Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Music Emotion Recognition for Internet of Things Acoustic Sensing Using Multiscale Convolutional Feature Fusion

Huanzi Li

Lihuanzi111@126.com

School of Art and Design,Yantai University of Science and Technology

Research Article

Keywords: Internet of Things, Acoustic perception, Music emotion recognition, Multi-scale convolution, Feature fusion

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11602062/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Music Emotion Recognition for Internet of Things Acoustic Sensing Using Multiscale Convolutional Feature Fusion

Huanzi Li*
School of Art and Design,Yantai University of Science and Technology,Penglai District,Yantai City,Shandong Province, 265600,P. R. China
Email：Lihuanzi111@126.com

Abstract
A multi-scale convolution feature fusion method was proposed to meet the needs of music emotion recognition under the conditions of noise disturbance, device difference and computing power limitation of acoustic nodes in the Internet of Things. The system collects monophone audio at 44.1 KHZ and 16 bits precision. Each frame is aligned by timestamp, and 128-dimensional Mel-spectrum is generated by filtering, framing and short time Fourier transform. The 1802 pieces of music and valence and arousal labels of the DEAM dataset were used to divide the training, validation and test sets by track at 70%, 15%, and 15%. The model constructs 3×3, 5×5 and 7×7 parallel depthwise separable convolution branches, extracts acoustic features by cross-scale attention, gated residual and temporal convergence, fuses chroma and rhythm description, and completes four types of emotion classification and valence and arousal regression. AdamW is used to optimize the weighted cross entropy, regression and consistency joint loss, and is deployed on edge devices through pruning and eight-bit integer quantization. In the comparison data, the accuracy of the model is 85.56%, the macro average F1 value is 0.856, the processing delay after the completion of the window under the condition of four nodes is 78ms, and the volume of the quantized model is 0.58 megabits. The ablation, noise and cross-device analysis are carried out to provide a basis for multi-node access and low-resource deployment of music emotion recognition.

Key words Internet of Things; Acoustic perception; Music emotion recognition; Multi-scale convolution; Feature fusion

1 Introduction

The combination of artificial intelligence and Internet of things perception technology extends music analysis from offline file processing to emotional discrimination of networked terminals. This process needs to balance the acquisition accuracy and inference efficiency. Musical emotion is not determined by single pitch or loudness. Melody fluctuation, rhythm density, harmony color and timbre change jointly affect the perception of valence and arousal. Louro et al. compared a variety of deep learning music emotion recognition methods, which provides reference for the selection of network structure [1]. Panda et al. sorted out the correspondence between acoustic features and emotional representation, indicating that different descriptors are complementary [2]. In the actual networking scenario, the pickup distance, room reverberation and terminal microphone response will change the time-frequency distribution, and the model trained only on fixed recording conditions is difficult to directly adapt to multi-node acquisition. Therefore, acoustic data sources, feature learning methods, and inference deployment should be considered as a continuous technical process.

Music emotion recognition has gradually shifted from manually designed features to deep representation and joint modeling. Talaghat et al. incorporated audio clips and lyrics into the multimodal network, expanding the information sources of music sentiment analysis [3]. Turchet et al. studied intelligent Musical Instruments with emotional perception and demonstrated the connection between recognition results and online music interaction [4]. This kind of research suggests that music emotion tags can not only serve music retrieval, but also be used for content organization and real-time response of intelligent terminals. At the same time, the data transmission between networked

music devices will introduce additional delay, and the connection mode will also affect the accuracy of node time synchronization. Vignati et al. analyzed the ability of mobile communication networks to support music Internet of things applications, which provided a basis for transmission performance evaluation [5]. When multiple acoustic nodes work at the same time, the recognition system also needs to coordinate the sampling time, network overhead and edge computing load, so as to keep the model output corresponding to the music segment, and reduce the invalid data traffic generated by repeated transmission.

This paper establishes a unified method around music acoustic acquisition and emotion classification in the Internet of things environment: perception nodes collect waveforms and generate timestamped data, edge computing units complete time-frequency transformation and deep feature learning, and classification modules map continuous emotional information into interpretable categories. Three kinds of convolutional receptive fields are used to capture the instantaneous timeliness, local rhythm and long time-frequency structure, and then the redundant channels are suppressed by attention gating, which takes into account the emotion discrimination ability and model complexity. The evaluation is based on public data and samples collected by networked nodes, and the classification quality, noise disturbance, cross-device migration and end-to-end inference cost are jointly investigated to avoid equating offline recognition accuracy with system application performance.

1.1 Research needs of acoustic perception of iot music

Networked music terminals need to convert dispersed collected audio into time-consistent, computable and reusable emotional features. Wiafe et al. studied music emotion classification algorithms, providing a basis for the comparison of different recognition mechanisms [6]. Sams et al. combined convolution, temporal network and text representation to carry out song sentiment analysis [7], and Dutta et al. used convolution and long short-term memory network to describe the characteristics of music sequence [8]. These methods reflect the value of the joint representation of short-time spectrum and context information, but in iot devices, it is necessary to consider the input bandwidth, computing power budget and collection environment differences. In this paper, mono 44.1 KHZ, 16-bit audio is used as a unified input specification, amplitude normalization and timestamp recording are completed at the node side, and spectrum calculation and fusion classification are carried out at the edge side. For different acquisition locations, the device number, sound pressure condition and segment start and end time are retained, so as to analyze the sensitivity of the model to device changes and quantify the resource occupation of terminal storage and transmission. Aswale et al. used deep convolutional networks to study Indian classical music recognition [9], and showed that convolutional representation can extract time-frequency rules related to music categories. When it is introduced into the emotion task, it should also distinguish the music style identification features from the emotional expression features to avoid the bias of the model depending on the recording source or genre cues.

