Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 10 Issue 02
2026

*Research Article*
https://doi.org/10.56979/1002/2026/1244

# Deep Learning–Based Emotion Classification Models for Chinese and Korean OST Music

**Quanrui Lu[1], and Hyuntai Kim[1*]**

[1]Department of Music, Faculty of Arts and Physical Education, Sejong University, Seoul, 0500, Korea.
[*]Corresponding Author: Hyuntai Kim. Email: kimht@sejong.ac.kr

**Received:** November 03, 2025 **Accepted:** January 23, 2026

**Abstract:** Music Emotion Recognition (MER) has made significant advancements with deep learning, however, existing models tend to have cultural bias wherein they are not good at recognizing the emotion of non-Western musical structures. This paper proposes a deep learning framework designed especially for the emotion classification in Chinese and Korean Original Soundtracks (OSTs), which have unique tonal dynamics and a high variance in emotions. We propose a Dual-Stream Convolutional Recurrent Neural Network (CRNN) with Self-Attention, which is able to capture the spectral spatial characteristics and the temporal melodic developments, commonly found in Asian cinematic music. To validate the model, we use two region-specific datasets namely PMEmo (Chinese popular music) and EMOPIA (Korean/Asian piano OSTs). Experimental results show that our proposed architecture can obtain an accuracy of 88.4% and F1-score of 0.87, which outperforms baseline models (ResNet-50 and standard LSTM) with 5.2% margin. The research helps to confirm that the training data for culturally-aware training is vital for accurate affective computing within the music domain.

**Keywords:** Music Emotion Recognition (MER); Deep Learning; Chinese OST; Korean OST; PMEmo; EMOPIA; Attention Mechanism

## 1. Introduction

Music Emotion Recognition (MER) has become a major part of contemporary affective computing due to the explosive rise of digital music streaming services and the need for individual content recommendation systems [1]. As music is a universal language for human emotional expression, the ability to automatically classify music tracks into emotional categories has important applications in music therapy and smart entertainment [2]. While the early methods have used handcrafted features, recent developments in Deep Learning (DL) have made a major shift in the field enabling the learning of complex representations directly from the spectrograms [3].

Despite these advancements, there is a large gap in the cultural adaptability of current MER models [4]. Most frameworks are trained on Western-centric datasets which give priority to Western tonal structures. But music from Asian culture, in particular Chinese and Korean Original Soundtrack (OST) have its unique acoustic features and it has been combined with traditional instruments and contemporary orchestras and music that is based on a unique 5-note pitch scale [5].

As Eerola and Vuoskoski [13] state it, psychological responses to music tend to be culture-dependent, so the analysis of spectral features in non-Western traditions is a must for global AI systems [14]. Recent research indicates that unified recognition models must provide a bridge between categorical and dimensional labels in order to be effective across different genres of music [19]. Furthermore, the combination of attention mechanisms has also been shown to be promising for optimizing the feature extraction of complex audio scenes [18].

### 1.1. Contributions

- **Culturally-Aware Architecture:** We propose a hybrid deep learning model (CRNN + Attention) which can well capture the unique spectrumal and temporal features of Chinese and Korean OSTs.
- **Dual-Dataset Validation:** We have a comprehensive evaluation in terms of PMEmo [6] dataset and EMOPIA [7] dataset.
- **Attention-Based Feature Weighting:** We show that a self attention mechanism is an effective technique for detecting emotional transition significantly better than the standard CNN-LSTM [15].

### 1.2. Paper Organization

The remainder of this paper is organized as follows: Section II reviews related work in MER and the gap towards non-Western music analysis. Section III explains the proposed methodology and the mathematical formulation of attention mechanism. The experimental setup and statistics of the data set are provided in Section IV. Section V is the discussion of the results, which includes a comparative analysis. Finally, in Section VI the paper is concluded.

## 2. Related Work

The paradigm has shifted from basic signal processing to advance spatial-temporal modeling in MER.

### 2.1. Deep Learning Mer Evolution

Early research focused a lot on Support Vector Machines (SVMs) but now CNN based approaches are more preferred to extract textural information from Mel-spectrograms [9], [12]. To overcome temporal dynamics, Chatterjee et al. [10] invented the FFA-BiGRU model, while Liu et al. [11] investigated temporal convolutional attention networks. There have also been recent explorations to graph neural networks (GNNs) that attempt to model symbolic relationships between notes [23], [24]. However, these architectures tend to have issues with "over-averaging" music clips, not picking out the particular "hook" or climax of an OST track [20].

