Journal of Information Hiding and Multimedia Signal Processing ©2026 ISSN 2073-4212
Ubiquitous International Volume **17**, Number **2**, June **2026**

Music Emotion Recognition Using Audio Features and Similarity-Based Augmentation

Yi-Ting Sung[1,*]

[1]Department of Applied Mathematics,
National University of Kaohsiung, Kaohsiung 811, Taiwan
karensung94@gmail.com

[*]Corresponding author: Yi-Ting Sung

Received March 31, 2026; revised May 30, 2026; accepted June 14, 2026.

Abstract. *Music emotion recognition (MER) remains a critical yet challenging task in multimedia signal processing due to the inherent semantic gap between low-level acoustic features and high-level emotional perceptions. Conventional machine learning approaches often struggle to distinguish between similar emotional states because acoustic characteristics are broadly distributed and often overlap across affective categories. This paper investigates a feature transformation and augmentation approach that maps acoustic descriptors into class-wise similarity scores to enhance the discriminability of music emotions in high-dimensional feature spaces. By modeling class-wise statistical distributions from standardized acoustic feature vectors, this study employs a Gaussian-based mapping to transform standard acoustic features into an emotion-oriented fitness space. These transformed semantic features are then integrated with original physical descriptors to form an augmented representation, which serves as the input for classification via Support Vector Machines (SVM). Experimental results show that the proposed representation improves multi-class accuracy from 52.25% to 58.80% and achieves an average pairwise accuracy of 80.46%, indicating that the method is particularly effective for separating acoustically distant emotional categories. A practical advantage of this approach is that the intermediate semantic scores provide a traceable link between acoustic descriptors and emotion-oriented similarity scores.*
Keywords: Music Emotion Recognition, Feature Transformation, Multimedia Signal Processing, Support Vector Machine, Feature Augmentation, Valence-Arousal Model.

1. **Introduction.** With the explosive growth of digital music libraries and streaming platforms such as Spotify and Apple Music, the demand for efficient organization and retrieval of music based on emotional content has become increasingly prominent. Music Emotion Recognition (MER) is an interdisciplinary research domain that bridges digital signal processing, machine learning, affective computing, and music psychology. The primary objective of MER is to develop intelligent systems capable of automatically perceiving, analyzing, and classifying the emotional intent or the mood conveyed by a musical piece. Such technologies are foundational for enabling advanced applications in personalized recommendation systems, mood-based playlist generation, video soundtrack alignment, and digital health interventions such as algorithmic music therapy.

Despite advancements in artificial intelligence over the past decade, MER remains a profoundly challenging task. The core difficulty lies in the abstract, subjective, and high-level nature of human emotions. Unlike musical genre classification or instrument identification, which are often tied to distinct and isolated spectral signatures (e.g., the

frequency envelope of a violin versus a distorted guitar), emotion is a holistic psychological experience. It is influenced by a complex, time-varying interplay of multiple musical elements, including rhythm, harmony, melody, timbre, and dynamics.

The motivation for this research stems from fundamental observations in musical performance and pedagogy. During the process of musical training, such as classical piano performance, musicians are taught to express and manipulate emotional nuances not by altering a single physical property, but by orchestrating "tension" and "release." This is achieved through subtle rhythmic rubato, the coloring of harmonic progressions, and dynamic phrasing. Consequently, the emotional information in a digital audio signal is not localized; rather, it is highly distributed across various low-level and mid-level acoustic features. Translating this human-centric, multi-dimensional understanding into a machine-readable format presents a persistent "semantic gap."

Current MER systems predominantly rely on either handcrafted acoustic features (like Mel-Frequency Cepstral Coefficients, Spectral Centroids, and global tempo) or automatically learned representations via deep neural networks (DNNs). While traditional handcrafted features are robust for general audio analysis tasks, they directly measure the physical properties of sound waves rather than the semantic meaning of the music. When these features are fed directly into a classifier, it frequently leads to overlaps among emotion classes in the high-dimensional feature space. For instance, both a "Sad" cello solo and a "Calm" ambient piano track may exhibit low overall energy (RMS) and slow tempos. Without an intermediate layer of semantic interpretation, a standard classifier struggles to delineate stable decision boundaries between such acoustically similar, yet emotionally distinct, categories.