1.2 Technical difficulties of music emotional feature recognition

Musical emotions are characterized by expressions across time scales. Short-term energy and spectral centroid can reflect local intensity and brightness, while beat organization and harmony advancement need longer context to be stable. Yoo et al. used cepstral features and machine learning for music sentiment classification [10], and Santoso et al. used cepstral parameters and nearest neighbor classification for instrumental music sentiment analysis [11], indicating that traditional features can provide a clear acoustic description, but a single statistical scale is difficult to cover complex musical structures. To this end, this paper organizes the MEL spectrum into two-dimensional input, and uses 3 $\times$3, 5$\times$5 and 7$\times$7 convolution kernels to extract texture in parallel, taking into account local and neighborhood dependence, learns the contribution weights of different scales with channel attention, and uses residual connections to maintain subtle band changes. Patil et al. reviewed the techniques, datasets and challenges of music sentiment analysis [12], suggesting that sentiment annotation differences directly affect model evaluation. In response to this feature, this paper uses valence and

arousal to construct a four-quadrant label, divides the training set and test set by tracks rather than slices, and uses class-weighted loss to reduce the imbalance of sample proportion, and records the prediction confidence and category confusion. The end-to-end inference is further constrained by memory, power consumption and transmission rate, so it is necessary to report the recognition performance after quantization and compression synchronously with the computational load.

1.3 Research content and Innovation contribution

This paper designs a hierarchical architecture composed of perception nodes, edge processing units and music emotion classifiers, forming a closed data chain of acquisition, preprocessing, feature fusion and result output. In the data part, the DEAM dataset containing 1802 pieces of music with valence and arousal annotations was used, and the cross-device test was organized by combining node acquisition and recording. The database of emotional music clips established by Strauss et al. provides another reference for the structured selection of emotional materials [13]. In the method part, MEL spectrum and rhythm description are used to construct complementary input, features are extracted by parallel multi-scale convolution, weighted fusion is completed by attention gating, and four types of emotional probabilities are output by the classification head. Model training was divided into tracks at the level of 70%, 15% and 15%, and conventional convolutional networks, single-scale networks and de-fusion modules were set as controls to distinguish the effects of multi-scale modeling and attention weighting with ablation tests. The deployment part records end-to-end delay, peak memory, model volume, and recognition metrics at different signal-to-noise ratios on edge devices, analyzes the resource allocation relationship between acoustic perception and artificial intelligence computing, and evaluates the processing throughput after node scale expansion. The above design extends the algorithm comparison to the network acquisition condition, so that the emotion recognition effect can be evaluated from two levels of classification reliability and terminal execution ability at the same time.

2 Literature Review

The acoustic perception of the Internet of things extends music emotion recognition from offline classification to acquisition, transmission and edge inference. Related research involves time-frequency representation, convolutional classification and end-to-side deployment. Timbre, tempo, and spectral dynamics span different time scales and are altered by device differences and noise. The review needs to take into account both feature representation and terminal computing power constraints. When the sampling frequency and transmission bandwidth of nodes are fixed, the long-term structure required for emotion discrimination will increase the cache overhead, forming a trade-off between recognition accuracy and real-time processing.

Gourisaria et al. compared the Mel-frequency cepstrum coefficient and short-time Fourier transform in audio classification, and detected the classification difference before and after noise reduction. The accuracy of its artificial neural network on the two sound datasets is 91.41% and 91.27% respectively [14]. This is a general sound classification result and cannot be directly extrapolated to music emotion recognition. Zaman et al. systematically collated deep learning audio classification methods, involving convolutional structure, data representation and model evaluation [15]. Different networks have different descriptions of local texture and temporal dependence, and stacked convolutions may not cover emotional changes across sentences. Ranmal et al. proposed a hardware-aware neural architecture search method for edge environmental sound classification, which combined the accuracy of convolution model and device resource consumption into the network structure selection [16]. This method reflects the hardware constraints, but the environmental sound category is not equivalent to the music emotion. The accuracy of the model on ESC-10 and ESC-50 datasets is 96.25%and 81.0%respectively, reflecting the influence of the change of classification difficulty on the effect of the edge model. Mohaimenuzzaman et al. studied the deployment process of deep acoustic network on very low resource devices, and realized sound classification at the microcontroller side through the adaptation of network structure and operating conditions [17]. This

flow helps with lightweight deployment, but the long-term dependencies of the music need to be considered. Gairi et al. reviewed environmental sound recognition on embedded devices and compared deep learning models with terminal implementation methods [18]. Their research suggests that storage, computation and latency should be evaluated simultaneously, rather than just offline accuracy. Mahyub et al. used signal latent subspace to represent environmental sounds and explored a representation method different from conventional manual acoustic features [19]. Latent features are able to compress redundancy, but whether they retain musical emotional cues still needs to be tested. Molina-Moreno et al. constructed a noise monitoring system by using domain adaptation to deal with changes in sound level measurement conditions [20]. This method can be used to analyze device differences, but noise estimation is different from emotion discrimination.

