

Attention-Assisted Multi-Scale Music Spectral Representation Learning for Intelligent Emotion Quantification

Jiajie Liu

jiajie0824@163.com

Guangzhou College of Technology and Business

Article

Keywords:

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11312569/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Attention-Assisted Multi-Scale Music Spectral Representation Learning for Intelligent Emotion Quantification

Jiajie Liu^{1,*}

¹School of Music, Guangzhou College of Technology and Business, Guangzhou, 510850, China

*Corresponding author: Jiajie Liu (jiajie0824@163.com)

ABSTRACT

Accurate quantification of musical emotions remains challenging due to the complex relationship between acoustic structures and human affective perception. Existing approaches mainly focus on fixed spectral representations or categorical emotion recognition, which limits their ability to capture hierarchical frequency-temporal patterns and continuous emotional variations in music signals. This paper proposes MASAN-Net, an attention-assisted multi-scale spectral representation learning framework for continuous music emotion quantification. The proposed framework employs multi-scale convolutional feature extraction to learn hierarchical spectral representations and introduces frequency-temporal attention mechanisms to adaptively identify emotion-related acoustic patterns. A regression-based prediction module is further developed to estimate continuous Valence and Arousal dimensions from music spectrograms. Extensive experiments are conducted on the DEAM dataset, with cross-dataset validation on the MTG-Jamendo dataset to evaluate generalization capability. Compared with traditional machine learning methods and advanced deep learning baselines, MASAN-Net achieves superior performance under multiple regression metrics, including MSE, RMSE, MAE, and Pearson correlation. Ablation studies and repeated experiments further verify the effectiveness and stability of the proposed framework. Attention visualization provides interpretable insights into the spectral-temporal regions associated with emotional prediction. The proposed method offers an effective and explainable solution for intelligent music emotion analysis.

1 Introduction

Music emotion quantification has become an important research topic in intelligent multimedia analysis, affective computing, and music information retrieval¹. Unlike conventional audio understanding tasks that focus on recognizing predefined categories, continuous emotion quantification aims to estimate the intensity and variation of emotional states expressed in music². The emotional perception of music is influenced by complex interactions among spectral characteristics, temporal evolution, harmonic structures, and acoustic textures³. Therefore, developing computational models capable of effectively learning the relationship between acoustic representations and human emotional perception is essential for intelligent music analysis and personalized multimedia applications⁴.

Early studies on music emotion analysis mainly relied on handcrafted acoustic features, including spectral descriptors, rhythmic characteristics, and timbral information, combined with traditional machine learning algorithms. Although these approaches provided initial solutions for computational emotion analysis, their performance was highly dependent on manually designed feature representations and often failed to capture complex relationships hidden in music signals⁵. With the rapid development of deep learning, neural network-based methods have demonstrated strong capabilities in automatically learning discriminative representations from audio data⁶. Convolutional neural networks (CNNs) have been widely adopted for processing Mel-spectrogram representations by capturing local spectral patterns, while recurrent neural networks (RNNs) and their variants have improved temporal modeling by learning sequential dependencies⁷. More recently, Transformer-based architectures have introduced attention mechanisms into audio analysis and achieved promising results by modeling long-range dependencies⁸.

Despite these advances, several limitations remain in continuous music emotion quantification. Many existing approaches rely on single-scale spectral representations, which may be insufficient for modeling hierarchical acoustic structures distributed across different frequency ranges and temporal resolutions⁹. Music signals contain diverse spectral components with different perceptual meanings, and effective emotion modeling requires representation learning strategies capable of integrating multi-scale information¹⁰. Although attention-based models provide powerful global dependency modeling capabilities, most existing attention mechanisms are designed for general sequence processing and do not explicitly consider the unique frequency-temporal characteristics of music spectrograms¹¹. The adaptive interaction between frequency-related acoustic information and

temporal emotional evolution remains insufficiently explored¹². Many existing studies focus on categorical emotion recognition, while continuous Valence-Arousal representation provides a more detailed description of emotional intensity and variation¹³. The limited interpretability of deep learning-based music analysis models remains a challenge, as the relationship between learned representations and emotionally relevant spectral regions is often unclear¹⁴.

To address these challenges, this paper proposes MASAN-Net, an attention-assisted multi-scale spectral representation learning framework for continuous music emotion quantification. The proposed framework first transforms music audio signals into Mel-spectrogram representations and employs multi-scale convolutional feature learning to capture hierarchical spectral characteristics. A frequency-temporal attention mechanism is then introduced to adaptively enhance emotion-related spectral and temporal patterns by assigning different importance weights to informative regions. Based on the refined representations, a regression-based prediction module is developed to estimate continuous Valence and Arousal values, enabling a more comprehensive description of musical emotions. In addition, attention visualization is incorporated to analyze the spectral-temporal regions contributing to emotional predictions and improve the interpretability of the proposed framework.

The main contributions of this study are summarized as follows:

- A multi-scale spectral representation learning framework is proposed to capture hierarchical acoustic characteristics from complex music signals and improve continuous emotion quantification performance.
- A frequency-temporal attention mechanism is developed to adaptively model emotion-related spectral structures and temporal variations, enabling more effective feature refinement for music emotion analysis.
- An attention-based interpretability strategy is introduced to reveal the relationship between learned spectral representations and emotional predictions, improving the transparency of deep learning-based music analysis.
- Extensive experiments on the DEAM dataset, together with cross-dataset validation on the MTG-Jamendo dataset, demonstrate the effectiveness, generalization capability, and stability of the proposed MASAN-Net framework.

The remainder of this paper is organized as follows. Section II introduces related works on music emotion analysis, spectral representation learning, and attention-based audio modeling. Section III presents the proposed MASAN-Net framework in detail. Section IV describes the experimental setup, including datasets, implementation details, baseline methods, and evaluation metrics. Section V reports experimental results and discussions. Finally, Section VI concludes this study and discusses future research directions.

2 Related Work

2.1 Music Emotion Recognition

Music emotion recognition has been extensively studied as an important research direction in music information retrieval and affective computing. The primary objective of this task is to computationally model the relationship between acoustic characteristics and human emotional perception¹⁵. Early studies mainly relied on handcrafted acoustic features, including spectral descriptors, rhythmic patterns, tempo-related information, and timbral characteristics, which were combined with traditional machine learning algorithms such as support vector machines and regression models. These approaches provided initial solutions for automatic music emotion analysis; however, their effectiveness was highly dependent on the quality of manually designed features and their ability to represent complex emotional variations in music signals¹⁶. With the development of deep learning, data-driven approaches have gradually replaced handcrafted feature-based methods by enabling automatic representation learning from audio signals. Convolutional neural networks (CNNs) have been widely adopted for music emotion analysis by treating Mel-spectrograms and other time-frequency representations as image-like inputs¹⁷. CNN-based models can effectively capture local spectral structures and learn discriminative acoustic patterns without requiring manually engineered features¹⁸. To further incorporate temporal information, recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and convolutional recurrent neural networks (CRNNs) have been introduced to model sequential dependencies in music signals. These methods demonstrate that both spectral characteristics and temporal evolution play important roles in determining perceived emotional states¹⁹. Transformer-based architectures have attracted increasing attention in music emotion analysis due to their ability to capture long-range dependencies through self-attention mechanisms. By modeling global relationships among audio features, Transformer-based methods have achieved competitive performance in various audio understanding tasks. However, most existing approaches are primarily designed for general representation learning and may not fully consider the unique characteristics of music emotion perception. In particular, music signals contain complex hierarchical structures distributed across different frequency ranges and temporal scales, which require more adaptive representation strategies²⁰. Furthermore, many existing studies formulate emotion recognition as a classification problem using discrete emotion categories, while continuous emotion representation in the Valence-Arousal space provides a more comprehensive

description of emotional intensity and variation. Therefore, developing models capable of jointly capturing multi-scale spectral patterns, temporal dynamics, and interpretable emotional representations remains an important challenge in intelligent music emotion quantification.