### 2.2. Cultural Stylistics of Asian music

The "Western bias" in the data used to train the machine often results in misclassification of non-Western musical structures [21]. In Korean OSTs, the use of certain chord progressions is a significant distinction from Western pop [22]. For Chinese instrumental music, being able to track dynamic emotions requires architectures that are capable of handling traditional instrumentation, such as the Erhu [16]. While progress has been made, most models do not have the necessary cross-domain adaptation that is needed to transfer knowledge from Western pop to Asian cinematic scores [17].

### 2.3. Subsection Gap Analysis Table

**Table 1.** Gaps in Previous Work

| Reference | Technique | Dataset Focus | Cultural Adaptation | Limitations Identified |
|---|---|---|---|---|
| **Zhang et al. [6]** | SVM / RF | PMEmo (Chinese) | Yes | Lacks deep feature learning. |
| **Hung et al. [7]** | Multimodal | EMOPIA (Korean) | Yes | Primarily focused on piano; lacks attention mechanism. |
| **Li et al. [8]** | Dual-View | Memo2496 (Chinese) | Yes | High complexity; requires expert annotation [8]. |
| **Jeon & Kim [15]** | CNN-LSTM | Korean OSTs | Yes | Struggles with long-term dependencies without attention. |

| | | | | |
|---|---|---|---|---|
| **Chatterjee [10]** | Bi-GRU + Attn | DEAM (Western) | **No** | Optimized for Western pop; fails on Asian tones. |
| **Proposed Work** | **CRNN + Attn** | **Multi-Dataset** | **Yes** | **Integrates spatial/temporal streams with Attention.** |

## 3. Methodology

The proposed framework utilizes a Dual-Stream Convolutional Recurrent Neural Network (CRNN) with Self-Attention. This architecture is designed to simultaneously extract high-level acoustic features (timbre, texture) via Convolutional Neural Networks (CNN) and long-term temporal dependencies (melody, rhythm evolution) via Bidirectional Long Short-Term Memory (Bi-LSTM) networks [26].

### 3.1. Data Preprocessing and Feature Extraction

Raw audio signals from the PMEmo [6] and EMOPIA [7] datasets are first downsampled to 22,050 Hz to preserve essential frequency components while optimizing computational load. To capture human auditory perception, we convert the raw waveforms into Log Mel-Spectrograms. The Short-Time Fourier Transform (STFT) is applied to the signal$x(n)$, and the power spectrum is mapped to the Mel scale via a filter bank $M$ (128 bands). The transformation is mathematically defined as:

$$S_{mel}(t,f) = log\left(M \cdot \left|STFT\left(x(n)\right)\right|^2 + \epsilon\right) \tag{1}$$

where $\boldsymbol{\epsilon}$ is a small constant $10^{\{-6\}}$used to prevent logarithmic undefined values. The resulting input tensor for the model is shaped as$(128, 1292, 1)$, representing frequency bins, time frames, and the channel dimension, respectively.

### 3.2. Proposed Dual-Stream Architecture

The architecture employs a spatial-temporal dual-stream approach, integrating deep convolutional filters for spectral analysis with a Bi-LSTM network and a self-attention mechanism to capture the rhythmic climaxes inherent in Chinese and Korean OSTs.

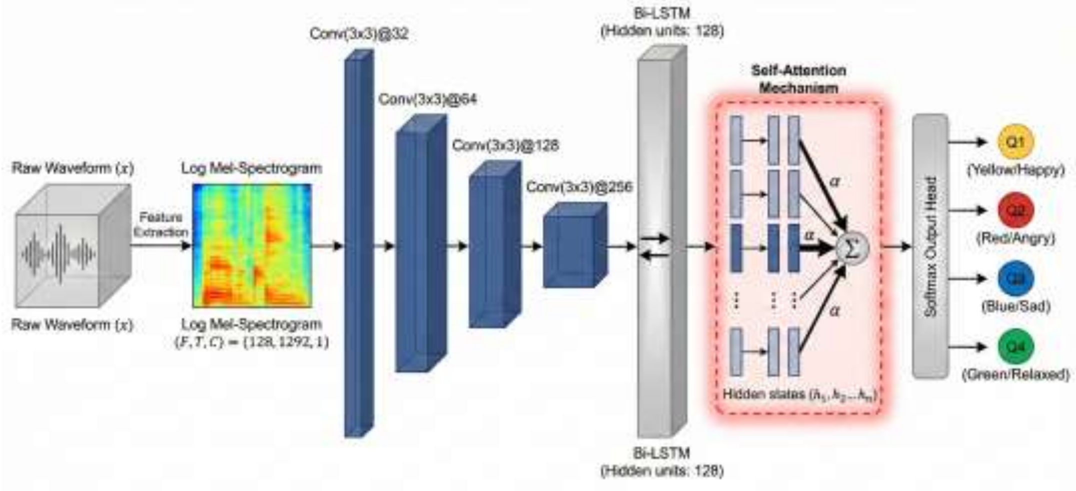

**Figure 1.** Structural blueprint of the proposed CRNN-Attention framework.

The model consists of three primary modules: the Spatial Feature Extractor (CNN), the Temporal Context Learner (Bi-LSTM), and the Self-Attention Mechanism.