Bonet-Sola et al. combined acoustic event detection with noise levels collected by wireless acoustic sensor networks for acoustic environment comfort prediction [21]. The scheme supports multi-node cooperation, but it needs to unify timestamp and device information. Malche et al. used a small machine learning method to build an iot terminal voice control system, focusing on keyword recognition and real-time response on embedded devices [22]. The speech task requires a short segment, and the music emotion judgment depends on a longer time-frequency context, so the network lightweight cannot simply use the same input window. Scarpiniti et al. used scalogram and convolutional neural network to recognize construction site audio, and used time-frequency images to realize complex sound pattern classification [23]. Scale representation has reference significance for multi-band structure, but music emotion recognition still needs to integrate beat, harmony and dynamic changes to avoid classification driven only by scene noise. Almulla used deep convolutional networks to carry out multimodal emotion recognition research [24]. Different modalities can provide complementary emotional cues, but if the Internet of Things music perception mainly relies on audio input, the difference between cross-scale fusion and cross-modal fusion should be clarified to prevent the two from being mixed into the same mechanism. Turchet et al. summarized the technical development of sound Internet of things and discussed the combination direction of sound computing and interconnected devices [25]. This research broads the system perspective of acoustic perception, but the specific music emotion model still needs to be independently verified under node communication, inference load and delay conditions. See Table 1 for a comparison of related studies.

Table 1 Comparison of related research methods

Reference	Research Setting	Algorithm or Model	Research Focus	Main Contribution	Limitations
[14]	Audio Classification	Cepstral Analysis and Fourier Transform	Time-Frequency Features	Compares noise reduction performance	Not focused on emotion recognition
[15]	Sound Analysis	Deep Learning Review	Network Architectures	Summarizes classification methods	Lacks edge-device validation
[16]	Edge Sensing	Neural Architecture Search	Model Selection	Jointly considers computational constraints	Focuses on environmental sounds
[17]	Microcontrollers	Lightweight Acoustic Network	On-Device Classification	Reduces deployment overhead	Emotional temporal dynamics are not evaluated
[18]	Embedded Systems	Deep Learning Recognition Review	Resource Constraints	Compares deployment approaches	Not focused on music emotion recognition
[19]	Environmental Sound	Latent Subspace	Signal Representation	Extracts latent features	Lacks emotion annotations

		Learning			
[20]	Noise Monitoring	Domain Adaptation	Cross-Condition Transfer	Addresses measurement differences	Outputs non-emotional categories
[21]	Wireless Sensor Networks	Sound Event Detection	Node Collaboration	Integrates sound level information	Primarily focuses on acoustic comfort
[22]	IoT Devices	Tiny Machine Learning	Speech Response	Enables embedded recognition	Limited to short-duration speech tasks
[23]	Construction Sound Fields	Scale-Based Graph Convolution	Time-Frequency Patterns	Classifies complex sounds	Not designed for music segments
[24]	Emotion Analysis	Multimodal Convolution	Emotion Feature Fusion	Integrates multiple input modalities	Does not address cross-scale modeling
[25]	Internet of Sounds	Systematic Review	System Architecture	Integrates sound computing technologies	Lacks a task-specific model

Research gap: Existing research has accumulated technology in audio feature design, edge classification and sound iot architecture, but there is still a lack of evaluation for cross-scale feature association of music emotion combined with Internet of things acquisition conditions. The single-scale network is easy to weaken the slowly changing melody and harmony cues, and the direct concatenation of different level features may add noise and redundant channels. Model compression in turn changes the ability to retain fine-grained spectral information. Taking acoustic node data as input, this paper uses parallel multi-scale convolution and attention gated fusion, and combines track-level data division, noise perturbation, cross-device migration and end-to-end delay testing to analyze the relationship between recognition performance and resource overhead. Under a fixed sampling period, the network transmission delay and the model inference delay were measured respectively.

3 Materials and Methods

This study designs the music emotion recognition process of iot acoustic perception and edge artificial intelligence collaboration. The node collects the music waveform and adds the timestamp, the gateway completes the synchronization verification, and the edge device performs spectrum processing and multi-scale convolution coding. The model extracts timbre, rhythm and cross-band information, and outputs the acoustic representation for subsequent emotion classification.

3.1 iot acoustic sensing nodes and data acquisition

The system has a three-layer structure of acoustic perception node, edge gateway and model inference service. The node consists of an omnidirectional digital microphone, a digital audio interface, an embedded processor and a wireless communication module, which uniformly collects monaural pulse code modulated audio at 44.1 KHZ sampling rate and 16 bit accuracy. Nodes add device identification, collection timestamp and sequence number to data, and set circular buffer to alleviate network congestion. An audio payload with 882 sampling points is formed every 20 ms, and the effective data size of a single frame is 1764 bytes. The raw audio data rate is about 706 kilobits per second, and the gateway configures the transmission buffer and access capacity accordingly. Each node transmits through the wireless local area network, records the packet loss mark and arrival time in the data frame, and retransmission is only used for offline archiving flow, and real-time reasoning flow limits

the waiting time to reduce the stacking delay [26].

The edge gateway establishes an independent time series cache according to the device identification, calibrates the acquisition time through the reference clock, and detects duplicate frames, missing frames and order inversion according to the sequence number. For the short missing segments, only the validity mask is added, and the spectrum is not filled by the unrecorded sound. A segment is marked as low-confidence input when successive losses reach a preset threshold. The system also saves the node sound pressure level, acquisition distance group, firmware version and environmental noise level for subsequent cross-device hierarchical analysis. The test end can be connected with microphones with different directivity and frequency response. During the acquisition process, the music source, playback gain and the start and end time of the segment can be controlled, and the difference between equipment and content can be separated. The edge processing and classification program is run in an independent thread, and the original data can be deleted according to the access policy after completing the time-frequency transformation, and only the necessary features and operation logs are saved [27]. The data interface is used to connect the subsequent models, which supports the statistics of effective frame ratio, link delay and throughput by node.