2.2 Spectral Representation Learning for Audio Analysis

Spectral representation learning has become a fundamental component in modern audio analysis due to its capability of transforming complex acoustic signals into structured representations suitable for deep learning models. Raw audio waveforms contain rich information in both frequency and temporal dimensions, but their high-dimensional and unstructured characteristics make direct modeling challenging²¹. Time-frequency representations, such as spectrograms and Mel-spectrograms, provide an effective solution by converting audio signals into two-dimensional representations that preserve important spectral characteristics and temporal variations. Among these representations, Mel-spectrograms have been widely adopted in music and speech analysis because they approximate human auditory perception by emphasizing perceptually meaningful frequency distributions²². With the development of deep learning, convolutional neural networks have demonstrated strong capabilities in learning hierarchical representations from spectrogram-based inputs. By applying convolution operations on time-frequency representations, CNN-based approaches can automatically capture local spectral patterns, including harmonic structures, timbral characteristics, and frequency variations associated with different acoustic events. These advantages have enabled CNN architectures to achieve significant improvements in various audio understanding tasks, including music classification, sound event detection, and emotion analysis²³. Conventional CNN-based approaches typically rely on fixed convolutional kernels and single-scale feature extraction strategies, which may limit their ability to model complex acoustic structures distributed across different frequency ranges and temporal resolutions. To overcome the limitations of single-scale representation learning, recent studies have explored multi-scale feature extraction strategies that integrate information from different receptive fields. Multi-scale architectures enable models to simultaneously capture fine-grained local patterns and broader contextual characteristics, providing richer representations for complex audio signals. In music analysis, this capability is particularly important because emotional expression is often associated with interactions among multiple acoustic factors, including short-term spectral fluctuations, long-term structural changes, and dynamic variations throughout a musical piece²⁴. Therefore, effective emotion modeling requires representation learning mechanisms that can adaptively integrate multi-scale spectral information. Although existing spectral representation learning methods have achieved promising performance, several challenges remain. Many approaches focus primarily on improving feature extraction capability while paying limited attention to the selective importance of different frequency and temporal regions. In complex music signals, not all spectral components contribute equally to emotional perception, and the relationship between specific acoustic patterns and emotional responses remains difficult to model. Consequently, further research is needed to develop adaptive representation learning approaches that can capture hierarchical spectral structures while identifying emotion-relevant frequency-temporal characteristics²⁵.

2.3 Attention Mechanisms in Audio Understanding

Attention mechanisms have become an important technique in deep learning-based audio understanding due to their ability to dynamically adjust the importance of different feature components. Unlike conventional neural networks that treat all extracted features equally, attention-based models assign adaptive weights to informative regions, enabling more effective representation learning. This property is particularly valuable for audio analysis tasks, where meaningful acoustic patterns may appear at different temporal positions and frequency ranges²⁶. The introduction of self-attention mechanisms and Transformer architectures has significantly advanced audio representation learning by enabling models to capture long-range dependencies without relying on recurrent structures. Transformer-based approaches have been successfully applied in various audio-related tasks, including speech recognition, sound event detection, and music understanding²⁷. By establishing relationships among different elements within feature sequences, self-attention mechanisms provide a powerful strategy for modeling complex interactions in audio signals. However, most Transformer-based methods employ general attention strategies that are not specifically designed for the structural characteristics of time-frequency audio representations²⁸. For spectrogram-based audio analysis, attention mechanisms have been further explored to enhance the selection of important frequency and temporal information²⁹. Frequency attention focuses on identifying informative spectral components, while temporal attention emphasizes significant moments or dynamic variations within audio sequences³⁰. These strategies have demonstrated potential for improving model performance by reducing the influence of irrelevant information and highlighting task-related acoustic patterns. In music emotion analysis, such adaptive weighting is especially meaningful because emotional perception is often associated with specific combinations of spectral characteristics and temporal evolution patterns³¹. Despite these advantages, existing attention-based audio models still face several limitations. Many approaches apply attention mechanisms as general feature enhancement modules without explicitly considering the complementary relationship between frequency structures and temporal dynamics in music signals. Existing methods often focus primarily on prediction performance, while providing limited interpretation of which acoustic regions contribute to emotional decisions. This lack of transparency restricts the understanding of deep learning models in affective computing applications³². Developing attention mechanisms that are specifically adapted

to music spectral representations remains an important research direction. An effective framework should not only capture global dependencies but also distinguish the relative importance of different frequency bands and temporal segments associated with emotional expression. These observations motivate the development of MASAN-Net, which integrates multi-scale spectral representation learning with frequency-temporal attention mechanisms to achieve more accurate and interpretable continuous music emotion quantification.

3 Methodology

3.1 Problem Formulation

In this study, music emotion quantification is formulated as a continuous regression problem, where the objective is to learn the mapping relationship between music audio signals and their corresponding emotional representations in the Valence-Arousal space. Given an input music signal, the task is to extract discriminative spectral characteristics and predict continuous emotional values rather than assigning discrete emotion categories. Let the input audio signal be represented as $x(t)$, where t denotes the temporal index. The corresponding time-frequency representation is obtained through spectral transformation and can be formulated as

$$S = \mathcal{F}(x(t)) \quad (1)$$

where S represents the generated spectrogram feature and $\mathcal{F}(\cdot)$ denotes the time-frequency transformation function. In this work, the Mel-spectrogram representation is adopted as the primary acoustic representation due to its effectiveness in capturing perceptually relevant frequency patterns in music signals. The objective of the proposed framework is to learn a nonlinear mapping function that estimates the continuous emotional state from the extracted spectral representation. Given the input spectrogram S , the model aims to predict the emotional vector

$$\hat{y} = f(S; \theta) \quad (2)$$

where $\hat{y} = [\hat{v}, \hat{a}]$ represents the predicted Valence and Arousal values, $f(\cdot)$ denotes the proposed MASAN-Net framework, and θ represents the trainable model parameters. The optimization objective is to minimize the difference between the predicted emotional values and the ground-truth annotations. Therefore, the learning process can be formulated as

$$\theta^* = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, \hat{y}_i) \quad (3)$$

where N denotes the number of training samples, y_i represents the ground-truth emotional annotation, $\hat{y}_i$ represents the predicted emotion vector, and $\mathcal{L}(\cdot)$ denotes the regression loss function. Through this formulation, MASAN-Net is designed to learn hierarchical spectral representations and capture emotion-related frequency-temporal patterns from music audio signals, providing a unified framework for continuous music emotion quantification.

3.2 MASAN-Net

The proposed framework follows a hierarchical representation learning pipeline, which transforms raw music signals into Mel-spectrogram representations and progressively extracts emotion-related acoustic features through multi-scale spectral encoding, frequency-temporal attention enhancement, and regression-based emotion prediction. As shown in Fig. 1, the framework consists of four main components: time-frequency representation construction, multi-scale spectral feature learning, frequency-temporal attention refinement, and continuous emotion quantification. This design enables MASAN-Net to jointly model hierarchical spectral characteristics and temporal emotional variations in complex music signals. After obtaining the multi-scale spectral representation, the proposed framework introduces a frequency-temporal attention mechanism to adaptively enhance emotion-related features. The frequency attention module focuses on identifying informative spectral regions associated with different acoustic components, while the temporal attention module captures dynamically changing patterns along the time dimension. Through the interaction between frequency-aware and temporal-aware representations, MASAN-Net is capable of modeling complex relationships between spectral structures and emotional expressions in music signals. The enhanced feature representation is aggregated and transmitted to a multilayer perceptron (MLP) regression head for continuous emotion prediction. Instead of assigning predefined emotion categories, the proposed framework estimates emotional values in the two-dimensional Valence-Arousal space, enabling a more detailed representation of emotional intensity and variation. The complete learning process can be formulated as

$$\hat{y} = g(\mathcal{A}(\mathcal{C}(S))) \quad (4)$$

where S denotes the input Mel-spectrogram, $\mathcal{C}(\cdot)$ represents the multi-scale convolutional feature extraction module, $\mathcal{A}(\cdot)$ represents the frequency-temporal attention enhancement module, $g(\cdot)$ denotes the MLP regression head, and $\hat{y}$ represents the predicted Valence-Arousal vector. Through this hierarchical architecture, MASAN-Net integrates spectral representation learning and attention-based feature refinement to effectively capture emotion-related acoustic patterns from complex music signals.

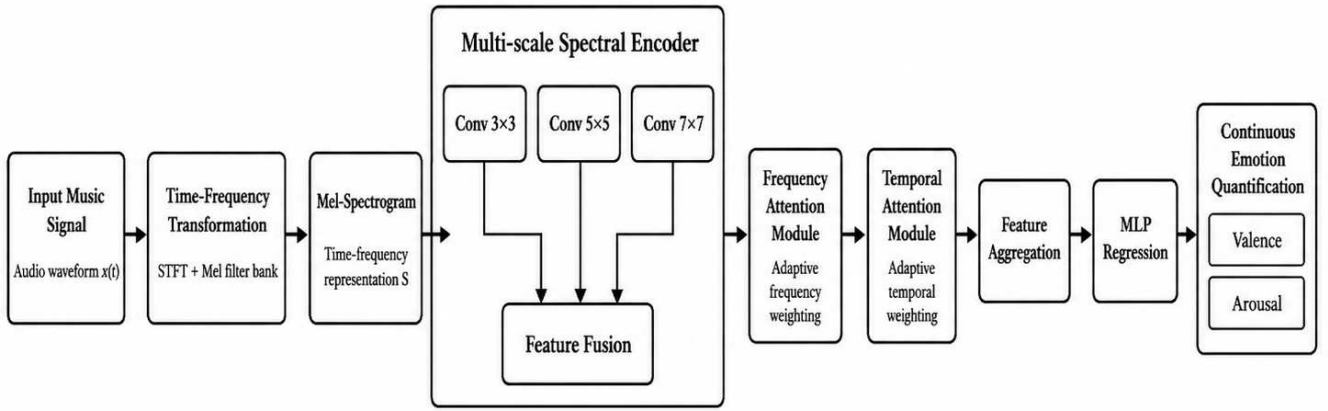


Figure 1. Overall framework of MASAN-Net. The proposed architecture transforms music signals into Mel-spectrogram representations and performs multi-scale spectral feature learning, frequency-temporal attention refinement, and continuous emotion quantification.

