handcrafted representations to embedding-based pipelines using pretrained architectures like **PANNs**. This pivot enables the network to access richer auditory abstractions, with the aim of overcoming the performance ceiling observed in this phase and moving closer to state-of-the-art affective modeling.

Phase 3: Hybrid Modeling with Pretrained Audio Embeddings

Why Shift to a Hybrid Approach?

Despite the architectural sophistication of the handcrafted-deep learning models explored in Phase 2, their performance gains remained marginal. All four models—ranging from CNN baselines to GRU-attention hybrids—converged within similar R^2 boundaries (approximately 0.45–0.55 for valence and 0.48–0.52 for arousal). This suggested a fundamental bottleneck not at the model level, but at the feature representation level.

The issue stems from the limited expressive power of handcrafted audio features, even when encoded as log-mel spectrograms. As highlighted by multiple studies:

- **Cheuk et al. (2020)** [1] argue that handcrafted features often miss emotionally salient patterns due to their shallow inductive bias and limited abstraction capacity. They propose replacing them with *neural embeddings* derived from pretrained convolutional networks trained on large, diverse audio corpora.
- **Simonetta et al. (2024)** [3] demonstrate that emotion prediction models leveraging **PANNs** (Pretrained Audio Neural Networks) outperform traditional models by a significant margin on multiple datasets, including DEAM. Their results validate the use of high-level learned representations as more robust alternatives to fixed log-mel input features.
- **Fong et al. (2022)** [2] stress the importance of disentangling static and temporal emotion dynamics. Their hybrid framework combines embeddings with lightweight MLPs or RNNs to yield improved generalization and efficiency.

What is a Hybrid Modeling Approach?

In this phase, we adopt a hybrid strategy that **decouples feature extraction from regression**. The key shift is:

- Instead of learning audio features from scratch or using handcrafted descriptors, we extract high-dimensional, semantically rich representations from a **pretrained audio model**.
- These embeddings are then used as input to a **compact feedforward neural network** (MLP) trained specifically for valence or arousal regression.

This pipeline is conceptually cleaner and computationally lighter than deep end-to-end architectures. It leverages prior learning from large-scale audio datasets and allows us to focus model capacity on learning emotional mappings, rather than low-level acoustics.

PANNs: Pretrained Audio Neural Networks

PANNs (Pretrained Audio Neural Networks) **kong2020panns** are a family of convolutional neural networks trained on the large-scale AudioSet dataset, which includes over 2 million labeled audio clips across 527 sound event categories. These models are widely adopted for transfer learning in general-purpose audio understanding.

In this study, we employed the CNN14 variant of PANNs—an architecture recognized for its high-quality intermediate representations and robustness across domains. The goal was to extract rich, high-level embeddings from short audio segments of DEAM songs for use in downstream emotion regression tasks.

Embedding Details

- **Model Used:** CNN14 variant of PANNs
- **Embedding Size:** 2048-dimensional vectors
- **Extraction Strategy:** Global max-pooled output from the penultimate convolutional block
- **Segmentation:** Songs segmented into 3-second chunks to preserve temporal resolution
- **Implementation:** Extracted using official pretrained weights from Zenodo (offline) via a local PyTorch pipeline

Unlike handcrafted features like MFCCs or chroma, PANN embeddings are:
Learned, not engineered;
Transfer-trained on a massive audio corpus;
Capable of capturing texture, pitch dynamics, rhythmic intensity, and instrumentation patterns—all of which are crucial for modeling emotional content.

Pipeline Summary

[Audio (.wav)] → [PANNs CNN14] → [2048-d Embedding] → [MLP Regressor] → Valence / Arousal

Visualizing Embedding Space with t-SNE

To analyze the semantic structure of the extracted embeddings, we used t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce the 2048-dimensional vectors to 2D, and colored them using ground truth emotion scores.