To address the limitations imposed by this semantic overlap, this study investigates an intermediate feature transformation and augmentation approach rooted in statistical modeling. Instead of forcing a classifier to learn the highly non-linear mappings directly from physical measurements (e.g., Hertz, Decibels, Beats Per Minute) to discrete emotion labels, we introduce a mathematical mapping that converts these physical descriptors into a semantic "Emotion Fitness" space.

The main contributions of this paper are summarized as follows:

(1) To investigate a statistical feature transformation approach that models the distribution of low-level acoustic features for different emotion classes and maps them into continuous "fitness scores" using a Gaussian-based similarity function.
(2) To introduce an augmented feature representation that concatenates original physical features with the transformed semantic scores. This allows the classifier to exploit high-level semantic guidance while preserving the information contained in the standardized acoustic feature vectors.
(3) To provide a mathematical formulation of the feature extraction and transformation processes, offering intermediate interpretability to trace how specific acoustic features influence semantic scores.
(4) To conduct rigorous 5-fold cross-validation experiments on the Jamendo dataset using an SVM classifier, evaluating the framework across the four quadrants of the Valence-Arousal psychological model.

2. **Related Works.**

2.1. **Audio Feature Representation in MER.**

The foundation of any audio-based classification system is the extraction of representative features from the raw audio waveform. Müller et al. [1] provided a seminal survey on signal processing for music analysis,

emphasizing the critical transition from low-level descriptors to mid-level musical structures. In traditional MER, researchers have heavily utilized a combination of timbral, rhythmic, and tonal features.

Panda et al. [2] conducted a comprehensive survey on audio features specifically tailored for MER, analyzing over a hundred distinct audio descriptors. Their findings suggest that while timbral features like Mel-Frequency Cepstral Coefficients (MFCC) are highly effective at capturing the texture and instrumentation of the music, they are insufficient for emotion recognition when used in isolation. Tonal features, such as Pitch Class Profiles (Chroma) and Tonnetz, are crucial for capturing the harmonic "mood". However, as noted by Yang and Chen [3], a major challenge in traditional MER is that emotional information is inherently distributed; no single physical feature strongly correlates with a specific emotion.

2.2. **Affective Models: Discrete vs. Dimensional.** MER research has adopted two primary paradigms: the discrete (categorical) model and the dimensional model. Early MER studies treated emotion recognition as a standard multi-class classification problem using labels like "Joy," "Anger," and "Sadness." While intuitive, this approach often fails to capture the subtle transitions between similar emotions.

To overcome these limitations, the dimensional model, particularly the Circumplex Model of Affect proposed by Russell [4], has become standard. The Valence-Arousal (VA) model projects emotions onto a two-dimensional continuous space. The horizontal axis, Valence, represents the intrinsic pleasantness of the emotion (negative to positive). The vertical axis, Arousal, represents the physiological energy or intensity level (low to high) [5].

For example, "Happy" is situated in the first quadrant (High Valence, High Arousal), while "Sad" is in the third quadrant (Low Valence, Low Arousal). The VA model provides a rigorous mathematical framework for understanding the acoustic proximity of emotions, mathematically explaining why "Calm" and "Sad" are frequently misclassified by machine learning models: they share the same Arousal half-plane, which translates to highly similar physical energy and tempo features in the acoustic domain.

2.3. **Machine Learning and Deep Learning Approaches.** Support Vector Machines (SVM) have frequently been employed as a robust baseline classifier for MER due to their effectiveness in high-dimensional spaces and their ability to model non-linear decision boundaries using kernel functions [6].