3.3 Multi-scale Frequency-Temporal Attention Representation Learning

The core component of MASAN-Net is the proposed multi-scale frequency-temporal attention representation learning module, which aims to enhance the model's capability of capturing emotion-related spectral and temporal patterns. The detailed structure of this module is illustrated in Fig. 2. The module first extracts multi-scale spectral features through parallel convolution operations with different receptive fields. The extracted representations are then adaptively refined by frequency attention and temporal attention mechanisms, enabling the model to emphasize informative frequency bands and emotionally relevant temporal segments. Given the input Mel-spectrogram feature map S , three convolution branches with different kernel sizes are employed to obtain multi-scale representations. The extracted features are concatenated to generate the integrated representation Z_{ms} , which can be formulated as:

Multiple convolutional operations with different kernel sizes are applied to capture acoustic patterns at various temporal and frequency resolutions. The extracted feature maps can be represented as

$$Z_i = \sigma(W_i * S + b_i) \quad (5)$$

where W_i and b_i denote the convolutional parameters of the i -th branch, $*$ represents the convolution operation, $\sigma(\cdot)$ denotes the nonlinear activation function, and Z_i represents the extracted feature representation from the corresponding scale. The multi-scale features are subsequently aggregated to construct a comprehensive spectral representation

$$Z_{ms} = \text{Concat}(Z_1, Z_2, \dots, Z_n) \quad (6)$$

where Z_{ms} denotes the fused multi-scale spectral feature. By integrating information from different receptive fields, the encoder is able to capture both local acoustic variations and global structural characteristics embedded in music signals.

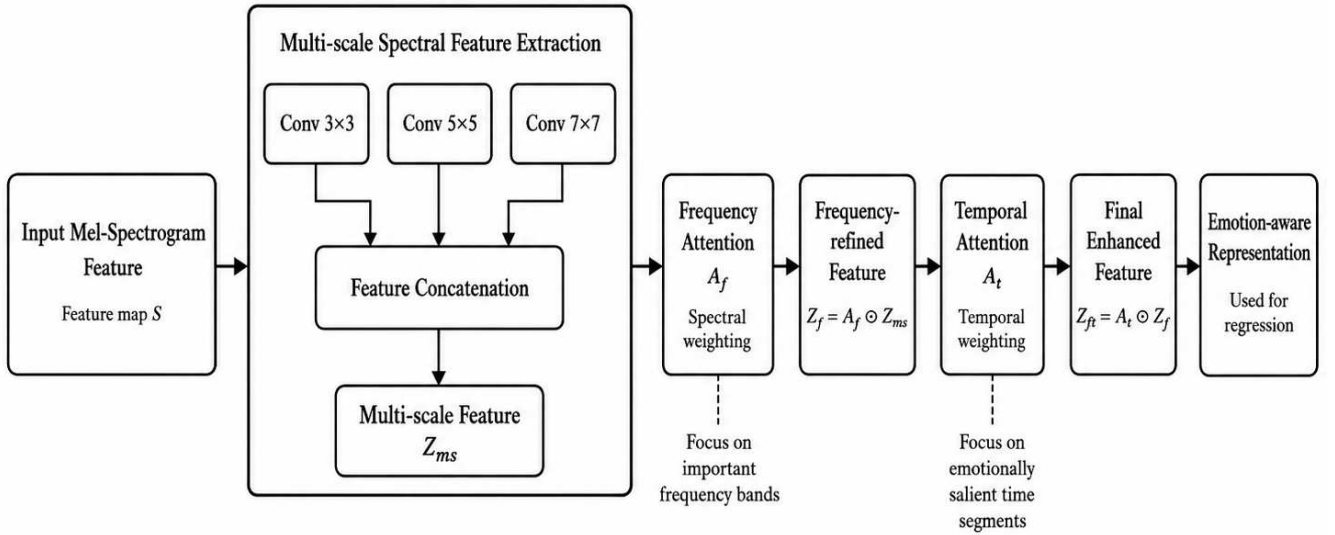


Figure 2. Detailed architecture of the multi-scale frequency-temporal attention mechanism in MASAN-Net. The module extracts hierarchical spectral representations through multi-scale convolution operations and adaptively enhances feature representations using frequency and temporal attention mechanisms.

After obtaining the multi-scale spectral representation, a frequency attention mechanism is introduced to adaptively model the importance of different frequency components. Since different frequency regions may correspond to distinct acoustic characteristics, the frequency attention module learns feature weights along the spectral dimension and emphasizes emotion-relevant frequency patterns. The attention calculation can be formulated as

$$A_f = \sigma(W_f Z_{ms} + b_f) \quad (7)$$

where A_f represents the learned frequency attention weights, and W_f and b_f are trainable parameters. The refined frequency-aware representation is obtained by applying the attention weights to the original spectral features

$$Z_f = A_f \odot Z_{ms} \quad (8)$$

where $\odot$ denotes element-wise multiplication.

Although frequency information provides important acoustic cues, emotional expression in music is also strongly related to temporal evolution and dynamic changes. Therefore, a temporal attention mechanism is further introduced to capture emotionally significant temporal patterns. The temporal attention weights are calculated as

$$A_t = \sigma(W_t Z_f + b_t) \quad (9)$$

where A_t represents the temporal attention map. The final frequency-temporal enhanced representation is obtained by combining both attention dimensions

$$Z_{ft} = A_t \odot Z_f \quad (10)$$

where Z_{ft} denotes the final emotion-oriented spectral representation. Through the joint optimization of multi-scale convolutional feature extraction and frequency-temporal attention enhancement, MASAN-Net is capable of selectively emphasizing informative spectral regions while suppressing irrelevant acoustic variations. This design enables the model to establish a more effective relationship between complex music structures and continuous emotional representations.

3.4 Emotion Quantification and Interpretability Analysis

After the frequency-temporal attention representation learning process, the enhanced feature representation is further transformed into continuous emotional predictions through a regression-based quantification module. The overall emotion quantification pipeline is illustrated in Fig. 3. The attention-enhanced feature representation Z_{ft} is first aggregated through global pooling to obtain compact feature vectors. Subsequently, an MLP regression module learns the nonlinear mapping between acoustic representations and emotional states, producing continuous Valence and Arousal estimations. This regression-based design enables MASAN-Net to model the gradual variations of musical emotions rather than assigning samples into discrete categories. Furthermore, the proposed framework incorporates an attention-based interpretability strategy to analyze the relationship between learned representations and emotional predictions. The generated attention maps provide insights into the spectral-temporal regions that contribute to the final emotion estimation. This analysis improves the transparency of MASAN-Net and helps reveal the acoustic characteristics associated with perceived emotional responses.

The final feature representation Z_{ft} is first aggregated through a global feature pooling operation and then mapped into the emotional space through a multilayer perceptron (MLP) regression head. The prediction process can be formulated as

$$\hat{y} = W_o \phi(W_h Z_{ft} + b_h) + b_o \quad (11)$$

where W_h and W_o represent the learnable weights of the hidden and output layers, b_h and b_o denote the corresponding biases, $\phi(\cdot)$ represents the nonlinear activation function, and $\hat{y} = [\hat{v}, \hat{a}]$ represents the predicted Valence and Arousal values. By optimizing the regression objective, the proposed framework is able to learn the continuous relationship between spectral characteristics and emotional perception.

The training objective of MASAN-Net is defined based on the mean squared error between the predicted emotional values and the ground-truth annotations. The regression loss is formulated as

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (12)$$

where N denotes the number of training samples, y_i represents the ground-truth emotional value, and $\hat{y}_i$ denotes the predicted value. Minimizing this objective encourages the model to generate accurate continuous emotional representations while preserving the subtle variations of musical affective expression.