3.2 Music acoustic signal preprocessing and emotion annotation

Before the acoustic data enters the model, the continuous waveform is restored according to the device timestamp, the duplicate data packets are eliminated, and the missing interval is located according to the validity mask. Dc bias cancellation, amplitude normalization, and bandpass filtering were performed on the available segments, and the filter passband was set from 40 Hz to 16 KHZ to mitigate low-frequency mechanical noise and high-frequency electronic noise effects. The original waveform is retained for rechecking to prevent excessive filtering from changing the timbre. For 44.1 KHZ audio, 1024 points Hanning window framing and 512 points frame shift are used. The duration of a single frame is about 23.2 ms, and the interval between adjacent frames is about 11.6 ms. The short time Fourier transform (STFT) was used to generate the complex spectrum, and the power spectrum was calculated and mapped to 128 sets of MEL filters, and then the logarithm was taken to obtain the two-dimensional acoustic matrix with time axis and frequency axis. The mean and standard deviation of each frequency band are estimated according to the training set, and then the same transformation is performed on the validation set and the test set to prevent statistical information leakage from the test set [28].

Emotion supervision uses two continuous dimensions, valence and arousal, to describe the positive degree and activation intensity of musical expression respectively. The DEAM dataset containing 1802 pieces of music was selected as the public training data, the continuous labels were linearly mapped to the interval from zero to one, and 0.5 was used as the division threshold of the two dimensions to form four labels: high-valence high-arousal, high-valence low-arousal, low-valence high-arousal, and low-valence low-arousal. For annotations with time axis, the mean value was calculated according to the time range covered by the audio segment. For the data with only segment-level annotation, the valid label of the track is used, and the manual label is not reversed by the model prediction result. Boundary samples near the threshold are encountered while retaining both continuous scores and discrete categories, facilitating the analysis of label uncertainty.

The sample processing takes the track as the minimum division unit, and assigns training, validation and testing tracks according to the proportion of 70%, 15% and 15%. Different segments of the same track are not allowed to cross the set. The segment length is uniform to 8 seconds. In the training phase, an overlapping window is constructed with a step of 4 seconds, and a non-overlapping window is used for validation and testing to avoid repeating segments to enlarge the evaluation sample size. Controllable background noise, slight gain change and equipment frequency response disturbance were added to the training set. The noise level was recorded according to the signal-to-noise ratio level, and a pure test subset was retained to distinguish the effect of the algorithm and the influence of acquisition conditions. All tags, segment numbers, node numbers and preprocessing parameters were used to establish corresponding indexes, and subsequent models performed repeatable data loading and error attribution.

3.3 Multi-scale convolution acoustic feature extraction

The performance of music emotion in frequency band and time has different scales. The instantaneous onset and overtone distribution form local texture, while the beat period and continuous timbral variation correspond to a wider time-frequency region. In order to reduce the receptive field bias caused by a single convolution kernel, we design a shared input projection and three parallel depth-wise separable convolution branches to map the preprocessed 128-dimensional Mel-spectrum into a unified channel space. The network takes 8-second segments as analysis units, and the convolution calculation is expanded along the frequency axis and the time axis simultaneously to extract emotional acoustic features under the condition of maintaining the same output resolution.

In order to compare textures at different time scales, the aligned MEL spectra are fed into three convolution branches to extract timbre, rhythm and frequency band associations, which are expressed as follows:

$$\begin{aligned} H_s &= \phi\big(BN\big[P_s *_{1\times1} DWConv_{k_s,d_s}(X)\big]\big), \\ (k_s, d_s) &\in \{(3,1),(5,3),(7,6)\}, \quad s \in \{1,2,3\} \end{aligned} \tag{1}$$

Where, X represents the input tensor after shared projection, T and F are the number of time frames and the number of MEL frequency bands, and C is the number of channels. s represents the scale branch, and its convolution kernel size is 3, 5, and 7 in turn. DWConv denotes per-channel convolution, P denotes per-point convolution, BN and ϕ denote batch normalization and nonlinear activation, respectively. The input projection is configured with 32 channels, and the output of each branch is 64 channels. The convolution uses symmetric filling to keep the time and frequency dimensions unchanged. The three-way convolutional dilation rates of one, three, and six correspond to receptive fields of three, thirteen, and thirty-seven time frames, covering structures of about thirty-five to four hundred and thirty milliseconds, respectively. The tone head and rhythm changes are extracted while maintaining the output size. Compared with directly imposing large convolution kernels on all channels, this structure reduces repeated computations across channels, and allows short phonotactic heads, persistent overtones, and slower spectral changes to enter the same feature coordinate system.

Schemata unfold three-way convolutions over the same input spectrum and connect gated residual units with cross-scale weights. The shared projection retains the baseline representation of the input, and the residual branch sends it to the fusion unit to avoid weakening the frequency band information with low energy but emotional recognition value after multiple convolutions. The network structure is shown in Fig. 1.

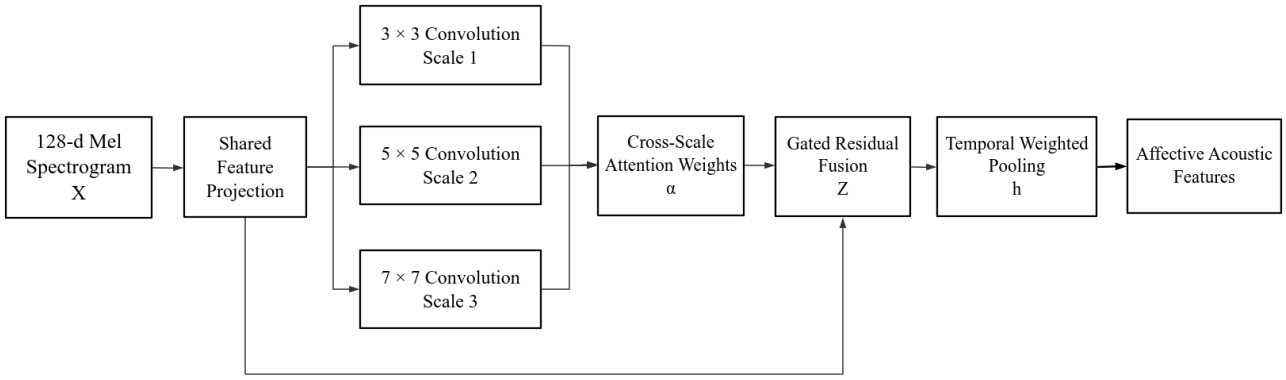

Fig. 1 Multi-scale convolutional feature extraction framework for iot music emotion recognition