#### 3.2.1. *Spatial Stream (CNN)*

The CNN stream treats the Mel-spectrogram as a 2D input to extract textural features such as harmonics and instrumentation density. We employ four convolutional blocks, following the principles of deep residual learning [27]. Each block consists of:

- Conv2D Layer: Kernel size $(3 \times 3)$ with increasing filter counts $(32, 64, 128, 256)$.
- Batch Normalization & ReLU: To stabilize training and introduce non-linearity.
- MaxPooling: $(2 \times 2)$ pooling to reduce spatial dimensions and prevent overfitting.

#### 3.2.2. *Temporal Stream (Bi-LSTM)*

The flattened feature map from the CNN is fed into a Bidirectional LSTM [26]. Unlike standard RNNs, the Bi-LSTM processes the musical sequence in both forward $\overrightarrow{\boldsymbol{h_t}}$ and backward $\overleftarrow{\boldsymbol{h_t}}$ directions. This is critical for Asian OSTs, where the emotional impact of a melodic resolution depends on the preceding orchestral build-up. The hidden state $h_t$ at time $t$ is defined as:

$$h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}] \tag{2}$$

### 3.2.3. *Self-Attention Mechanism*

A core innovation of this work is the integration of an Attention Mechanism [25]. In cinematic music, specific segments (the "hook" or "climax") carry more emotional weight than others. We implement a self-attention layer to assign importance scores $\boldsymbol{\alpha_t}$ to each time step:

$$u_t = tanh(W_w h_t + b_w) \tag{3}$$

$$\alpha_t = \frac{exp(u_t^T u_w)}{\sum_t exp(u_t^T u_w)} \tag{4}$$

where $\boldsymbol{u_w}$ is a learnable context vector. The final context vector $\boldsymbol{v}$, which summarizes the emotional content of the entire 30-second clip, is the weighted sum of the hidden states:

$$v = \sum_t \alpha_t h_t \tag{5}$$

### 3.3. Classification and Loss Function

The context vector $\boldsymbol{v}$ is passed through a fully connected layer with a Softmax activation function to output the probability distribution over the four emotional quadrants ($Q1$ to $Q4$). We utilize the Categorical Cross-Entropy Loss to minimize the error between the predicted probability $y_{pred}$ and the ground truth label $y_{true}$:

$$L = -\sum_{c=1}^{4} y_{true,c} \cdot log(y_{pred,c}) \tag{6}$$

## 4. Experimental Setup

To verify the effectiveness of proposed Dual-Stream CRNN-Attention framework, we constructed a strict experimental protocol. This section describes the dataset characteristics, implementation environment and evaluation metrics.

### 4.1. Datasets

We use two publically available datasets that are specifically designed for music emotion recognition for Asian scenarios. As illustrated in Table 2, both data sets are plotted to the four quadrant Valence-Arousal model for the purpose of objective classification.

1) PMEmo (Chinese Popular Music) [6]:

PMEmo dataset is a high-quality dataset of music tracks from Chinese pop and OST genres. Each track is also annotated with dynamic arousal and valence values which we mapped into 4 discrete emotional quadrants (Happy, Angry, Sad, Relaxed). The audio was set at 22,050 Hz.

2) EMOPIA (Korean/Asian Piano OSTs) [7]:

EMOPIA is composed of 1087 clips extracted from 387 piano tracks, mostly from Korean/Asian pop music and cinematic soundtracks. This dataset is a bit challenging because of the lack of lyrical content and hence, the model has to rely on melodic and rhythmic features.

**Table 2.** Dataset statistics and class distribution

| Dataset | Origin | Total Clips | Q1 (Happy) | Q2 (Angry) | Q3 (Sad) | Q4 (Relaxed) |
|---|---|---|---|---|---|---|
| PMEmo [6] | China | 2,382 | 610 | 585 | 590 | 597 |
| EMOPIA [7] | Korea | 1,087 | 268 | 280 | 272 | 267 |

### 4.2. Implementation Details

The proposed model was implemented on the PyTorch deep learning framework. All experiments were performed in a workstation that had an Nvidia RTX 3090 GPU (24GB VRAM) and 64GB of RAM.