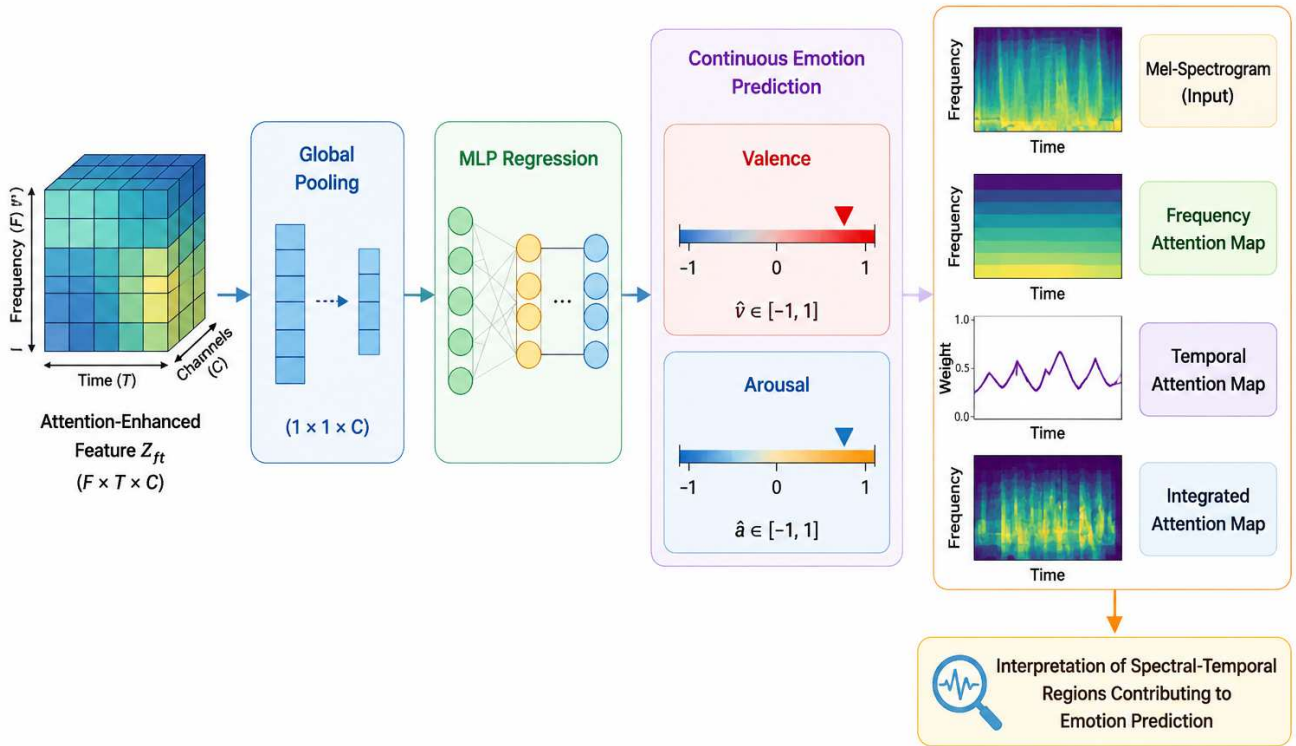


Figure 3. Emotion quantification and interpretability pipeline of MASAN-Net. The attention-enhanced feature representation is transformed into continuous Valence and Arousal predictions through global pooling and MLP regression, while attention visualization reveals the spectral-temporal regions contributing to emotional estimation.

In addition to prediction accuracy, interpretability is an important consideration for deep learning-based music analysis models. To investigate how MASAN-Net identifies emotion-related acoustic patterns, the attention weights generated by the frequency-temporal attention modules are utilized to construct attention maps. The learned frequency attention map and temporal attention map are projected back to the original spectrogram space to visualize the contribution of different spectral regions during emotion prediction. The attention representation can be expressed as

$$M = A_t \odot A_f \quad (13)$$

where M represents the integrated attention map, and A_t and A_f denote the temporal and frequency attention weights, respectively. Higher attention values indicate that the corresponding time-frequency regions contribute more significantly to the emotional prediction process. Through this interpretability analysis, MASAN-Net not only provides quantitative emotion estimation but also reveals the underlying spectral characteristics associated with emotional expression in music.

4 Experimental Setup

The overall experimental procedure adopted in this study is summarized in Fig. 4. Music datasets are prepared and transformed into suitable time-frequency representations. Subsequently, the processed features are used to train MASAN-Net, followed by continuous emotion prediction and attention-based interpretability analysis. The model performance is evaluated through regression metrics, ablation experiments, cross-dataset validation, and statistical stability analysis.

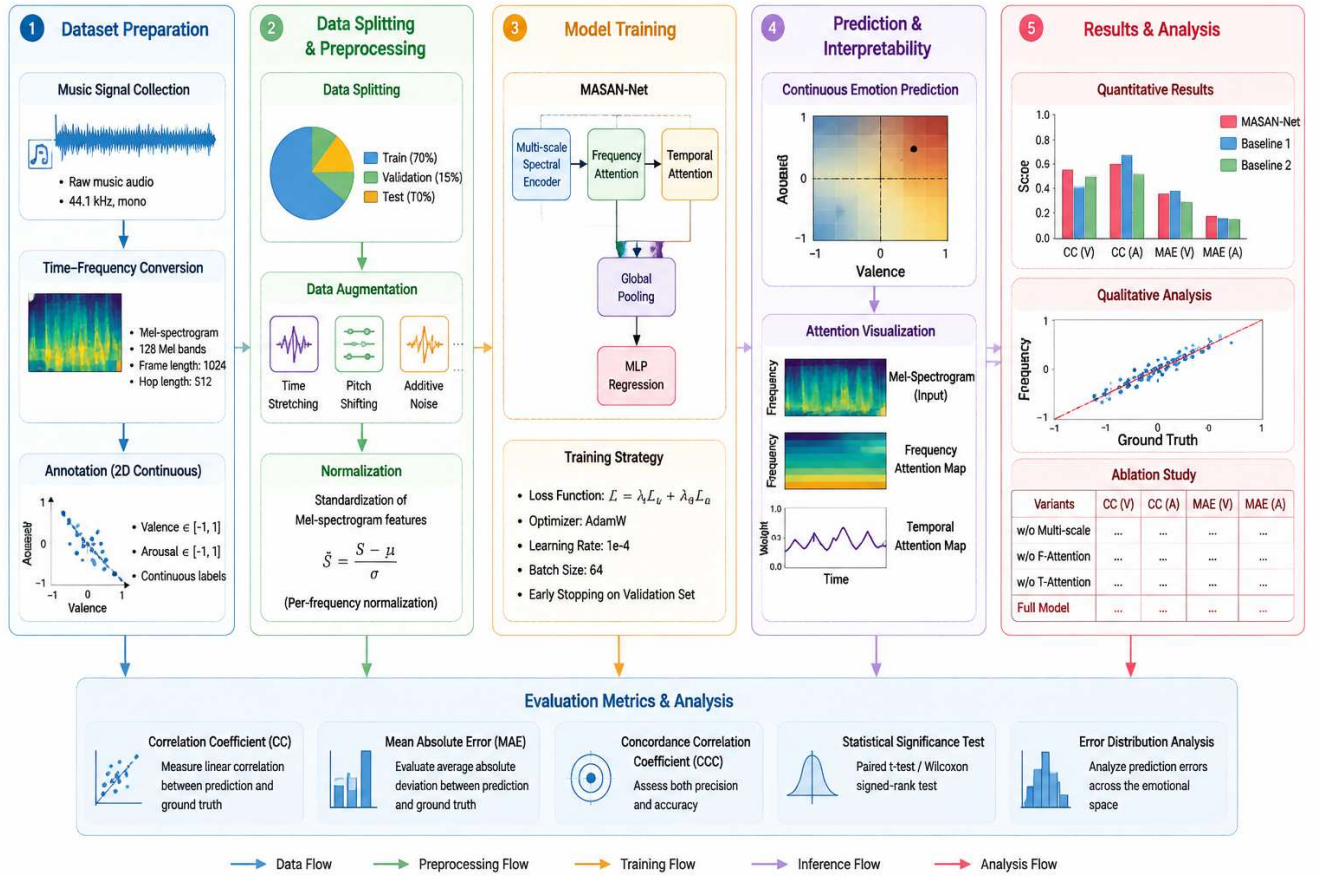


Figure 4. Experimental pipeline of MASAN-Net. The workflow includes dataset preparation, data preprocessing, model training, continuous emotion prediction with interpretability analysis, and comprehensive evaluation using multiple experimental strategies.