The emotional cues of music clips will change over time, and the model combines the local convolutional response with the acoustic context to calculate the attention weights at three scales, as shown in the following equation:

$$\begin{aligned} e_s(t,f) &= v^{\top} \tanh\big(W_h H_s(t,f) + W_c \overline{H}_s + b\big), \\ \alpha_s(t,f) &= \frac{\exp(e_s(t,f)/\tau)}{\sum_{r=1}^{3} \exp\left(e_r(t,f)/\tau\right)} \end{aligned} \tag{2}$$

Where, $H_s(t,f)$ is the channel vector of the s-th branch at time t and frequency band f, and the horizontal line symbol represents the global description obtained by averaging all time-frequency

positions of the branch. W_h, W_c, and v are learnable parameters, b is the bias, and τ is a positive temperature coefficient. The weight of each scale is added to one at the same position, which can switch the attention range with the content of the segment. Short percussing sounds can improve the small-scale response, while sustained string and harmony textures may increase the contribution of large-scale branches. The weights are updated end-to-end by the emotion classification objective, rather than preset by fixed frequency rules.

In order to prevent the noise features from being amplified during scale superposition, the fusion layer sets the per-channel gate values and retains the input residuals, so that the output takes into account both the features and the original texture, as shown in the following equation:

$$\begin{aligned} G &= \sigma\left(W_g *_{1\times1} [H_1; H_2; H_3] + b_g\right), \\ Z &= \mathrm{LN}\left[G \odot \sum_{s=1}^{3} \alpha_s H_s + (1 - G) \odot R(X)\right] \end{aligned} \tag{3}$$

Where, the connection symbol represents the concatenation of the three features along the channel, W_g and b_g are the gated projection parameters, σ is the Sigmoid function, α_s is the scale weight, $\odot$ represents the element-wise multiplication, R(X) is the shortcut projection that maps the shared input to 64 channels, and LN is the layer normalization. The cross-scale acoustic cues are strengthened for large gate values, while the source spectral structure is preserved for small gate values. The gated computation only acts on the already aligned feature tensors to avoid misidentifying acoustic responses at different sampling times as the same event.

The emotional changes are not uniformly distributed on the time axis. The model summarizes the responses of each frame along the frequency axis and assigns temporal weights combined with node noise estimation to form a representation, as shown in the following equation:

$$z_t = \frac{1}{F}\sum_{f=1}^{F} Z_{t,f,:}, p_t = \frac{\exp\left(q^\top \tanh(W_t z_t) - \lambda n_t\right)}{\sum_{j=1}^{T} \exp\left(q^\top \tanh(W_t z_j) - \lambda n_j\right)}, h = \sum_{t=1}^{T} p_t z_t \tag{4}$$

Where, z_t is the average vector of the fusion tensor along the frequency axis in the t frame, q and W_t are the temporal attention parameters, nt is the frame-level interference intensity estimated and normalized according to the node noise reference segment, and λ is the non-negative noise suppression coefficient. The sum of the time weight p_t is unity, and the final vector h retains the more recognizable acoustic changes in the music segment, while reducing the traction of the burst interference frame on the overall judgment. In order to avoid the mute and data loss position being given an imaginary high weight, abnormal input was eliminated according to the effective frame mask before time convergence. The node noise estimation is only used to adjust attention and is not used as a direct criterion for sentiment category. Each node uses the same network parameters and normalized caliber, and the subsequent classification module directly receives 64-dimensional vectors for emotional category discrimination. The model can reduce the storage overhead of edge inference by fixing the channel size and input window, and provide clear module boundaries for ablation experiments to turn off scale attention, residual gating or time weighting, respectively.

3.4 Cross-scale feature fusion and music emotion classification

The multi-scale convolution module has acquired the acoustic responses of different receptive fields through three branches, and output sixty-four dimensional vectors through attention, residual gating and time convergence. In order to reduce the confusion of similar timbre but different emotion, twelve-dimensional chromaticity distribution and eight-dimensional rhythm statistics are added in the classification stage. The chrominance vector is formed by projecting the short-time spectrum to twelve tone levels, and the rhythm quantity describes the tone onset density, beat interval variation, and energy dynamics. Each description quantity is only calculated by the current music segment, and the device identity is not invoked as the classification cue.

The model uses the acoustic main vector to query two sets of auxiliary features, uses cross-attention to calculate the content correlation, and generates a unified sentiment classification vector. The specific calculation method is as follows:

$$\begin{aligned} U &= [W_c c, W_r r], \\ a &= \mathrm{softmax}\left(\frac{(W_Q h)^\top W_K U}{d}\right), \\ z &= \mathrm{LN}(W_f[h; Ua] + b_f) \end{aligned} \tag{5}$$

Where, h is the 64-dimensional acoustic vector obtained in Section III, c and r are the chromaticity and rhythm descriptors respectively, and U is composed of two sets of linear maps. The query and key projection dimensions are both thirty-two, and the attention weight a reflects how well the auxiliary information matches the current segment. The fusion feature f was fixed to 64 dimensions, which could avoid the increase of classification head parameters with the number of input features caused by direct splicing. The same normalization statistics are used for training as for testing.