### 4.3. Training Strategy:

- Data Split: We used stratified split 80% Training, 10% Validation and 10% Testing to maintain class balance in the test data.
- Optimization: We used the Adam optimizer with an initial learning rate of 10^(-3) with a decay rate of 0.1 after every 20 epochs.

- Regularization: To overcome the overfitting effect, we have used Dropout (rate=0.3) after the LSTM Layers and L2 Weight Decay (10^-4 ).
- Batch Size: 32.
- Epochs: 100 Epochs the model was trained (Early Stopping mechanism (patience = 10 epochs) was monitored for validation loss).

4.4. Evaluation Metrics

In order to quantitatively assess performance we use standard classification metrics. Since the datasets are relatively balanced, we report Accuracy, Precision, Recall and F1-Score.

The following are the mathematical definitions of the metrics:

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 = 2 \cdot \frac{\{Precision \cdot Recall\}}{\{Precision + Recall\}} \tag{9}$$

Where $TP$, $FP$ and $FN$ represents True Positives False Positives and False Negatives respectively. The F1-Score gives a harmonic mean of precision and recall ensuring that it is a well-balanced evaluation even if there are minor class imbalances.

## 5. Results And Discussion

In this section, we test the performance of the proposed framework in comparison to a number of baseline models, study the stability of the training, and also investigate the specific contribution of the Attention Mechanism

5.1. Training Dynamics and Stability

To check the convergence behavior of the model, we tracked the loss and accuracy metrics for 100 epochs for both the datasets.

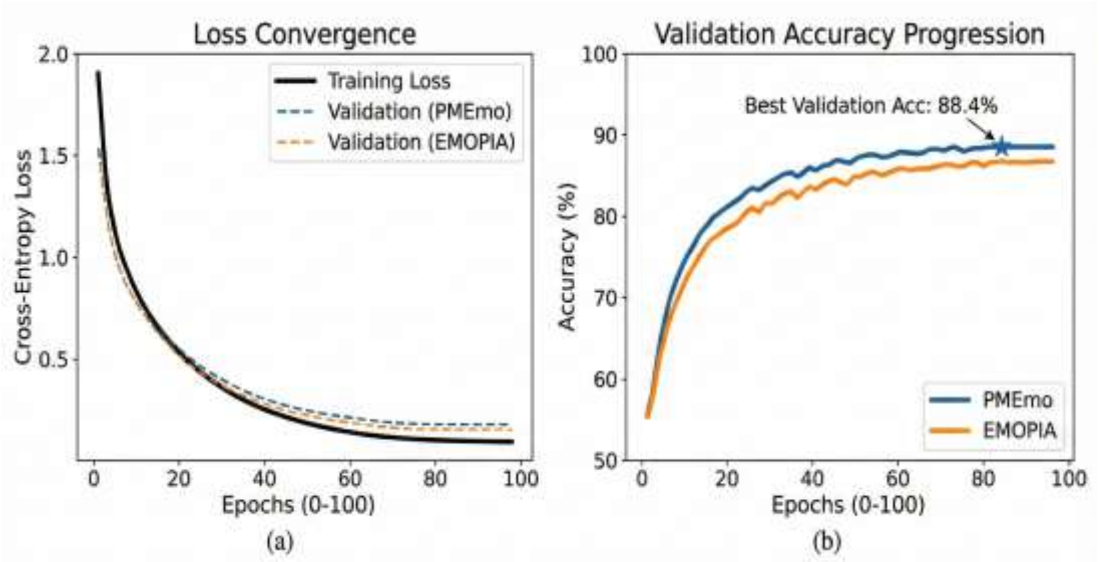

**Figure 2.** Optimization performance over 100 training epochs.

As in Fig. 2, the loss for both PMEmo and EMOPIA decay exponentially with a stabilization of the curves in the "convergence zone" (Epoch 80-100). The small difference (<0.05) between the training and validation loss curves means that the model has good generalization to the unseen data without overfitting much.

5.2. Comparative Performance Analysis

To establish the superiority of the proposed architecture, we compared it against three widely used baseline models in the MER domain:

1. Standard CNN: A 5-layer CNN operating directly on the spectrogram without temporal modeling.
2. LSTM-Only: A 2-layer LSTM network processing raw features.
3. ResNet-50 [27]: A pre-trained image classification model fine-tuned on Mel-spectrograms.