4.1 Dataset Description

To evaluate the effectiveness of the proposed MASAN-Net framework for continuous music emotion quantification, two publicly available music datasets are employed in this study, including the Database for Emotional Analysis of Music (DEAM)³³ and the

MTG-Jamendo dataset³⁴. These datasets are selected because they provide large-scale music recordings with emotion-related annotations, enabling the evaluation of deep learning models for audio-based affective computing tasks. The DEAM dataset is used as the primary benchmark dataset for model training and evaluation because it provides continuous emotional annotations in the Valence-Arousal space, which is consistent with the objective of this study. Unlike categorical emotion recognition tasks that assign music pieces into predefined emotional classes, the continuous Valence-Arousal representation describes emotional states as numerical dimensions, allowing models to capture subtle variations in musical expression. Each audio sample in DEAM is associated with affective annotations reflecting perceived emotional characteristics, providing an appropriate foundation for investigating the relationship between spectral structures and emotional responses. In this study, the audio recordings from DEAM are transformed into Mel-spectrogram representations and used as input for MASAN-Net to learn hierarchical frequency-temporal patterns related to emotional perception. In addition to the main experimental evaluation, the MTG-Jamendo dataset is adopted as an external validation dataset to further examine the generalization capability of the proposed framework. Compared with the DEAM dataset, MTG-Jamendo contains a broader collection of music recordings covering diverse musical styles and acoustic characteristics, making it suitable for evaluating whether the learned spectral representations can be transferred beyond the original training distribution. Although the annotation strategy and task definition of MTG-Jamendo differ from those of DEAM, its diverse music content provides additional evidence regarding the robustness of the proposed model in handling complex audio signals. Therefore, the DEAM dataset is mainly utilized for supervised continuous emotion regression, while MTG-Jamendo is employed to assess the adaptability of MASAN-Net under a more diverse music environment. The combination of these two datasets provides a comprehensive evaluation protocol for the proposed framework. The DEAM dataset enables detailed analysis of emotion prediction performance through quantitative regression metrics, while the MTG-Jamendo dataset provides an additional perspective on model robustness and generalization. By conducting experiments across datasets with different musical characteristics, this study aims to verify that MASAN-Net can effectively learn emotion-related spectral representations rather than relying on dataset-specific patterns. This experimental design supports a more reliable assessment of the proposed attention-assisted spectral representation learning framework for intelligent music emotion quantification.

4.2 Data Preprocessing

To provide consistent input representations for MASAN-Net, all audio samples are processed through a unified preprocessing pipeline before model training and evaluation. Since raw audio signals contain redundant information and variations caused by recording conditions, the original waveforms are first normalized to reduce amplitude differences among different music recordings. The processed audio signals are then transformed from the time domain into time-frequency representations, which preserve both spectral characteristics and temporal evolution information. In this study, Mel-spectrograms are adopted as the primary input representation because they provide a perceptually meaningful frequency representation and have been widely applied in audio understanding tasks. The Mel-spectrogram can be generated by applying short-time Fourier transform (STFT) followed by Mel filter bank conversion, which can be formulated as

$$M(m, t) = \sum_{k=0}^{K-1} H(m, k) |X(k, t)|^2 \quad (14)$$

where $X(k, t)$ represents the short-time Fourier transform of the audio signal, $H(m, k)$ denotes the Mel filter bank, m indicates the Mel frequency index, and t represents the temporal frame. The resulting Mel-spectrogram captures frequency-dependent acoustic patterns while maintaining temporal information required for emotion-related analysis. To ensure compatibility with the convolutional feature extraction module, all generated spectrogram representations are converted into unified input dimensions. This process enables MASAN-Net to learn consistent hierarchical spectral features across different music samples. During preprocessing, the same transformation strategy is applied to both the DEAM dataset and the MTG-Jamendo dataset to reduce distribution differences caused by feature extraction procedures. For model training, the DEAM dataset is divided into training and testing subsets following a consistent experimental protocol, while the MTG-Jamendo dataset is processed independently and used for external validation to evaluate the generalization capability of the proposed framework. To improve the robustness of the learned representations, the preprocessing pipeline preserves the essential frequency and temporal structures of music signals rather than applying aggressive information reduction. This design allows the subsequent multi-scale convolutional encoder and frequency-temporal attention modules to effectively capture emotion-related acoustic patterns. Through this standardized preprocessing procedure, the input data maintain sufficient spectral information for continuous Valence-Arousal prediction and provide a reliable foundation for evaluating the performance of MASAN-Net.

4.3 Implementation Details

The proposed MASAN-Net framework is implemented using the PyTorch 2.1.0 deep learning framework and trained on a workstation equipped with an NVIDIA RTX 3090 GPU with 24 GB memory. The experiments are conducted under the CUDA

12.1 environment to accelerate model training and ensure computational efficiency. All experiments are performed using the same hardware and software configurations to provide a fair comparison between MASAN-Net and the baseline methods. During the training process, the AdamW optimizer is adopted to update the network parameters due to its effectiveness in optimizing deep neural networks with complex architectures. The initial learning rate is set to 1×10^{-4} , and a cosine annealing learning rate scheduler is employed to gradually adjust the learning rate during training. The batch size is set to 32, and the maximum training epoch is configured as 100. An early stopping strategy is applied according to the validation performance to prevent overfitting and select the optimal model parameters. Since the proposed task is formulated as a continuous emotion regression problem, the mean squared error (MSE) loss is utilized as the optimization objective. During training, the model learns the mapping relationship between Mel-spectrogram representations and continuous Valence-Arousal annotations by minimizing the difference between predicted and ground-truth emotional values. The optimization process can be expressed as updating model parameters through gradient-based backpropagation until convergence. To ensure the reliability of experimental results, each model is independently trained five times with different random initialization settings. The final performance is reported as the mean value with standard deviation across multiple runs. All baseline methods are implemented under the same experimental conditions, including identical preprocessing procedures, training strategies, and evaluation protocols. This experimental configuration provides a consistent and reproducible environment for evaluating the effectiveness and robustness of the proposed MASAN-Net framework for intelligent music emotion quantification.

4.4 Baseline Methods

To comprehensively evaluate the effectiveness of MASAN-Net, several representative baseline methods are selected for comparison. These baseline models cover different categories of approaches, including traditional machine learning methods, convolutional neural network-based models, recurrent neural network-based models, and attention-based architectures. All baseline models are trained and evaluated using the same preprocessed audio representations and experimental protocols as the proposed framework to ensure a fair comparison. Support Vector Machine (SVM) and Random Forest (RF) are selected as traditional machine learning baselines. These methods represent conventional approaches that rely on manually extracted or predefined acoustic features for music emotion analysis. SVM is widely used for audio classification and regression tasks due to its strong generalization capability in high-dimensional feature spaces, while Random Forest provides an ensemble learning strategy that can capture nonlinear relationships between acoustic features and emotional responses. Although these methods do not perform automatic feature extraction, they provide important references for evaluating the effectiveness of deep representation learning. A Convolutional Neural Network (CNN) model is included as a deep learning baseline to investigate the contribution of automatic spectral feature extraction. CNN-based methods have demonstrated strong performance in audio analysis tasks by learning local patterns from time-frequency representations. However, conventional CNN architectures mainly focus on local spectral structures and may have limited capability in modeling long-range temporal dependencies. To further consider temporal information, Convolutional Recurrent Neural Network (CRNN) and CNN-LSTM models are introduced as sequential modeling baselines. CRNN combines convolutional feature extraction with recurrent layers to jointly capture spectral and temporal patterns, while CNN-LSTM integrates CNN-based acoustic representation learning with long-term dependency modeling through LSTM networks. A Transformer-based model is selected as an advanced attention-based baseline. Transformer architectures have shown promising performance in audio understanding by modeling global dependencies through self-attention mechanisms. This baseline enables an evaluation of whether the proposed frequency-temporal attention mechanism can provide additional advantages compared with general attention-based representation learning. Compared with these existing approaches, MASAN-Net further incorporates multi-scale spectral feature extraction and specialized frequency-temporal attention enhancement to adaptively identify emotion-related acoustic patterns from complex music signals. The comparison among these baseline methods provides a comprehensive evaluation of the proposed framework from different perspectives, including traditional feature-based learning, local spectral modeling, sequential dependency modeling, and global attention-based representation learning. This experimental design allows the effectiveness of each component in MASAN-Net to be systematically analyzed for continuous music emotion quantification.

4.5 Evaluation Metrics

Since the proposed task is formulated as a continuous emotion quantification problem, the performance of MASAN-Net is evaluated using regression-based metrics rather than classification accuracy measures. Four widely adopted evaluation metrics are employed to comprehensively assess the prediction performance, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (PCC). These metrics evaluate different aspects of prediction quality, including absolute prediction error and the consistency between predicted and ground-truth emotional values.

Mean squared error measures the average squared difference between the predicted emotional values and the ground-truth annotations. It is sensitive to larger prediction deviations and is defined as

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (15)$$

where N represents the number of samples, y_i denotes the ground-truth emotional value, and $\hat{y}_i$ represents the predicted value. A lower MSE indicates that the model produces predictions closer to the reference annotations.

Root mean squared error is calculated as the square root of MSE and provides an error measurement in the same scale as the original emotional values. It is formulated as

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (16)$$

Compared with MSE, RMSE provides a more intuitive interpretation of prediction deviation and maintains sensitivity to large errors.

Mean absolute error calculates the average absolute difference between predictions and ground-truth values, which provides a more robust measurement of overall prediction accuracy. The MAE is defined as

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (17)$$

where smaller MAE values indicate better prediction performance and lower deviation from human emotional annotations.

In addition to error-based measurements, Pearson correlation coefficient is employed to evaluate the linear correlation between predicted and actual emotional values. This metric reflects whether the model can accurately capture the variation trend of emotional states and is defined as

$$PCC = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}} \quad (18)$$

where $\bar{y}$ and $\bar{\hat{y}}$ represent the mean values of the ground-truth and predicted emotional scores, respectively. A higher PCC value indicates stronger consistency between the predicted emotional trajectory and the actual affective perception.