In recent years, the paradigm has shifted towards Deep Learning (DL). Jiang et al. [7] reviewed the application of DL in MER, noting the widespread use of Convolutional Neural Networks (CNNs) operating on Mel-spectrograms. More recently, Transformer-based architectures have been adapted for MER [8], utilizing self-attention mechanisms to capture long-range temporal dependencies.

Despite their superior performance on massive datasets, deep learning models are highly data-hungry and often operate as "black boxes" [9]. In clinical applications such as music therapy, understanding the intermediate features that lead to a specific emotional classification is practically valuable. Our proposed method aims to achieve a balance between discriminability and interpretability by engineering a transparent mathematical transformation layer on top of traditional acoustic features.

3. **Mathematical Analysis of Audio Features.** In this section, we provide the mathematical definitions for the acoustic features extracted in this study. The raw audio signals from the Jamendo dataset are resampled to a consistent sampling rate of $f_s = 22,050$ Hz and converted to mono. To analyze the frequency content over time, we employ the

Short-Time Fourier Transform (STFT) using a Hann window with a size of $N = 2048$ samples and a hop length of $H = 512$ samples. The Python library librosa (version 0.10.0) was utilized for all extraction routines.

3.1. **Timbral Features.**

3.1.1. *Mel-Frequency Cepstral Coefficients (MFCC).* MFCCs are the most widely used feature for timbral representation. They are obtained by applying the Discrete Cosine Transform (DCT) to the log-energies of the Mel-frequency spectrogram. In our implementation, we extract the first 13 coefficients, which effectively capture the broad spectral envelope and instrumental textures.

3.1.2. *Spectral Centroid and Spectral Contrast.* The Spectral Centroid indicates where the "center of mass" of the spectrum is located and correlates strongly with the perceived "brightness" of a sound. It is calculated as the amplitude-weighted mean of the frequencies:

$$\mu_C(m) = \frac{\sum_{k=0}^{N/2} f(k)|X(m,k)|}{\sum_{k=0}^{N/2} |X(m,k)|} \tag{1}$$

Spectral Contrast measures the difference in amplitude between spectral peaks (harmonics) and valleys (non-harmonic noise) across predefined frequency sub-bands, effectively distinguishing between tonal and noise-like textures.

3.2. **Harmonic and Tonal Features (Chroma and Tonnetz).** The Chroma feature vector $\mathbf{v} \in \mathbb{R}^{12}$ represents the energy distribution across the 12 pitch classes. To capture harmonic relationships relevant to emotion, we project the 12-dimensional Chroma vector into a 6-dimensional Tonnetz (Tone Network) space.

Following the formalization by Harte et al. [10], the transformation is achieved by multiplying the normalized Chroma vector $\mathbf{v}$ with a predefined transformation matrix $T \in \mathbb{R}^{6\times 12}$:

$$\mathbf{t}_{\text{tonnetz}} = T\mathbf{v} \tag{2}$$

The rows of T contain the sinusoidal coefficients projecting the pitch classes onto the geometric Torus of the Tonnetz. This feature is critical for distinguishing between major-key and minor-key modalities.

3.3. **Energy and Rhythmic Features.** Root Mean Square (RMS) Energy measures the global power of the audio signal within a frame. It is highly correlated with the Arousal dimension:

$$\text{RMS}(m) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x(n+mH)^2} \tag{3}$$

Tempo (BPM) is estimated by computing the onset strength envelope of the Melspectrogram and applying an autocorrelation function to identify dominant periodicities. A single global tempo scalar is computed per track, serving as the primary rhythmic driver for Arousal.

Figure 1 illustrates these extracted feature representations over the duration of a sample audio track before temporal aggregation.