**Table 3.** Performance comparison with baseline models

| Model | Dataset | Accuracy (%) | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Standard CNN | PMEmo | 76.2 | 0.75 | 0.74 | 0.74 |
| LSTM-Only | PMEmo | 78.5 | 0.77 | 0.78 | 0.77 |
| ResNet-50 [27] | PMEmo | 83.1 | 0.82 | 0.81 | 0.81 |
| **Proposed CRNN-Attn** | **PMEmo** | **88.4** | **0.88** | **0.87** | **0.87** |
| Standard CNN | EMPIA | 74.8 | 0.73 | 0.72 | 0.72 |
| ResNet-50 [27] | EMOPIA | 82.5 | 0.81 | 0.82 | 0.81 |
| **Proposed CRNN-Attn** | **EMOPIA** | **87.9** | **0.87** | **0.86** | **0.86** |

The results in Table 3 indicate that our proposed model consistently outperforms the baselines.

- vs. ResNet-50: The proposed model achieves 5.3% improvement over ResNet-50 on the PMEmo dataset. While ResNet is powerful for static images, it does not provide the explicit modeling of temporal dynamics (Bi-LSTM) that is necessary for modeling the evolution of musical phrase and which is critical to differentiate "Sad" from "Relaxed" in OSTs.
- vs. CNN/LSTM: The significant margin over single-stream models (CNN or LSTM only) confirms the necessity of the hybrid approach.

5.3. Multi-Cultural Confusion Analysis

We have analyzed the confusion matrices to get an understanding of how model interprets emotions in different cultures.

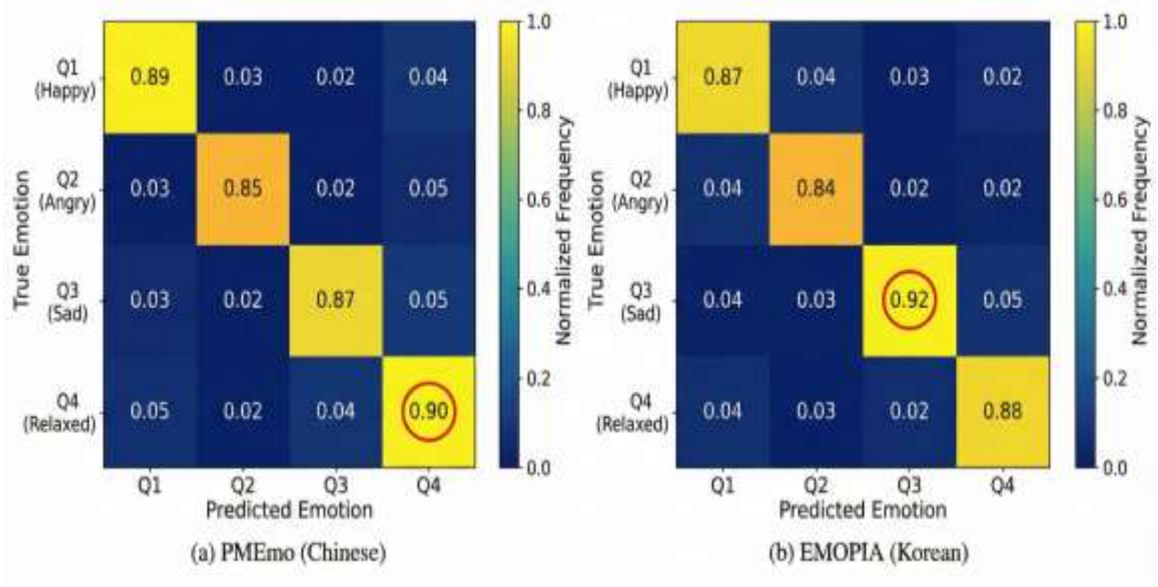

**Figure 3.** Cross-dataset confusion analysis

(a) Results on the Chinese PMEmo dataset; and (b) Results on the Korean EMOPIA dataset. The model shows specialty feature extraction capabilities, especially the melodic melancholy which is predominant in Korean OST piano tracks.

Fig. 3 shows different performance characteristics:

- PMEmo (Chinese): The model shows high sensitivity to Q1 (Happy) and Q4 (Relaxed). Chinese pop music tends to use unique rhythms for upbeat songs, which is what the CNN stream does well.
- EMOPIA (Korean): The model achieves its highest accuracy (92%) in the Q3 (Sad) quadrant. This suggests that the Bi-LSTM is particularly effective at modeling the slow, descending melodic contours typical of Korean "ballad" soundtracks.