For all regression metrics, lower values of MSE, RMSE, and MAE indicate better prediction accuracy, while higher PCC values represent stronger correlation with human-provided emotional annotations. By combining error-based and correlation-based evaluation criteria, the effectiveness of MASAN-Net can be comprehensively assessed from both quantitative accuracy and emotional representation consistency perspectives.

5 Results and Discussion

5.1 Overall Performance Comparison

The overall performance comparison between MASAN-Net and the selected baseline methods on the DEAM dataset is presented in Table 1. Traditional machine learning approaches, including SVM and Random Forest, obtain relatively limited performance, with PCC values below 0.45 for both Valence and Arousal prediction. This limitation is mainly caused by their dependence on manually designed acoustic representations, which cannot fully capture the hierarchical spectral structures and temporal variations contained in music signals. Deep learning-based methods achieve substantial improvements by automatically learning discriminative representations from Mel-spectrogram inputs. The CNN model improves the Valence PCC to 0.486 and the Arousal PCC to 0.526, demonstrating the effectiveness of convolutional feature extraction for modeling local spectral patterns. However, its performance remains limited due to the lack of temporal dependency modeling. By introducing sequential representation learning, CRNN and CNN-LSTM further improve the PCC values, reaching 0.536/0.574 and 0.558/0.601 for Valence and Arousal, respectively. These results confirm the importance of temporal dynamics in continuous music emotion prediction. The Transformer baseline achieves competitive performance with PCC values of 0.581 and 0.624, benefiting from self-attention-based global dependency modeling. Nevertheless, conventional Transformer architectures are not specifically optimized for the frequency-temporal characteristics of music spectrograms and may treat different spectral components with insufficient task-specific adaptation. Compared with all baseline approaches, MASAN-Net achieves the best performance

on both emotional dimensions, obtaining RMSE values of 0.212 and 0.190, together with PCC values of 0.636 and 0.681 for Valence and Arousal prediction. Compared with the strongest Transformer baseline, MASAN-Net reduces the Valence RMSE from 0.228 to 0.212 and improves PCC by 0.055, while the Arousal RMSE decreases from 0.207 to 0.190 with a PCC improvement of 0.057. The performance gains can be attributed to the proposed multi-scale spectral representation learning and frequency-temporal attention enhancement strategy. The multi-scale convolutional module enables MASAN-Net to capture acoustic patterns from different receptive fields, while the frequency and temporal attention mechanisms adaptively emphasize emotion-related spectral regions and dynamic musical variations. These results demonstrate that the proposed framework establishes a more effective mapping between complex acoustic structures and continuous emotional states, providing a robust solution for intelligent music emotion quantification.

Table 1. Performance comparison of different methods on the DEAM dataset for continuous emotion quantification. The best results are highlighted in bold.

Method	Valence				Arousal			
	MSE ↓	RMSE ↓	MAE ↓	PCC ↑	MSE ↓	RMSE ↓	MAE ↓	PCC ↑
SVM ³⁵	0.084	0.290	0.228	0.362	0.073	0.270	0.214	0.401
Random Forest ³⁶	0.078	0.279	0.219	0.395	0.068	0.261	0.205	0.432
CNN ³⁷	0.064	0.253	0.198	0.486	0.055	0.235	0.181	0.526
CRNN ³⁸	0.058	0.241	0.187	0.536	0.049	0.221	0.169	0.574
CNN-LSTM ³⁹	0.055	0.235	0.182	0.558	0.046	0.214	0.163	0.601
Transformer ⁴⁰	0.052	0.228	0.176	0.581	0.043	0.207	0.157	0.624
MASAN-Net	0.045	0.212	0.163	0.636	0.036	0.190	0.143	0.681

5.2 Overall Average Performance

To provide a comprehensive evaluation of continuous music emotion quantification capability, the average performance across Valence and Arousal dimensions is summarized in Table 2. Compared with the dimension-specific results reported in Table 1, the average evaluation provides an integrated measurement of prediction accuracy and emotional consistency across different affective dimensions. The results show that traditional machine learning approaches achieve relatively limited overall performance, with average PCC values of 0.382 and 0.414 for SVM and Random Forest, respectively. This indicates that handcrafted acoustic representations are insufficient to fully describe the complex spectral and temporal patterns associated with continuous emotional variations. Deep learning-based models demonstrate consistent improvements by learning discriminative representations directly from Mel-spectrogram inputs. CNN achieves an average PCC of 0.506, confirming the effectiveness of convolutional feature extraction for capturing local spectral characteristics. However, its average RMSE of 0.244 remains higher than sequential and attention-based approaches, indicating that single-scale convolution cannot sufficiently model the dynamic evolution of musical emotions. By incorporating temporal dependency modeling, CRNN and CNN-LSTM further reduce the average RMSE to 0.231 and 0.225, respectively, while improving the average PCC to 0.555 and 0.580. These results demonstrate the importance of temporal information for continuous emotion representation. The Transformer baseline obtains an average PCC of 0.603 and an average RMSE of 0.218, benefiting from global dependency modeling through self-attention mechanisms. Nevertheless, its performance remains limited because the conventional attention mechanism does not explicitly consider the hierarchical frequency structures and temporal characteristics of music spectrograms. In comparison with all baseline methods, MASAN-Net achieves the best overall performance, obtaining the lowest average RMSE of 0.201, the lowest average MAE of 0.153, and the highest average PCC of 0.659. Compared with the Transformer baseline, MASAN-Net improves average PCC by 0.056 while reducing average RMSE by 0.017. These improvements can be attributed to the joint optimization of multi-scale spectral representation learning and frequency-temporal attention enhancement. The multi-scale feature extraction module enables the model to capture acoustic patterns from different receptive fields, while the attention mechanism adaptively emphasizes emotion-relevant spectral regions and temporal variations. The consistent improvement across all average evaluation metrics demonstrates that MASAN-Net provides a more robust and comprehensive representation learning strategy for continuous music emotion quantification.

5.3 Ablation Study

To investigate the contribution of each component in MASAN-Net, an ablation study is conducted on the DEAM dataset, and the results are reported in Table 3. The complete MASAN-Net consistently achieves the best performance on both Valence and Arousal prediction, demonstrating that the proposed multi-scale spectral representation learning and frequency-temporal attention mechanisms provide complementary improvements for continuous music emotion quantification. The CNN baseline obtains PCC values of 0.486 and 0.526 for Valence and Arousal prediction, respectively, which are substantially lower than

Table 2. Overall average performance comparison of different methods on the DEAM dataset.

Method	Average RMSE ↓	Average MAE ↓	Average PCC ↑
SVM	0.280	0.221	0.382
Random Forest	0.270	0.212	0.414
CNN	0.244	0.190	0.506
CRNN	0.231	0.178	0.555
CNN-LSTM	0.225	0.173	0.580
Transformer	0.218	0.167	0.603
MASAN-Net	0.201	0.153	0.659

the complete MASAN-Net. This performance gap indicates that conventional single-scale convolutional feature extraction is insufficient for modeling the hierarchical acoustic structures of music signals. After removing the multi-scale spectral feature extraction module (MASAN-Net w/o MS), the Valence and Arousal PCC values decrease from 0.636 to 0.602 and from 0.681 to 0.651, respectively. The corresponding RMSE increases by 0.014 and 0.014, confirming that multi-scale representation learning plays a fundamental role in capturing acoustic patterns distributed across different frequency and temporal resolutions. By integrating multiple receptive fields, MASAN-Net can simultaneously preserve local spectral variations and global contextual information. The removal of the frequency attention module results in further performance degradation, with Valence PCC decreasing from 0.636 to 0.614 and Arousal PCC decreasing from 0.681 to 0.662. This demonstrates that adaptive frequency selection is important for identifying emotion-related spectral components. Since different frequency regions contain distinct acoustic characteristics, the frequency attention mechanism enables MASAN-Net to selectively enhance informative spectral patterns while suppressing irrelevant information. Removing the temporal attention module reduces the PCC values to 0.621 and 0.667 for Valence and Arousal, respectively. The performance decrease indicates that temporal dependency modeling is essential for capturing the dynamic evolution of musical emotions. In particular, the larger degradation observed in Arousal prediction suggests that temporal variations play a more important role in representing emotional intensity changes. The consistent improvements achieved by the complete MASAN-Net verify that multi-scale spectral learning, frequency attention, and temporal attention contribute independently while working synergistically. The ablation results demonstrate that the proposed architecture effectively integrates hierarchical spectral analysis with adaptive feature refinement, leading to more accurate and robust continuous emotion quantification.

Table 3. Ablation study of different components in MASAN-Net on the DEAM dataset.