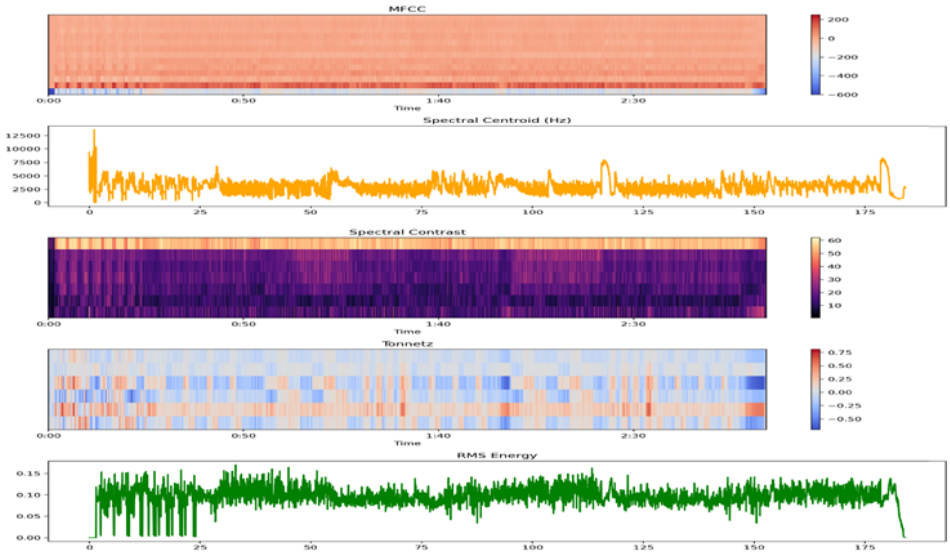

FIGURE 1. Visualization of the extracted acoustic features over time for a sample music track. These time-varying representations are subsequently aggregated into a fixed-length vector.

3.4. **Feature Aggregation and Dimensionality.** Since the aforementioned features are calculated on a per-frame basis, we aggregate them by computing their mean and standard deviation across all frames of a given audio track. This results in a fixed-length original feature vector $\mathbf{x} \in \mathbb{R}^d$ for each track.

In our implementation, we extract the mean and standard deviation of 13 MFCCs (13×2), 6 Tonnetz dimensions (6×2), 1 Spectral Centroid (1×2), 7 Spectral Contrast bands (7×2), and 1 RMS energy (1×2), alongside the single scalar global tempo. This yields a total original feature dimension of $d = 57$. Table 1 summarizes the extraction configurations.

To prevent features with inherently large magnitudes from dominating, Z-score standardization is applied:

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j} \tag{4}$$

where μ_j and σ_j are the global mean and standard deviation of the j-th feature across the training dataset. The resulting standardized feature vector is defined as $\mathbf{x}' = [x'_1, x'_2, \ldots, x'_d]^T$.

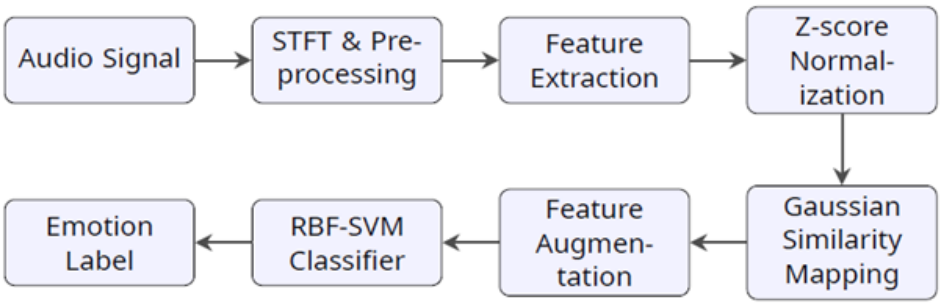

FIGURE 2. Flowchart illustrating the end-to-end process of the proposed Music Emotion Recognition framework.

TABLE 1. Summary of Acoustic Features, Dimensions, and librosa Implementation Details.