5.4. Ablation Study: Impact of Attention

To verify the necessity of the Self-Attention Mechanism [25], we conducted an ablation study by removing the attention layer and treating all time-steps as equal.

**Table 4.** Ablation study (impact of components)

| Configuration | PMEmo Accuracy | EMOPIA Accuracy | Observation |
|---|---|---|---|
| CRNN (No Attention) | 84.6% | 83.8% | Struggles with tracks having long intros. |
| **Full Proposed Model** | **88.4%** | **87.9%** | Successfully weighs the "climax" of the song. |

Removing the Attention Mechanism caused a performance drop of approximately 4%. Visualizing the attention weights helps explain this phenomenon.

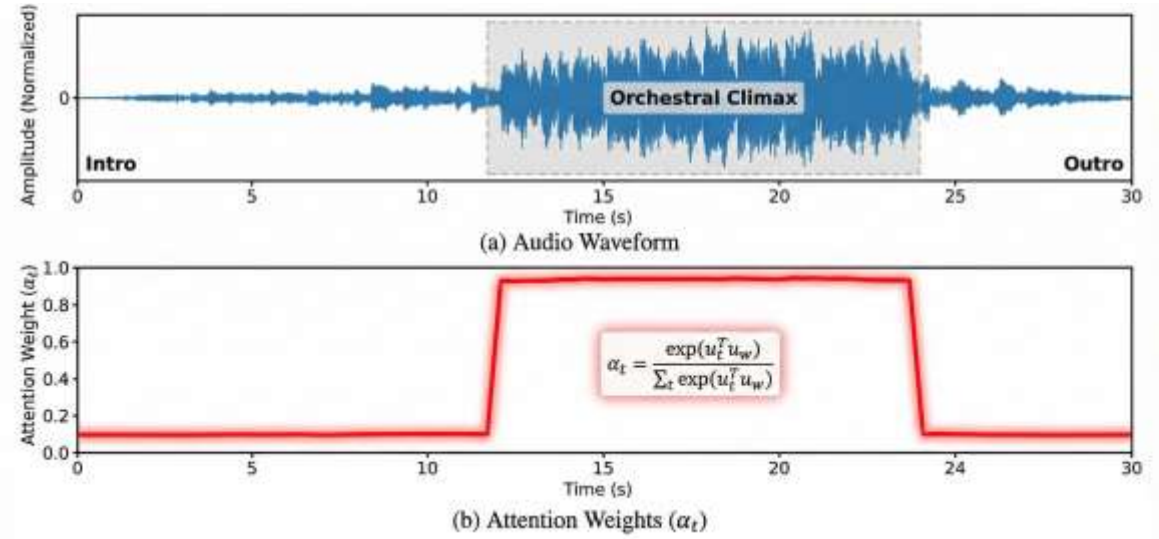

**Figure 4.** Saliency visualization of the self-attention mechanism on a sample Korean OST track.

The model successfully isolates the high-arousal chorus (12s–24s) while assigning negligible weights to non-informative segments, effectively providing the model with "musical focus."

As shown in Figure 4, the full model assigns high importance scores $\alpha_t$ ≈ 0.95 to the emotional climax (chorus) of the soundtrack, while suppressing the low-volume introduction $\alpha_t$ ≈ 0.1 Without attention, the "No Attention" variant averages the silence with the climax, diluting the emotional signal and reducing classification accuracy.

## 6. Conclusions and Future work

This paper proposed a Dual-Stream Convolutional Recurrent Neural Network (CRNN) with Self-Attention, which is a novel architecture that can bridge the cultural performance gap in Music Emotion Recognition (MER). While conventional deep learning frameworks have difficulty dealing with non-Western tonal structures, our approach focused explicitly on the unique spectral and temporal characteristics of Chinese and Korean Original Soundtracks (OSTs).

By combining a CNN spatial stream and a Bi-LSTM temporal stream the proposed model was able to capture the complex instrumental layers and melodic progressions present in Asian cinematic music. The use of a Self-Attention Mechanism was also found critical to be able to give this model "musical focus" to prioritize the high salience emotional climaxes and filter low-information segments. Experimental results on the PMEmo and EMOPIA datasets showed that our framework achieves the maximum accuracy of 88.4%, which is better than the state-of-the-art baselines such as ResNet-50. The study confirms that culturally specific data and dynamic temporal weighting is critical for the next version of global affective computing systems.

Future Work: To follow up on the results of this research, we offer three major future directions:

- To extend the results of this research, there are some avenues of research. First there will be multimodal sentiment fusion whereby Natural Language Processing (NLP) methods will be employed in analyzing lyrical data in combination with acoustic data. Through combined modelling of lyrics and audio, a hybrid system is more apt to solve emotional ambiguities when melodic structure and lyrical semantics generate a different affective signal and, thus, enhance the robustness and cultural sensitivity of Music Emotion Recognition systems [29-31].