Model	Valence		Arousal	
	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
CNN Baseline	0.253	0.486	0.235	0.526
MASAN-Net w/o MS	0.226	0.602	0.204	0.651
MASAN-Net w/o FA	0.220	0.614	0.198	0.662
MASAN-Net w/o TA	0.217	0.621	0.196	0.667
MASAN-Net	0.212	0.636	0.190	0.681

5.4 Cross-dataset Generalization

To further evaluate the generalization capability of MASAN-Net, a cross-dataset validation experiment is conducted on the MTG-Jamendo dataset. Different from the in-domain evaluation on the DEAM dataset, all models are trained using DEAM and directly evaluated on MTG-Jamendo without additional fine-tuning. The results are summarized in Table 4. Since MTG-Jamendo contains diverse musical styles and different acoustic distributions, the performance degradation compared with the DEAM evaluation is expected and provides a more challenging assessment of model robustness. The CNN baseline achieves PCC values of 0.421 and 0.468 for Valence and Arousal prediction, respectively, indicating that models relying mainly on local spectral feature extraction are more sensitive to distribution variations. Although CRNN and CNN-LSTM improve the generalization performance by incorporating temporal dependency modeling, their PCC values remain limited to 0.472/0.523 and 0.495/0.548, respectively. These results suggest that temporal modeling alone cannot fully address the variations in spectral structures and emotional expressions across different music datasets. The Transformer-based model achieves stronger cross-dataset performance, obtaining PCC values of 0.526 and 0.574 for Valence and Arousal, respectively. This improvement demonstrates the advantage of global dependency modeling through self-attention mechanisms. However, the performance remains affected by the absence of music-specific frequency-temporal adaptation, which limits its ability to

selectively capture transferable emotion-related acoustic patterns. Compared with all baseline methods, MASAN-Net achieves the best generalization performance on MTG-Jamendo, with RMSE values of 0.231 and 0.211 and PCC values of 0.571 and 0.621 for Valence and Arousal prediction. Compared with the Transformer baseline, MASAN-Net improves PCC by 0.045 and 0.047 on the two emotional dimensions while reducing RMSE by 0.016 and 0.015, respectively. These improvements demonstrate that the proposed framework learns more transferable spectral representations under distribution shifts. The enhanced generalization capability can be attributed to the joint design of multi-scale spectral representation learning and frequency-temporal attention mechanisms. The multi-scale feature extraction module reduces the dependence on specific spectral resolutions by capturing acoustic patterns from different receptive fields, while the attention mechanisms dynamically emphasize emotion-related frequency regions and temporal variations. These results confirm that MASAN-Net provides a robust representation learning strategy for practical music emotion quantification scenarios.

Table 4. Cross-dataset generalization performance on the MTG-Jamendo dataset. Models are trained on the DEAM dataset and evaluated on the MTG-Jamendo dataset.

Method	Valence		Arousal	
	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑
CNN	0.276	0.421	0.259	0.468
CRNN	0.261	0.472	0.241	0.523
CNN-LSTM	0.254	0.495	0.233	0.548
Transformer	0.247	0.526	0.226	0.574
MASAN-Net	0.231	0.571	0.211	0.621

5.5 Statistical Stability Analysis

To evaluate the stability and reliability of the proposed framework, a statistical analysis based on five independent training runs with different random initialization settings is conducted. The average performance and standard deviation of RMSE and PCC values are reported in Table 5. This analysis provides an additional evaluation perspective by examining whether the performance improvements are consistent across different optimization processes rather than resulting from a particular initialization condition. The results show that all evaluated models maintain relatively stable performance across repeated experiments. However, noticeable differences can be observed in the variation magnitude among different architectures. The CNN-LSTM model achieves competitive prediction capability but exhibits the largest performance fluctuation, with RMSE standard deviations of 0.009 and 0.008 and PCC standard deviations of 0.014 and 0.013 for Valence and Arousal prediction, respectively. This variation can be attributed to the sequential optimization characteristics of recurrent architectures, where the learned temporal representations may be sensitive to parameter initialization. The Transformer-based model demonstrates improved stability compared with CNN-LSTM, reducing the PCC standard deviation to 0.011 and 0.010 for Valence and Arousal prediction. This improvement benefits from the self-attention mechanism, which provides more effective global dependency modeling. Nevertheless, the absence of music-specific frequency-temporal adaptation may still introduce variations when learning complex acoustic patterns. MASAN-Net achieves both the highest prediction accuracy and the lowest performance variation among all evaluated methods. The proposed framework obtains RMSE standard deviations of only 0.004 and 0.003 and PCC standard deviations of 0.006 and 0.005 for Valence and Arousal prediction, respectively. Compared with the Transformer baseline, MASAN-Net reduces PCC variation by 45.5% for Valence and 50.0% for Arousal prediction. The improved stability can be attributed to the joint design of multi-scale spectral representation learning and frequency-temporal attention enhancement. The multi-scale convolutional module provides more comprehensive hierarchical acoustic representations, while the attention mechanisms adaptively refine emotion-related features and reduce the influence of irrelevant spectral variations during optimization. These results demonstrate that MASAN-Net provides reliable and reproducible continuous emotion quantification performance across different training conditions. The consistent improvements observed over multiple independent runs further validate the robustness of the proposed attention-assisted representation learning framework for practical music emotion analysis applications.

5.6 Attention-based Interpretability Analysis

Although deep learning models have achieved promising performance in music emotion quantification, interpretability remains a critical requirement for practical intelligent music analysis applications. To investigate how MASAN-Net identifies emotion-related acoustic patterns, an attention-based interpretability analysis is conducted by visualizing the attention responses generated during the prediction process. The visualization results are presented in Fig. 5, where the original Mel-spectrogram, the frequency attention map, the temporal attention map, and the integrated attention heatmap are compared to reveal the time-frequency regions that contribute most significantly to emotion prediction. MASAN-Net does not assign uniform importance to

Table 5. Statistical significance analysis of different models on the DEAM dataset over five independent runs. Results are reported as mean $\pm$ standard deviation.

Method	Valence		Arousal	
	RMSE $\downarrow$	PCC $\uparrow$	RMSE $\downarrow$	PCC $\uparrow$
CNN-LSTM	0.235 ± 0.009	0.558 ± 0.014	0.214 ± 0.008	0.601 ± 0.013
Transformer	0.228 ± 0.007	0.581 ± 0.011	0.207 ± 0.006	0.624 ± 0.010
MASAN-Net	0.212 ± 0.004	0.636 ± 0.006	0.190 ± 0.003	0.681 ± 0.005

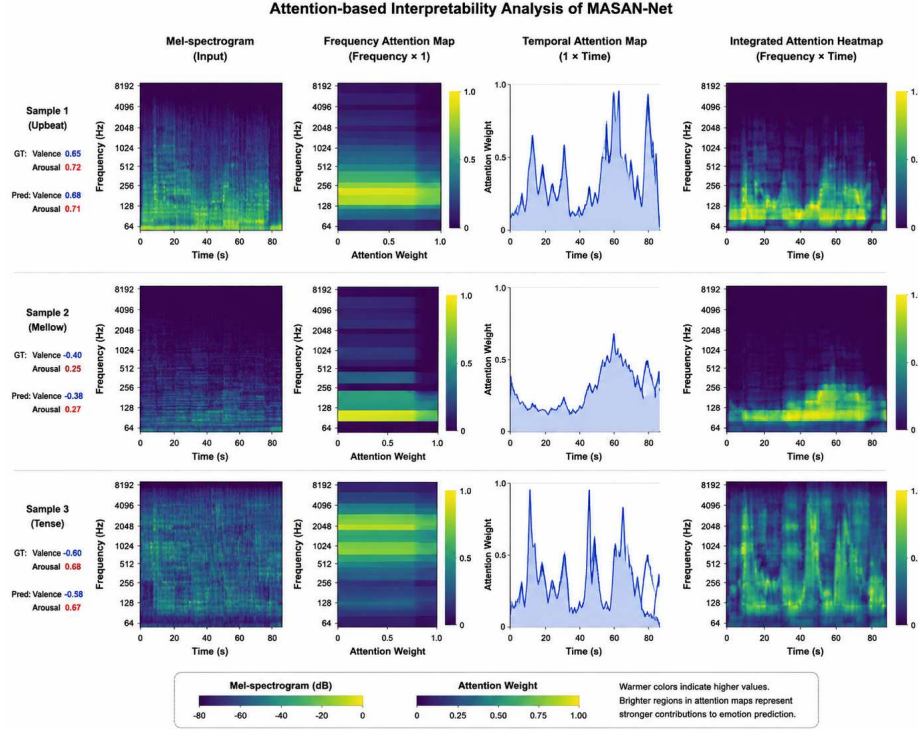


Figure 5. Attention-based interpretability analysis of MASAN-Net. From left to right: original Mel-spectrogram, frequency attention map, temporal attention map, and integrated attention heatmap. Warmer regions indicate higher attention weights and stronger contributions to continuous Valence-Arousal prediction.

all regions of the input spectrogram. Instead, the learned attention distributions selectively emphasize specific spectral bands and temporal segments that are more closely associated with emotional expression. The frequency attention map shows that the model concentrates on informative spectral regions rather than treating all frequency components equally. This behavior indicates that the proposed frequency attention mechanism can effectively distinguish discriminative acoustic cues embedded in music signals, including harmonic structures, timbral variations, and spectral energy patterns that are relevant to emotional perception. Such adaptive frequency selection helps the model suppress less informative components and enhances its capability to learn emotion-oriented spectral representations.