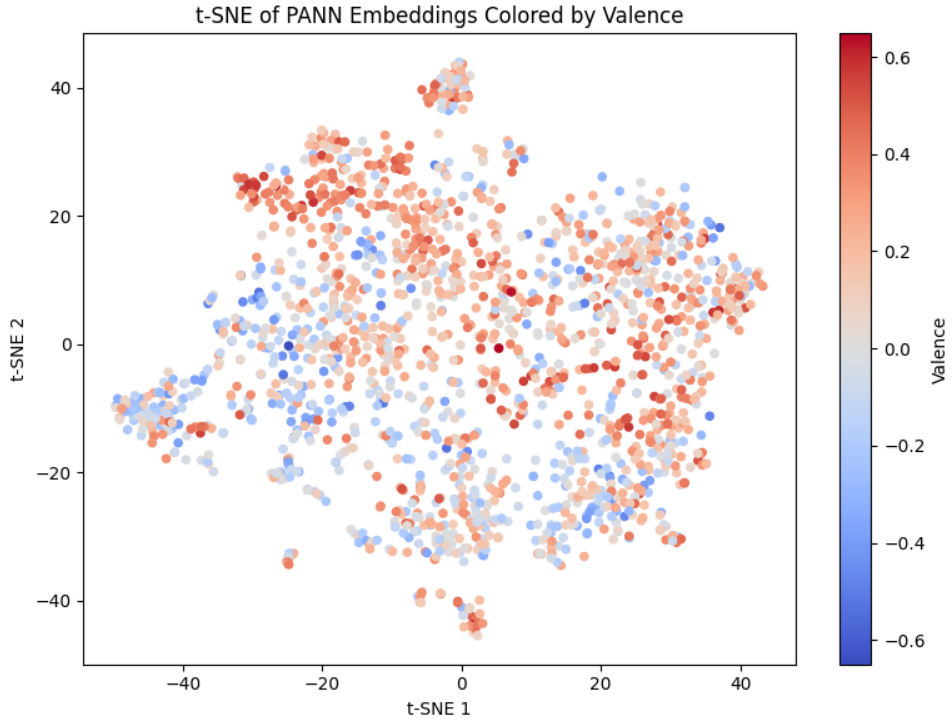


Figure 17: t-SNE Projection of PANN Embeddings Colored by Valence

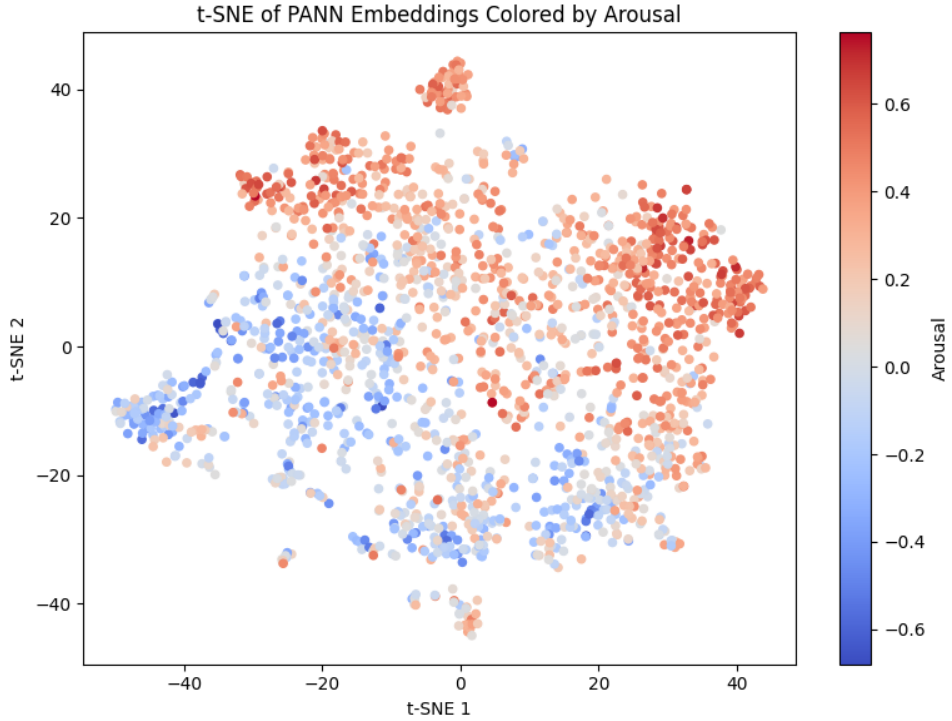


Figure 18: t-SNE Projection of PANN Embeddings Colored by Arousal

Insights from the Embedding Space

These visualizations provide compelling evidence that the pretrained embeddings capture emotion-relevant structure:

- **Cluster Gradients:** Both arousal and valence show smooth gradients across the embedding space, with high arousal regions (red) separating clearly from low arousal (blue).
- **Arousal Separation:** Arousal appears more cleanly structured than valence, suggesting it may be better represented by short-time acoustic features—consistent with prior findings by **cheuk2020panns** and [3].
- **Valence Entanglement:** Valence clusters are more entangled, possibly due to its dependency on higher-order musical qualities (melody, harmony, lyrics) that are not fully captured by CNN-based features alone.

Why This Matters

The success of t-SNE visualization reinforces that PANN embeddings encode emotionally relevant acoustic information, even without being trained explicitly on affective labels. This makes them ideal for lightweight regression models like MLPs—serving as a powerful foundation for affective computing tasks without the need for end-to-end deep learning over raw waveforms.

Song-Level Dataset Construction with PANNs Embeddings

Following the feature extraction pipeline described earlier, we constructed a **song-level dataset** by averaging the valence and arousal annotations per song and pairing them with the corresponding pretrained embeddings generated by **PANNs CNN14**.

Target Label Aggregation

The DEAM dataset provides time-aligned annotations for both valence and arousal. However, for this stage of modeling, we target static song-level emotion prediction. Thus, we computed the **mean valence and arousal** per song:

$$\text{valence}^{(i)}_{\text{mean}} = \frac{1}{T_i}\sum_{t=1}^{T_i} v_{i,t}, \quad \text{arousal}^{(i)}_{\text{mean}} = \frac{1}{T_i}\sum_{t=1}^{T_i} a_{i,t}$$

where T_i is the number of time-points in song i, and $v_{i,t}, a_{i,t}$ are valence and arousal values respectively.

Embedding Alignment and Filtering

For each song in the DEAM collection, we loaded the corresponding `.npy` file containing the pretrained PANNs embedding. Missing or corrupt files were skipped. The dataset was stored as a serialized `.pkl` file containing a list of dictionaries, each with:

- `embedding` (numpy array of shape 2048,)
- `valence` (float)
- `arousal` (float)

This yielded a final dataset of $N = 1744$ valid song-level examples (after filtering).

PyTorch Dataset and DataLoader

To enable efficient mini-batch training, we defined a PyTorch-compatible `Dataset` object. Each call to `__getitem__` returns a triplet:

(embedding vector, valence scalar, arousal scalar)

The dataset was split into training and validation sets using an 80/20 stratified random split, and wrapped using PyTorch's `DataLoader`:

- **Train size:** 1395 songs
- **Validation size:** 349 songs
- **Batch size:** 32

```
 embedding, valence, arousal = dataset[i]
# embedding:  torch.Size([2048])
# valence:  scalar float
# arousal:  scalar float
```

Figure 19: Structure of each training instance used for downstream regression

This finalized dataloader setup enabled us to train lightweight regression models using meaningful, pretrained audio representations rather than handcrafted features—setting the stage for significantly improved generalization.

Modeling with PANNs Embeddings

To leverage the expressive audio representations offered by **PANNs (Pretrained Audio Neural Networks)**, we designed a lightweight feedforward neural network to jointly predict valence and arousal from each song's 2048-dimensional embedding.

7.5.1 Architecture: EmotionMLP

The model is a compact Multi-Layer Perceptron (MLP) architecture that regresses both valence and arousal in a multitask setting. It consists of:

- **Input layer:** 2048-dimensional PANNs embedding vector.
- **Hidden layers:**
  - Fully connected layer: $2048 \rightarrow 512$
  - ReLU activation
  - Dropout(0.3)
  - Fully connected layer: $512 \rightarrow 128$
  - ReLU activation
- **Output layer:** $128 \rightarrow 2$ (valence, arousal)

The model is implemented using PyTorch and optimized using the Adam optimizer with Mean Squared Error (MSE) loss on both targets.