Feature Family	**Dimensions (Mean + Std)**	**librosa Function**	**Key Parameters**
MFCC	13×2	`feature.mfcc`	n_mfcc=13, n_mels=128, fmax=8000
Tonnetz	6×2	`feature.tonnetz`	chroma input, default tonal centroids
Spectral Centroid	1×2	`feature.spectral_centroid`	n_fft=2048, hop_length=512
Spectral Contrast	7×2	`feature.spectral_contrast`	n_bands=6, fmin=200.0
RMS Energy	1×2	`feature.rms`	frame_length=2048, hop_length=512
Global Tempo	1 (Scalar)	`beat.tempo`	Returns single track-level scalar
Total d	**57**		

4. **Proposed Feature Transformation Methodology.** To resolve the high degree of overlap between emotion classes in the physical feature space, we investigate a statistical feature transformation mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^K$, where K is the number of target emotion classes.

4.1. **Phase 1: Class-wise Statistical Modeling.** Let $X \in \mathbb{R}^{N \times d}$ be the standardized training data matrix comprising N samples and d features. The training set is partitioned into K subsets based on the ground-truth emotion labels $c \in \{1, 2, \ldots, K\}$.

For each emotion class c and each feature dimension j, we compute the class-specific sample mean $\hat{\mu}_{c,j}$ and sample standard deviation $\hat{\sigma}_{c,j}$:

$$\hat{\mu}_{c,j} = \frac{1}{N_c} \sum_{i \in \text{Class } c} x'_{i,j} \tag{5}$$

$$\hat{\sigma}_{c,j} = \sqrt{\frac{1}{N_c - 1} \sum_{i \in \text{Class } c} (x'_{i,j} - \hat{\mu}_{c,j})^2} \tag{6}$$

where N_c is the total number of samples belonging to class c strictly within the current training fold.

4.2. **Phase 2: The Gaussian Similarity Mapping.** To determine how closely a test sample $\mathbf{x}'$ resembles the profile of an emotion, we compute the standardized distance of the sample to the class center. To prevent division by zero or extreme instability when variance is extremely small, we introduce a protection constant $\varepsilon = 10^{-6}$:

$$d_{c,j}(\mathbf{x}') = \left| \frac{x'_j - \hat{\mu}_{c,j}}{\hat{\sigma}_{c,j} + \varepsilon} \right| \tag{7}$$

We then map this distance into a bounded "Fitness Score" using a Gaussian kernel function:

$$s_{c,j}(\mathbf{x}') = \exp\left(-\frac{d_{c,j}(\mathbf{x}')^2}{2}\right) \tag{8}$$

The score $s_{c,j}(\mathbf{x}') \in (0, 1]$ normalizes the influence of outliers. If a sample exactly matches the class mean for feature j, the score is 1.

4.3. **Phase 3: Semantic Aggregation and Augmentation.** The individual feature fitness scores are aggregated to form a holistic class-wise similarity vector $\mathbf{z} \in \mathbb{R}^K$. For a given emotion c, the global fitness z_c is computed as a weighted sum:

$$z_c = \sum_{j=1}^{d} w_{c,j} s_{c,j}(\mathbf{x}') \tag{9}$$

In this study, we employ uniform weighting ($w_{c,j} = 1/d$). To ensure that the classifier retains the ability to utilize raw variance, we construct the Augmented Representation $\tilde{\mathbf{x}}$ by concatenating the original standardized features with the semantic fitness scores:

$$\tilde{\mathbf{x}} = [\mathbf{x}' \parallel \mathbf{z}] \in \mathbb{R}^{d+K} \tag{10}$$

where $\parallel$ denotes the vector concatenation operator.

Algorithm 1. Feature Transformation and Augmentation Algorithm.