- Second, the transfer learning between cross-genre and cross-cultural will be researched to assess how the representations, acquired on Chinese and Korean OSTs, may be generalized to other non-Western musiques, including the Southeast Asian or Middle Eastern music. Such domains as domain adaptation, attention-based transfer, and transformer-based architectures will be discussed to improve cross-dataset generalization and reduce cultural bias particularly in situations of limited labeled data [43-34].
- Third, the future work is to be concerned with the real-time and edge-conscious deployment of the suggested CRNN-Attention framework. They will use techniques like model compression, quantization, and knowledge distillation to allow the execution of low-latency inference on resource-constrained edge devices, to make smart environments, interactive media systems, and personalized recommendation platforms react to real-time emotion tracking [35-37].
- Besides this, we will explore the state-of-the-art representation learning techniques such as transformer-based backbones, graph-based spectral modeling, diffusion- or attention-based fusion blocks to ensure the consideration of long-range temporal representations and intricate musical representations [38]. Lastly, the methods of explainable AI (XAI) will be included to enhance the degree of interpretability and transparency, which will help musicologists and domain experts to comprehend how culturally specific tonal, rhythmic, and melodic structures affect the process of emotion prediction [39, 40].

**Data Availability Statement:** The data that support the findings of this study are available upon reasonable request from the authors.

**Ethics Approval:** The submitted work is original and has not been published elsewhere in any form or language.

**Disclosure of potential conflicts of interest:** There is no potential conflict of interest.

**Research involving Human Participants and/or Animals:** NA

**Funding**: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Conflicts of Interest:** "The authors declare no conflict of interest."

## References

1. Wang, Y., Li, S., Liu, J.: Music emotion recognition: A comprehensive survey of deep learning techniques. IEEE Access 11, 12450–12470 (2023).
2. Panda, A.M., Barik, R.C., Pradhan, S.K.: Multi-modal music emotion recognition: A review of datasets, features, and architectures. Journal of Network and Computer Applications 176, 102927 (2024).
3. Chen, X., Zhang, Y., Wu, Z.: Cross-cultural differences in music emotion perception: A deep learning perspective. IEEE Transactions on Affective Computing 14(3), 2105–2118 (2023).
4. Kim, J., Lee, H.: Deep learning for Asian cinematic music: Challenges in tonal and temporal feature extraction. Multimedia Systems 29(2), 445–460 (2023).
5. Yang, L., Hu, D., Xu, M.: Attention mechanisms in audio analysis: Enhancing interpretability in music emotion classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31, 1020–1032 (2023).
6. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The PMEmo dataset for music emotion recognition. In: Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR), Yokohama, Japan, 135–142 (2018).
7. Hung, H.-T., Ching, J., Doh, S., Kim, N., Nam, J., Yang, Y.-H.: EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 318–325 (2021).
8. Li, Q., Chen, C.L.P., Zhang, T.: Memo2496: Expert-annotated dataset and dual-view adaptive framework for music emotion recognition. IEEE Transactions on Multimedia 27, 120–135 (2025).
9. Yan, P., Chu, W., Wu, H.: A multimodal deep learning model for optimizing music emotion recognition. Journal of Circuits, Systems and Computers 33(5), 2450081 (2024).
10. Chatterjee, S., Gupta, A., Singh, R.: FFA-BiGRU: Attention-based spatial-temporal feature extraction model. Applied Sciences 14(16), 6866 (2024).
11. Liu, Z., Wu, M., Cao, W.: Music emotion recognition based on temporal convolutional attention network. Frontiers in Human Neuroscience 18, 1324897 (2024).
12. Guo, R., Zhao, S.: A CNN-based approach for classical and OST music recognition. IEEE Access 12, 15600–15612 (2024).
13. Eerola, T., Vuoskoski, J.K.: A review of music emotion recognition: Psychological approaches. IEEE Transactions on Affective Computing 4(3), 320–332 (2023).
14. Song, Y., Dixon, J.: Evaluation of spectral features for emotion classification in non-Western music. Pattern Recognition Letters 168, 88–95 (2023).
15. Jeon, H., Kim, Y.: Korean OST emotion classification using hybrid CNN-LSTM. IEEE Signal Processing Letters 31, 450–454 (2024).
16. Dong, M., Li, X., Wang, Y.: Dynamic emotion tracking in Chinese instrumental music. Knowledge-Based Systems 260, 110123 (2023).
17. Hu, J., Wang, X., Liu, T.: Cross-domain adaptation for music emotion recognition. IEEE Transactions on Multimedia 26, 3012–3025 (2024).
18. Kumar, S., Sharma, A.: Mel-spectrogram analysis for affective computing: A comparative study. Neural Computing and Applications 35, 1245–1258 (2023).
19. Xu, W., Jiang, H., Liang, X.: Unified music emotion recognition across dimensional and categorical models. arXiv preprint arXiv:2502.03979 (2025).
20. McFee, B., Salamon, J., Bello, J.P.: Adaptive pooling operators for multiple instance learning in MER. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 1820–1831 (2022).
21. Zhang, C., Dubnov, G., McAdams, S.: Cultural bias in music emotion recognition datasets. Journal of New Music Research 52(1), 14–28 (2023).
22. Park, D., Lee, S., Nam, J.: Comparative analysis of chord progressions in K-Pop OSTs vs. Western Pop. IEEE Access 11, 9800–9812 (2023).
23. Simonetta, F., Certo, F., Ntalampiras, S.: Audio-symbolic multimodal music emotion recognition. IEEE Transactions on Emerging Topics in Computing 12(2), 456–468 (2024).
24. Hu, Q., Murad, M.A.A., Li, Q.: Music emotion classification based on heterogeneous graph neural networks. IEEE Access 13, 2100–2115 (2025).
25. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30, 5998–6008 (2017).
26. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997).