The temporal attention visualization further demonstrates that MASAN-Net dynamically adjusts the importance of different musical moments during emotion prediction. Compared with static feature extraction approaches, the temporal attention mechanism enables the model to focus on emotionally salient segments, such as transitions, expressive changes, and dynamically evolving passages within a music excerpt. This observation suggests that music emotion perception is influenced not only by static spectral characteristics but also by the temporal evolution of acoustic events. By combining the frequency and temporal attention responses, the integrated attention heatmap provides a more comprehensive interpretation of how MASAN-Net establishes the relationship between acoustic structures and continuous emotional states. Regions with higher attention intensity correspond to spectral-temporal patterns that contribute more strongly to the predicted Valence and Arousal values. Overall, the visualization results demonstrate that the proposed framework learns meaningful and interpretable acoustic representations rather than relying on arbitrary feature correlations. This attention-based interpretability analysis provides additional evidence for the

effectiveness of MASAN-Net and further enhances the transparency and practical applicability of the proposed framework.

6 Conclusion

In this study, an attention-assisted multi-scale spectral representation learning framework, termed MASAN-Net, is proposed for intelligent music emotion quantification. Different from conventional music emotion analysis approaches that mainly focus on handcrafted acoustic features or discrete emotion classification, this work formulates emotion prediction as a continuous regression problem in the Valence-Arousal space and aims to establish a more effective relationship between spectral structures and emotional perception. The proposed framework integrates multi-scale convolutional feature extraction with frequency-temporal attention mechanisms to capture hierarchical acoustic characteristics and adaptively identify emotion-related spectral-temporal patterns from music signals.

Extensive experiments conducted on the DEAM dataset demonstrate that MASAN-Net achieves superior performance compared with traditional machine learning methods and advanced deep learning baselines across multiple regression metrics. The performance improvements indicate that multi-scale representation learning and task-specific attention enhancement can effectively improve the modeling capability of complex music structures. Furthermore, the ablation experiments verify the individual contributions of multi-scale feature extraction, frequency attention, and temporal attention, demonstrating that these components provide complementary benefits for continuous emotion quantification. The cross-dataset evaluation on the MTG-Jamendo dataset further confirms the generalization capability of MASAN-Net under different musical characteristics and data distributions. In addition, the attention-based interpretability analysis provides visual evidence of the spectral and temporal regions contributing to emotional prediction, improving the transparency of the proposed framework. Although MASAN-Net achieves promising results, several limitations remain to be addressed. Future research will focus on extending the framework to larger and more diverse music collections, incorporating additional multimodal information such as lyrics, musical structures, and contextual metadata, and exploring more advanced representation learning strategies for complex emotional expressions. These efforts are expected to further improve the robustness, interpretability, and practical applicability of intelligent music emotion analysis systems.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Conceptualization, JL; methodology, JL; software, JL; validation, JL; formal analysis, JL; investigation, JL; data curation, JL; writing—original draft preparation, JL; writing—review and editing, JL; visualization, JL; supervision, JL; funding acquisition, JL; All authors have read and agreed to the published version of the manuscript.

Funding

No Funding

Data Availability Statement

The datasets generated and/or analysed during the current study are available in: <https://github.com/Jiajie0824/MASAN-Net.git>

References

1. Kim, Y. E. *et al.* Music emotion recognition: A state of the art review. In *Proc. ismir*, vol. 86, 937–952 (2010).
2. Jiang, X., Zhang, Y., Lin, G. & Yu, L. Music emotion recognition based on deep learning: A review. *IEEE Access* **12**, 157716–157745 (2024).
3. Zhang, S. *et al.* Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert. Syst. with Appl.* **237**, 121692 (2024).
4. Koh, E. & Dubnov, S. Comparison and analysis of deep audio embeddings for music emotion recognition. *arXiv preprint arXiv:2104.06517* (2021).
5. Hizlisoy, S., Yildirim, S. & Tufekci, Z. Music emotion recognition using convolutional long short term memory deep neural networks. *Eng. science technology, an international journal* **24**, 760–767 (2021).
6. Panda, R., Malheiro, R. & Paiva, R. P. Musical texture and expressivity features for music emotion recognition. In *ISMIR*, 383–391 (2018).
7. Li, X. & Li, X. Atst: Audio representation learning with teacher-student transformer. *arXiv preprint arXiv:2204.12076* (2022).
8. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
9. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).
10. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114 (PmLR, 2019).
11. Lian, H. *et al.* A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy* **25**, 1440 (2023).
12. Purwins, H. *et al.* Deep learning for audio signal processing. *IEEE J. Sel. Top. Signal Process.* **13**, 206–219 (2019).
13. Bathigama, T. H. & Madushika, M. K. S. Music emotion recognition using deep neural networks. Uva Wellassa University of Sri Lanka (2021).
14. Han, D., Kong, Y., Han, J. & Wang, G. A survey of music emotion recognition. *Front. Comput. Sci.* **16**, 166335 (2022).
15. Dong, Y., Yang, X., Zhao, X. & Li, J. Bidirectional convolutional recurrent sparse network (bcrsn): an efficient model for music emotion recognition. *IEEE Transactions on Multimed.* **21**, 3150–3163 (2019).
16. Gong, Y., Khurana, S., Rouditchenko, A. & Glass, J. Cmkd: Cnn/transformer-based cross-model knowledge distillation for audio classification. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2025).
17. Kong, Q. *et al.* Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, Lang. Process.* **28**, 2880–2894 (2020).

18. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. neural information processing systems* **33**, 12449–12460 (2020).
19. Hsu, W.-N. *et al.* Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, language processing* **29**, 3451–3460 (2021).
20. Rouard, S., Massa, F. & Défossez, A. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2023).
21. Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks learning systems* **32**, 4793–4813 (2020).
22. Li, Y., Cao, W., Xie, W., Li, J. & Benetos, E. Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes. *IEEE Transactions on Multimed.* **26**, 1346–1360 (2023).
23. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).
24. Meyer, M., Cavigelli, L. & Thiele, L. Efficient convolutional neural network for audio event detection. *arXiv preprint arXiv:1709.09888* (2017).
25. Qiao, Y., Mu, J., Xie, J., Hu, B. & Liu, G. Music emotion recognition based on temporal convolutional attention network using eeg. *Front. human neuroscience* **18**, 1324897 (2024).
26. Geroulanos, A. & Giannakopoulos, T. Emotion recognition in music using deep neural networks. In *Advances in Speech and Music Technology: Computational Aspects and Applications*, 193–213 (Springer, 2022).
27. Li, Y. *et al.* Mert: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations*, vol. 2024, 12181–12204 (2024).
28. Ehrlich, S., Guan, C. & Cheng, G. A closed-loop brain-computer music interface for continuous affective interaction. In *2017 international conference on orange technologies (ICOT)*, 176–179 (IEEE, 2017).
29. Huang, Q. *et al.* Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415* (2022).
30. Elizalde, B., Deshmukh, S., Al Ismail, M. & Wang, H. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2023).
31. Chen, S. *et al.* Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058* (2022).
32. Agostinelli, A. *et al.* Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).
33. Soleymani, M., Aljanaki, A. & Yang, Y.-H. Deam: Mediaeval database for emotional analysis in music. *Geneva, Switz.* **128**, 131 (2016).
34. Bogdanov, D., Won, M., Tovstogan, P., Porter, A. & Serra, X. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)* (Long Beach, CA, United States, 2019).
35. Han, B.-j., Rho, S., Dannenberg, R. B. & Hwang, E. Smers: Music emotion recognition using support vector regression. In *ISMIR*, 651–656 (2009).
36. Salman, H. A., Kalakech, A. & Steiti, A. Random forest algorithm overview. *Babylonian J. Mach. Learn.* **2024**, 69–79 (2024).
37. Hizlisoy, S., Yildirim, S. & Tufekci, Z. Music emotion recognition using convolutional long short term memory deep neural networks. *Eng. science technology, an international journal* **24**, 760–767 (2021).
38. Lee, S., Hussein, R., Ward, R., Wang, Z. J. & McKeown, M. J. A convolutional-recurrent neural network approach to resting-state eeg classification in parkinson's disease. *J. neuroscience methods* **361**, 109282 (2021).
39. Hizlisoy, S., Yildirim, S. & Tufekci, Z. Music emotion recognition using convolutional long short term memory deep neural networks. *Eng. science technology, an international journal* **24**, 760–767 (2021).
40. Gong, Y., Chung, Y.-A. & Glass, J. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).