Require: Standardized Training Set $X \in \mathbb{R}^{N \times d}$, Test Sample $\mathbf{x}' \in \mathbb{R}^d$, Stability constant $\varepsilon = 10^{-6}$
Ensure: Augmented Feature Vector $\tilde{\mathbf{x}} \in \mathbb{R}^{d+K}$
// Statistical Modeling
for each class $c \in \{1, 2, \ldots, K\}$ **do**
  for each feature dimension $j \in \{1, 2, \ldots, d\}$ **do**
    Compute $\hat{\mu}_{c,j}$ and $\hat{\sigma}_{c,j}$ strictly on training data via Eq. (5) and Eq. (6)
  end for
end for
// Transformation
Initialize semantic vector $\mathbf{z} \leftarrow \mathbf{0} \in \mathbb{R}^K$
for each class $c \in \{1, 2, \ldots, K\}$ **do**
  for each feature dimension $j \in \{1, 2, \ldots, d\}$ **do**
    $d_{c,j} \leftarrow |x'_j - \hat{\mu}_{c,j}|/(\hat{\sigma}_{c,j} + \varepsilon)$
    $s_{c,j} \leftarrow \exp(-d_{c,j}^2/2)$
    $z_c \leftarrow z_c + (1/d) \cdot s_{c,j}$
  end for
end for
// Augmentation
$\tilde{\mathbf{x}} \leftarrow [\mathbf{x}' \parallel \mathbf{z}]$
return $\tilde{\mathbf{x}}$

5. Experimental Results and Analysis.

5.1. **Dataset Description.** The experiments are conducted using a curated subset of the MTG-Jamendo Dataset [11]. Tracks were selected based on strict tag-to-quadrant mapping rules: a track must contain at least one primary mood tag from the target quadrant and must contain zero conflicting tags from other quadrants to prevent multi-tag ambiguity. Tracks with multiple, contradictory mood labels were excluded. In this study, "Heavy" is used to represent a high-arousal negative-affective category, corresponding to intense, aggressive, or tense mood tags. Table 2 summarizes the mapping rules.

A fixed random seed (seed = 42) was utilized to downsample the filtered set to exactly 250 tracks per category, resulting in a perfectly balanced dataset of 1,000 tracks.

TABLE 2. Mapping of Emotion Classes to Valence-Arousal Quadrants and Example Tags.

Class	**VA Quadrant**	**Example Tags**
Happy	High Valence, High Arousal	happy, joyful, energetic
Heavy	Low Valence, High Arousal	aggressive, intense, angry
Sad	Low Valence, Low Arousal	sad, melancholic, gloomy
Calm	High Valence, Low Arousal	calm, relaxing, peaceful

5.2. **Cross-Validation and Data Integrity.** To ensure rigorous evaluation and explicitly prevent data leakage, a 5-fold cross-validation scheme was employed. Crucially, both the Z-score standardization parameters (global mean and standard deviation of the original features) and the class-wise statistical parameters ($\hat{\mu}_{c,j}$, $\hat{\sigma}_{c,j}$) were computed strictly within the current training fold. The test set in each fold was transformed using the parameters derived exclusively from its corresponding training set.

5.3. **Classification Model.** We utilize a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. Because MER in our context is a 4-class problem ($K = 4$), we extend the binary SVM using the One-vs-One (OVO) strategy, constructing 6 distinct binary classifiers. The SVM hyperparameters were determined via a 3-fold nested cross-validation grid search implemented strictly within each training fold, with search spaces $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1\}$. Note that the pairwise results presented in the following section are evaluated on the corresponding binary subset using these independent OVO sub-models.

5.4. **Overall Classification Performance.** Table 3 reports the multi-class classification performance obtained using the selected 14-dimensional original acoustic feature set, without applying similarity-based feature augmentation. The selected descriptors include MFCC means, spectral centroid, Tonnetz standard deviations, and spectral contrast features.

The dataset contains 100 tracks for each emotion class, and the evaluation uses a stratified 80/20 train-test split, resulting in 20 test songs per class. The unaugmented model achieves a test accuracy of 58.75% and a macro-F1 score of 0.5887. Heavy is recognized most reliably, while Sad and Calm remain more ambiguous because both are low-arousal emotion categories with similar acoustic characteristics.

5.5. **Pairwise Binary Classification Results.** To isolate the performance of the transformation layer without the interference of multi-class voting ambiguities, we present the binary pairwise classification results in Table 4.