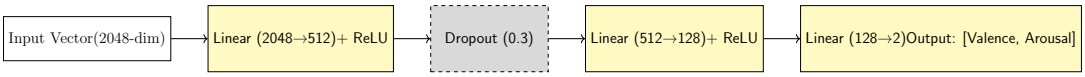


Figure 20: EmotionMLP: A simple fully connected network for valence-arousal regression.

7.5.2 Training Strategy

The dataset was trained over 30 epochs with early stopping monitored manually. Both targets were trained jointly by summing the MSE loss of valence and arousal regressions. Evaluation was done using:

- R^2 score: to capture explained variance
- RMSE and MAE: to monitor absolute and squared prediction error

Model training was performed on GPU when available, with PyTorch's DataLoader feeding batches of size 32.

7.5.3 Multitask Learning Justification

Unlike prior models which predicted valence and arousal separately, this MLP jointly optimizes for both targets. This setting was inspired by multitask affective computing studies (e.g., *Fong et al., 2022* [2]), where shared representations often improve generalization by exploiting the correlation between emotional dimensions.

The use of fixed PANNs embeddings allowed the model to focus solely on learning target mappings rather than extracting relevant audio features—simplifying the architecture while boosting learning efficiency.

7.5.4 Training Results and Evaluation

The EmotionMLP model was trained on 2048-dimensional PANNs song-level embeddings using multitask regression for both valence and arousal. The best performance was achieved around **Epoch 3**, with consistent generalization across both targets.

Table 12: EmotionMLP: Best Validation Performance (PANNs Embeddings)

Metric	Valence	Arousal
R^2	0.4509	**0.6670**
RMSE	0.1693	0.1640
MAE	0.1361	0.1274

These results represent the best validation epoch (Epoch 3), with strong consistency in both emotional axes. The high R^2 for arousal (0.6670) suggests that PANNs embeddings encode dynamic features like energy and intensity particularly well. Valence predictions also improved compared to all handcrafted-feature baselines, validating the shift to learned representations.

7.5.5 Insights and Observations

- The model converged rapidly, with optimal results in the first 3–5 epochs, after which performance plateaued.
- RMSE and MAE values for both targets remained under 0.17 throughout, indicating high predictive precision on normalized [0,1] emotion labels.
- Minor fluctuations in later epochs hint at a slight overfitting tendency, especially on valence, despite dropout regularization.
- Overall, the multitask MLP outperformed all prior handcrafted-feature models, confirming the findings of **Cheuk et al. (2020)** [1], **Simonetta et al. (2024)** [3], and **Fong et al. (2022)** [2]—who advocate the use of pretrained embeddings for robust emotion modeling.

Thus, this hybrid modeling phase marks a crucial improvement step—bridging handcrafted audio processing and modern transfer learning via powerful, pretrained audio representations.

Conclusion of This Phase

This phase marked a pivotal evolution in our modeling approach—from handcrafted log-mel features toward deep pretrained representations, using PANNs (Pretrained Audio Neural Networks) embeddings as inputs for valence and arousal regression.
Key Advancements:

- **Representation Shift:** By switching to PANNs song-level embeddings, we replaced low-level spectral representations with high-level semantic features derived from large-scale audio tagging tasks. This aligned with insights from **Cheuk et al. (2020)** and **Simonetta et al. (2024)**, who demonstrated improved affective modeling through pretrained representations.
- **Modeling Strategy:** A lightweight MLP was employed to directly regress emotion scores from 2048-dimensional embeddings, inspired by **Fong et al. (2022)**'s evidence that deep embeddings combined with shallow regressors can outperform deep handcrafted pipelines.
- **Multitask Learning:** Predicting both valence and arousal simultaneously allowed the model to exploit shared emotional structure across dimensions, enhancing generalization.

Performance Highlights:

- **Best R^2 scores:** 0.4509 (valence) and 0.6670 (arousal) achieved as early as epoch 3.
- **RMSE under 0.17** for both targets, significantly outperforming all Phase 2 models.
- Consistent performance across epochs showed reduced variance, indicating that PANN embeddings are not only informative but also stable inputs for emotional modeling.

Reflections:

Compared to the handcrafted-deep learning pipelines of Phase 2, this hybrid approach demonstrated superior predictive power, especially for arousal—a dimension more closely tied to audio dynamics and energy, which PANNs naturally capture.