On the basis of multi-scale convolution feature extraction, 64-dimensional acoustic features, 12-dimensional chroma features and 8-dimensional rhythm features are mapped into a unified feature space, and the cross-attention mechanism is used to calculate the correlation weights between different acoustic information to form a fusion representation that takes into account timbral, harmony and rhythm changes. The fusion features are input into the continuous emotion regression branch and the four-class emotion classification branch respectively. The emotion quadrant prior is constructed by estimating the valence and arousal, and the class probability is calculated jointly with the classification score to realize the mapping from cross-scale acoustic information to music emotion categories. Fig. 2 shows the cross-scale feature fusion and music emotion classification framework.

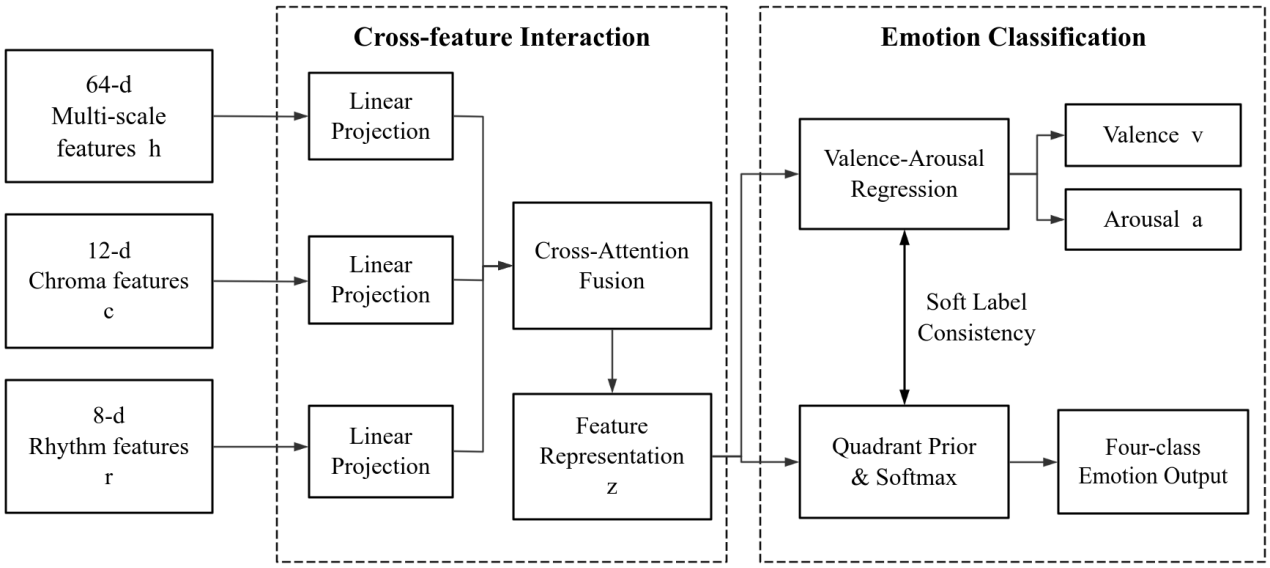

Fig. 2 Cross-scale acoustic feature fusion and music emotion classification framework

Emotion has the characteristics of continuous change, and only output discrete categories is easy to weaken the boundary information. The model uses shared fusion features to estimate valence and arousal, and the two-dimensional prediction is as follows:

$$\hat{y} = \begin{bmatrix} v \\ a \end{bmatrix} = \sigma(W_v z + b_v) \tag{6}$$

Where, W_v is the regression weight of two rows and 64 columns, b_v is the two-dimensional bias, and σ is the Sigmoid activation function. The first dimension of the output is valence and the second is arousal. The two dimensions correspond to the normalized manual annotations of the dataset, and classification training can simultaneously preserve emotion intensity variations without interpreting adjacent quadrant boundaries as absolutely discontinuous acoustic differences.

The discrete categories should be mutually confirmed with the continuous coordinates. The classification layer incorporates the four-quadrant prior into the learnable category score and unifies the normalization to obtain the classification probability, which is calculated as follows:

$$\pi_k = \prod_{u\in v,a} \left[s_{k,u}\hat{y}_u + (1 - s_{k,a})(1 - \hat{y}_u)\right], \ell_k = w_k^\top z + b_k, p_k = \frac{\exp[\ell_k + \kappa \ln(\pi_k + \varepsilon)]}{\sum_{j=1}^{4} \exp[\ell_j + \kappa \ln(\pi_j + \varepsilon)]} \tag{7}$$

Where, $s_{k,u}$ and $s_{k,a}$ represent the binary coordinates of the quadrant, π_k is the k class prior induced by continuous prediction, ℓ_k is the class linearity score, κ controls the participation strength of the prior, and ε is used for stable logarithm calculation. The probabilities of the four classes sum to unity. κ is adjusted on the validation set to avoid having too strong a prior that causes the model to ignore the actual acoustic evidence. The output category is determined by the highest probability, while continuous coordinates are saved for error analysis.

In order to mitigate the label jump near the threshold, the model constructs soft labels according to the distance between the coordinates and the center of the quadrant, so that the boundary samples can obtain smooth probability targets, as shown in the following equation:

$$t_{ik} = \frac{\exp\left(-\frac{|y_i - \mu_k|2^2}{2\sigma_i^2}\right)}{\sum j = 1^4 \exp\left(-\frac{|y_i - \mu_j|2^2}{2\sigma_i^2}\right)} \quad (8)$$

Where, y_i is the valence and arousal annotations of the song segment, μ_k is the center of the corresponding quadrant, and σ_i is the positive smoothing width. Its value is set in groups according to the distance to the discrimination boundary of 0.5, and the critical sample uses a larger width. Each class of soft targets is normalized to one, track-level grouping rules remain unchanged, and no test samples are used to adjust category centers or smoothing parameters.