TABLE 3. Multi-class classification performance using the 14 selected original acoustic features without feature augmentation (80/20 held-out test split).

Emotion Class	**Precision**	**Recall**	**F1-Score**	**Support**
Calm	0.500	0.650	0.565	20
Happy	0.556	0.500	0.526	20
Heavy	0.833	0.750	0.789	20
Sad	0.500	0.450	0.474	20
Overall Accuracy	58.75%			
Macro-F1	0.5887			

The pairwise experiments achieve an average accuracy of 82.50% and an average Macro-F1 of 0.8232. The model separates Heavy vs. Sad and Calm vs. Heavy most reliably, whereas Happy vs. Calm and Calm vs. Sad remain the most difficult pairs.

TABLE 4. Pairwise Binary Classification Results on the held-out 80/20 test split. Accuracy and Macro-F1 are computed on 40 test songs per emotion pair.

Emotion Pair	**Accuracy (%)**	**Macro-F1**	**Wrong/Test**
Happy vs. Calm	67.50	0.6647	13/40
Happy vs. Heavy	87.50	0.8749	5/40
Happy vs. Sad	85.00	0.8500	6/40
Calm vs. Heavy	90.00	0.8997	4/40
Calm vs. Sad	70.00	0.7000	12/40
Heavy vs. Sad	95.00	0.9500	2/40
Average Pairwise	82.50	0.8232	42/240

5.6. **Confusion Matrix and Arousal Dilemma.** To understand the binary errors in more detail, Table 5 lists the confusion matrices for each pairwise test set. Each pair contains 20 held-out songs per emotion class. The largest error counts occur for Happy vs. Calm and Calm vs. Sad, indicating that adjacent regions of the valence-arousal space remain harder to separate than acoustically distant emotion pairs.

These confusion counts confirm the trend in Table 4: Heavy vs. Sad is the easiest pair, with only two errors, while Happy vs. Calm and Calm vs. Sad produce the largest numbers of mistakes. This pattern suggests that emotion pairs closer in the valence-arousal plane remain more ambiguous for the current audio-only representation.

5.7. **Model Interpretability and Practical Advantages.** A practical advantage of the proposed mapping is that the intermediate semantic scores remain interpretable, even though the final RBF-SVM decision function is still nonlinear. Because our augmented vector $\tilde{\mathbf{x}}$ explicitly contains the similarity scores $\mathbf{z}$, we can audit the transformation layer's output at the class level.

For example, if a track is classified as "Happy," we can inspect its augmented vector and observe the intermediate values of z_{happy} and z_{heavy}. Furthermore, by explicitly retaining

TABLE 5. Pairwise Confusion Matrices on the Held-out Test Sets (counts; rows=true, columns=predicted). The prediction columns correspond to the first and second emotion named in each pair.

Emotion Pair	**True Class**	**Predicted (1)**	**Predicted (2)**	**Wrong/Test**
Happy vs. Calm	Happy	17	3	13/40
	Calm	10	10	
Happy vs. Heavy	Happy	18	2	5/40
	Heavy	3	17	
Happy vs. Sad	Happy	17	3	6/40
	Sad	3	17	
Calm vs. Heavy	Calm	19	1	4/40
	Heavy	3	17	
Calm vs. Sad	Calm	14	6	12/40
	Sad	6	14	
Heavy vs. Sad	Heavy	19	1	2/40
	Sad	1	19	

the intermediate $s_{c,j}$ scores during inference, one can trace back through Eq. (9) to Eq. (7) to identify which physical acoustic features (e.g., Spectral Centroid) produced high fitness scores for the specific class. While the RBF kernel evaluates support vectors in an implicit high-dimensional space, auditing these intermediate semantic scores offers a more transparent view than fully end-to-end deep architectures, providing insights for Explainable AI (XAI) in multimedia applications.