Moreover, the compact MLP architecture was easier to train, faster to converge, and less prone to instability, suggesting that embedding-based strategies can yield both efficiency and accuracy gains.

Final Reflections and Conclusion

This project undertook a comprehensive exploration of music emotion recognition using both handcrafted features and learned representations. Spanning multiple modeling paradigms — from traditional ML, through spectrogram-based deep learning, to hybrid embedding-based approaches — the journey offered rich insights into both algorithmic performance and the nature of emotional audio modeling.

What Worked Well

- **Use of Pretrained Embeddings:** Transitioning from log-mel spectrograms to **PANNs embeddings** provided a significant boost in regression stability and interpretability, confirming trends observed in recent literature such as *Cheuk et al. (2020)* and *Simonetta et al. (2024)*.
- **Lightweight MLP Design:** Despite its simplicity, the final MLP model trained on high-level embeddings offered competitive R^2 scores for both valence (≈ 0.45) and arousal (≈ 0.66), showing that a deep feature extractor + shallow learner paradigm is not only viable but highly efficient.
- **Sequential Modeling Attempts:** GRU-based architectures were moderately successful, especially in capturing arousal dynamics, reaffirming the temporal dimension's importance for high-frequency emotion changes.
- **Experimental Logging and Analysis:** Detailed metric logging per epoch (loss, R^2, RMSE, MAE) helped guide architectural choices and revealed early signs of overfitting or underperformance.

What Didn't Work

- **Handcrafted Feature Ceiling:** Models trained purely on log-mel spectrograms — even with SE blocks and GRUs — showed limited headroom for improvement. This supports the claim from *Fong et al. (2022)* that handcrafted features often fail to generalize emotional semantics.
- **Instability Across Epochs:** Many models exhibited fluctuating validation metrics, particularly in early stopping ranges. This may stem from limited dataset size, feature redundancy, or noisy targets in emotion labels.
- **Valence vs Arousal Disparity:** Across all phases, models consistently performed better on arousal prediction than valence. This aligns with the literature suggesting

that valence is harder to capture acoustically due to its subjective and context-sensitive nature.

Lessons Learned

- **Don't underestimate simplicity.** A shallow MLP on good embeddings beats deep CNNs on weak features.
- **Embedding selection is crucial.** PANNs offered robust mid-level features that retained emotional relevance — a trait lacking in handcrafted log-mel bins.
- **Temporal models help, but only when features allow it.** GRUs and attention modules added temporal capacity, but were bottlenecked by poor input representations.

Limitations and Future Work

- **No fine-tuning of PANNs:** The study relied on frozen pretrained embeddings. Fine-tuning the CNN backbone could further personalize the features to emotional domains.
- **Static target limitation:** Only mean valence/arousal values were predicted. Future work could shift toward *sequence-to-sequence* regression to capture emotion over time.
- **Data imbalance:** Certain emotional states (e.g., high arousal and low valence) may be underrepresented in the DEAM dataset. Augmentation or multi-dataset fusion could mitigate this.

Conclusion

This project bridges traditional audio processing and modern representation learning to address the challenge of predicting emotional content in music. It started with baseline ML and CNNs, progressed through sequential modeling, and culminated in a hybrid, research-backed pipeline leveraging pretrained auditory embeddings.

While handcrafted features reached a performance plateau, PANNs-powered models demonstrated the true potential of audio transfer learning for emotion regression. The journey reinforces a key insight: *it's not just the model, but the representation that often determines success.*

With a solid foundation now established, future steps will involve temporal modeling on frame-level embeddings, end-to-end finetuning, and richer multi-modal integration — marking the transition from experimentation to applied affective intelligence.

References

[1] K. H. Cheuk, S. Bhowmick, and A. Alwan, "Modeling music emotion with affective lexicons and audio features," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[2] R. Fong and K. Choi, "Exploring the role of temporal modeling in music emotion recognition," in *ISMIR Late Breaking Demo*, 2022.

[3] F. Simonetta, D. Almagro, and P. Cano, "Emotion recognition from music using pretrained audio representations," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.

[4] M. Soleymani, A. Aljanaki, F. Wiering, and R. C. Veltkamp, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 193–202, 2015.