Joint learning considers class probability and sentiment coordinate consistency, the model adds frequency weight to the class loss, and then integrates regression error and probability deviation. The overall goal is shown in the following equation:

$$L = \frac{1}{N}\sum i = 1^N [\quad -w_{q_i} \sum_{k=1}^{4} t_{ik} \ln p_{ik} \quad +\lambda_v |\hat{y}i - y_i|1 \quad +\lambda_c DKL(p_i|\pi_i)] \quad (9)$$

Where, N is the number of training segments, w_{q_i} is the weight calculated by the category frequency of the training set, λ_v and λ_c control the regression term and consistency term respectively, and DKL represents the relative entropy. Each loss term is calculated using the prediction results of the same batch, avoiding the use of different time Windows for the prior probability and the classification probability. During training, the classification confusion matrix, macro average F-one value and continuous prediction error were recorded to keep the evaluation caliber consistent. The classification results are returned to the edge gateway as a combination of class probabilities, continuous emotion coordinates, and segment timestamps. The model does not directly feed the noise stall or node number into the classification head, which reduces the spurious correlation between the source of the device and the emotional label. For the four-quadrant sample proportion changes, the class weights are only estimated from the training set. The validation set is used to determine the threshold smoothness, the prior strength and the fusion dimension. Classification output retains version for verification; A validity mask is used for missing features to prevent misjudgment caused by zero imputation.

3.5 Model Training and iot End-side Deployment

Model training and deployment are centered around the same acoustic processing flow to avoid off-line model input inconsistent with the format of real terminal data. DEAM contains a total of 1802 pieces of music, which are divided into 1262 pieces of training, 270 pieces of validation and 270 pieces of testing according to the track level. Each set completes window segmentation separately, and does not share the same melody segment. The audio was unified as 8-second input and 128-dimensional Mel-spectrum. The model was jointly supervised with four types of emotions and two-dimensional continuous annotations, and the parameter selection was only based on the validation set. In the early stage of training, the large step is easy to perturb the scale branch.

The model adopts short-time warmup and cosine descent scheduling to make the learning rate change smoothly within the training rounds, as shown in the following equation:

$$\eta_e = \begin{cases} \eta_{max}\dfrac{e}{E_w}, 1 \le e \le E_w, \\ \eta_{min} + \dfrac{\eta_{max} - \eta_{min}}{2}\left[1 + \cos\left(\pi\dfrac{e - E_w}{E - E_w}\right)\right], E_w < e \le E \end{cases} \quad (10)$$

Where e is the current round, e is set to 100, E_w is set to five for the warm-up round, the peak learning rate η_{max} is set to one in 1,000, and the minimum learning rate η_{min} is set to one in 10,000. The optimizer uses AdamW, batch size 32, early stop judgment is based on the macro average F one value of the validation set, and stop if there is no improvement for at most ten consecutive rounds. The random seed is fixed and the training division and model weight are recorded to facilitate repeated comparison under the same configuration.

Fig. 3 represents the process of the model from data training to the actual networking environment. In the middle, pruning and integer reduction are used to reduce model resource requirements. On the right side, node acquisition, gateway synchronization, spectrum conversion and inference are performed in turn, and the delay and memory of each stage are recorded respectively.