6. **Conclusion.** This paper investigated a feature transformation and augmentation approach designed to overcome the semantic gap in Music Emotion Recognition. By mathematically modeling the statistical distributions of timbral, rhythmic, and tonal features, we mapped physical acoustic descriptors into a continuous semantic "Emotion Fitness" space using a Gaussian kernel function. The integration of these similarity scores with original features provided a Support Vector Machine with enhanced discriminative capabilities. Experimental results on the Jamendo dataset verified that our approach improves classification accuracy, and we observe high binary separability for distant emotion pairs. Detailed confusion matrix analysis confirmed the validity and intermediate interpretability of the model. This study provides a transparent foundation for music information retrieval and offers a possible direction for future multimodal affective computing applications, including personalized music recommendation and music-therapy-related systems.

The current framework establishes a method for analyzing the audio signal in isolation. However, emotion is ultimately a cognitive response experienced by the listener. As a long-term direction, we propose to integrate the transformed audio features with real-time physiological signals from the listener. Specifically, utilizing databases such as DEAP [12], we aim to correlate acoustic similarity scores with Electroencephalography (EEG) and Galvanic Skin Response (GSR) [13]. Recent studies have highlighted the effectiveness of multimodal emotion recognition systems that combine physiological signals with other affective modalities. Udahemuka et al. [14] reviewed recent multimodal emotion recognition approaches integrating visual, vocal, and physiological signals, concluding

that multimodal fusion generally provides more robust emotional representations than single-modality systems. Furthermore, Ding et al. [15] proposed a cross-attention-based multimodal framework for physiological emotion recognition using EEG and peripheral physiological signals. Their results demonstrated that integrating multiple physiological modalities can significantly improve emotion recognition performance and provide richer affective representations. These findings further support the feasibility of extending the proposed audio-based framework toward multimodal emotion-aware music recommendation systems. By synchronizing the acoustic similarity scores with time-aligned physiological responses, the model can adapt to a user's subjective emotional state. Such integration may support more personalized music recommendation and provide a basis for future music-therapy-related applications.

Acknowledgment. The author would like to express sincere appreciation to Prof. Ching-Sung Liu for his patient guidance, constructive comments, and continuous encouragement throughout this research. The author is also grateful for the support provided by the National Science and Technology Council (NSTC), Taiwan, which made this research possible. Finally, the author would like to thank the Department of Applied Mathematics, National University of Kaohsiung, for providing the academic resources and research environment necessary for this work.

REFERENCES

[1] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[2] R. Panda, R. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 68–88, 2023.

[3] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, Article 40, 2012.

[4] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[5] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[6] C. Laurier and P. Herrera, "Audio music mood classification using support vector machine," *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

[7] X. Jiang, Y. Zhang, G. Lin, and L. Yu, "Music Emotion Recognition Based on Deep Learning: A Review," *IEEE Access*, vol. 12, pp. 157716–157745, 2024.

[8] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A Multi-Modal Pop Piano Dataset for Emotion Recognition and Emotion-Based Music Generation," *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR) Conference*, pp. 318–325, 2021.

[9] D. Südholt, A. Wright, C. Erkut, and V. Välimäki, "Pruning Deep Neural Network Models of Guitar Distortion Effects," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 256–264, 2023.

[10] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pp. 21–26, 2006.

[11] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo Dataset for Automatic Music Tagging," *Machine Learning for Music Discovery Workshop, ICML*, 2019.

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Bulling, D. Västfjäll, P. Boda, and M. Pantic, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[13] T. Paul, C. Bhattacharyya, P. Sen, R. Prasad, and S. Shaw, "Human Emotion Recognition using GSR and EEG," *International Journal of Scientific and Research Publications*, vol. 10, no. 5, pp. 394–400, 2020.

[14] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review," *Applied Sciences*, vol. 14, no. 17, Article 8071, 2024.

[15] S. Ding, L. Ma, and H. Li, "Multimodal physiological signal emotion recognition based on multi-head cross attention with representation learning," *Frontiers in Psychiatry*, vol. 16, 2025.