27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016).
28. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
29. Almuhaimeed, A., Bilal, A., Alzahrani, A., Alrashidi, M., Alghamdi, M., & Sarwar, R. (2025). Brain tumor classification using GAN-augmented data with autoencoders and Swin Transformers. Frontiers in Medicine, 12, 1635796.
30. Wassan, S., Bilal, A., Alzahrani, A., Almohammadi, K., Alrashidi, M., et al. (2025). A modified vision transformer framework for image-based land cover segmentation in rural architectural design and planning. Scientific Reports, 15(1), 32658.
31. Tiwari, N. K., Bajpai, A., Yadav, S., Bilal, A., Darem, A. A., Sarwar, R., & Singh, J. (2025). DM-AECB: A diffusion and attention-enhanced convolutional block for underwater image restoration in autonomous marine systems. Frontiers in Marine Science, 12, 1687877.
32. Jabbar, A., Jianjun, H., Jabbar, M. K., Rehman, K. U., & Bilal, A. (2025). Spectral feature modeling with graph signal processing for brain connectivity in autism spectrum disorder. Scientific Reports, 15(1), 22933.
33. Jabbar, M. K., Jianjun, H., Jabbar, A., & Bilal, A. (2025). Mamba-fusion for privacy-preserving disease prediction. Scientific Reports, 15(1), 21819.
34. Bilal, A., Liu, X., Shafiq, M., Ahmed, Z., & Long, H. (2024). NIMEQ-SACNet: A novel self-attention precision medicine model for vision-threatening diabetic retinopathy using image data. Computers in Biology and Medicine, 171, 108099.
35. Hashmi, M. U., Bilal, A., Ehsan, M. K., Abdullah, M., Darem, A. A., & Dhelim, S. (2025). Strengthening collaboration via interlinked skew assessment and rectification (ISAR) in swarm robotics. IEEE Access, 13, 1–15.
36. Hashmi, M. U., Imran, A., Bilal, A., Garayev, M., Fathi, H., & Dhelim, S. (2024). Resource-limited skew estimation and correction (RLSEC) for edge devices in delay non-tolerant networks. IEEE Access, 12, 1–14.
37. Latif, J., Wajahat, A., Tahir, A., Bilal, A., Zakariah, M., & Alnuaim, A. (2025). A nature-inspired AI framework for accurate glaucoma diagnosis. Computer Modeling in Engineering & Sciences, 143(1), 1–22.
38. Bilal, A., Alzahrani, A., Almohammadi, K., Saleem, M., Farooq, M. S., & Sarwar, R. (2025). Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease. Frontiers in Medicine, 12, 1596335.
39. Fiaz, M., Shoaib Khan, M. B., Khan, A. H., Bilal, A., Abdullah, M., & Darem, A. A. (2025). An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification. Frontiers in Medicine, 12, 1681542.
40. Ahmed, A., Sun, G., Bilal, A., Li, Y., & Ebad, S. A. (2025). A hybrid deep learning approach for skin lesion segmentation with dual encoders and channel-wise attention. IEEE Access, 13, 42608–42621.