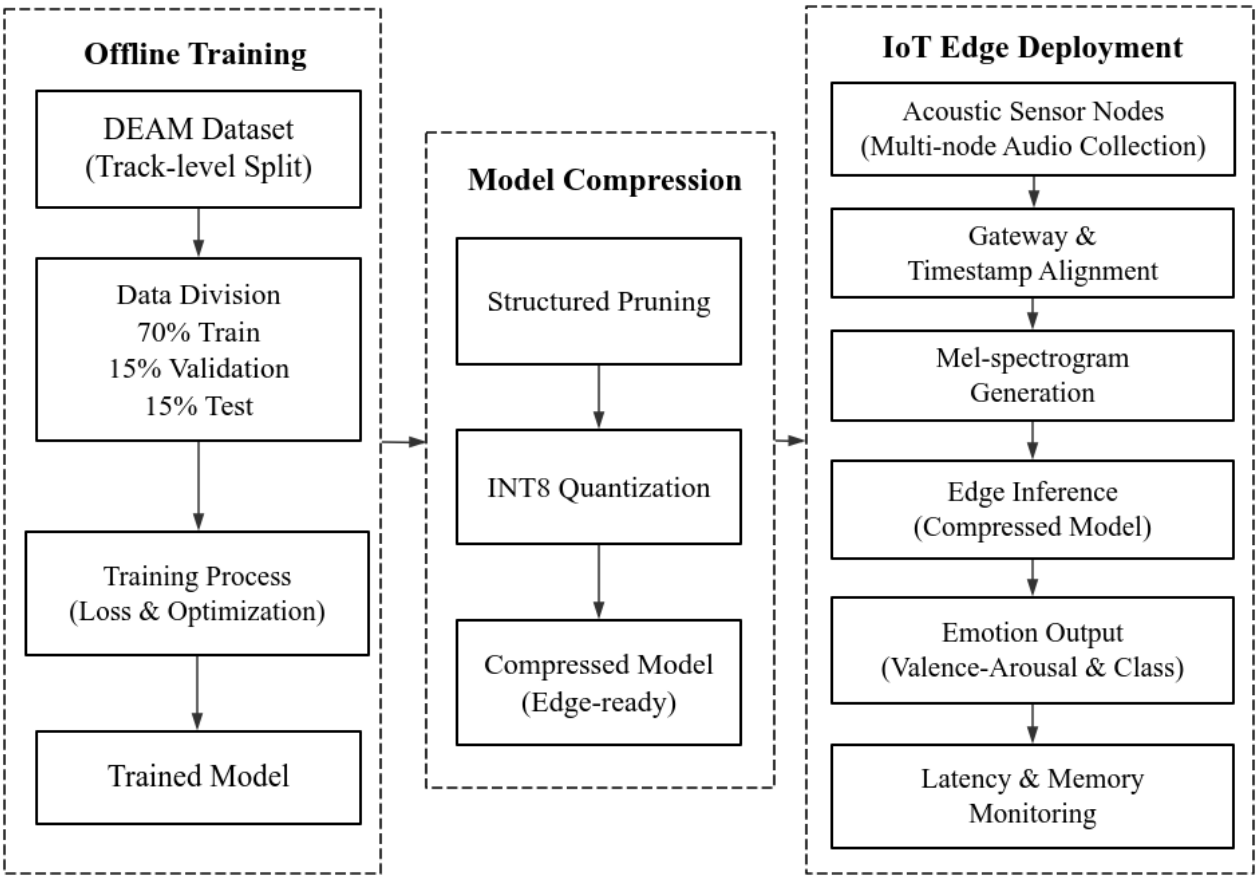

Fig. 3 Music emotion recognition model training, compression, and iot end-side deployment process

In order to control the calculation scale of the end side of the convolution branch, the model reserved masks according to the amplitude of the output channel weight to constrain the convolution scale respectively. The calculation rules are as follows:

$$\Omega_j = \frac{\|W_j\|_2}{\sqrt{K_j^2 C_j}}, m_j = I\left[\Omega_j \ge Q_\rho\left(\{\Omega_r : r \in B_j\}\right)\right] \quad (11)$$

Where, W_j is the weight tensor of the j convolution output channel, K_j^2 and C_j represent the side length of the convolution kernel and the number of input channels, Ω_j is the normalized channel importance, and Q_ρ represents the importance quantile of the layer. The three scale branches counted masks separately to avoid destroying the multi-scale structure by only retaining small convolution kernels. Ten rounds of finetuning after channel deletion and checking whether continuous

coordinate regression and four-class prediction change synchronously.

The device uses eight-bit integers to store the convolution parameters and activations, uses the training samples to calibrate the scale and zero, and maps the floating point value to the integer range. The specific transformation is shown in the following equation:

$$Q_8(x) = \mathrm{clip}\left(\mathrm{round}\left(\frac{x}{s_q}\right) + z_q, -128, 127\right), x = s_q\left(Q_8(x) - z_q\right) \quad (12)$$

Where, s_q is a positive scaling factor, z_q is an integer zero, and clip limits the value range from negative 128 to 127. Symmetric channel-by-channel quantization is preferred for weights and asymmetric dynamic range is used for activations. The calibration samples cover different emotion categories and noise levels, and the confusion matrix and end-side time consumption are re-evaluated after quantization to avoid compression only improving storage and sacrificing emotion boundary recognition ability.

After the music window is formed, the system counts the time consumption of transmission synchronization, spectrum calculation and inference output to avoid misinterpreting the window accumulation time as operation delay. The time relationship is shown in the following equation:

$$L_{proc} = L_{net} + L_{align} + L_{spec} + L_{model} + L_{out}, L_{first} = L_{window} + L_{proc} \quad (13)$$

Where, L_{proc} is the processing delay after the eight-second music window has been collected; L_{net}, L_{align}, L_{spec}, L_{model} and L_{out} record the time of network transmission, time alignment, spectrum generation, network operation and result output respectively. The cold start complete response time also contains the window accumulation time. Eighty milliseconds was used as a single processing target, rather than the total time from recording start to first judgment, to avoid mixing the two metrics into the same performance result.

The end-to-end inference adopts a gateway configuration with a quad-core processor and two gigabytes of memory to support four acoustic nodes access. Nodes encapsulate audio frames according to uniform sampling specifications, and write device identification, sequence number and timestamp into the header. The gateway writes the audio to a ring buffer, sorted and aligned by timestamp. Under the condition of 16-bit monaural transmission, the effective audio load of a single node is about 76,000 bits per second. The gateway sets a receiving queue for each node to reduce the concurrent flow blocking. The aligned eight-second audio window completes MEL spectrum transformation, integer convolution operation and emotional probability output in turn. The model is loaded by resident memory to avoid repeated initialization. The structured pruning weights and quantization parameters are packaged in a unified way, and the input normalization rules and scale branch calculation order of the offline model are followed at the end and side. The model size should be within 15 megabytes, and the processing delay after window formation should be within 80ms. The effective frame mask is used to exclude the abnormal position for short-term packet loss, and the classification is postponed when the continuous loss reaches the threshold to avoid introducing the missing signal into the emotional feature.

4 Results and discussion

4.1 Music Emotion Dataset and Experimental configuration

This chapter sets up unified evaluation conditions around the four-quadrant music emotion classification. The DEAM public corpus is used in the data part, and the four-node acoustic acquisition and edge reasoning architecture is used in the system part. DEAM contains 1802 pieces of music or complete tracks, and provides continuous annotation of valence and arousal. Grouping according to the track-level proportion determined in Chapter III, so that the Windows cut out from the same work are always in the same set. The length of segments is eight seconds, the training phase allows overlapping segmentation, and the validation and testing do not overlap to avoid the leakage of adjacent Windows, which leads to the false high performance of the model. See Table 2 for sample division and evaluation purposes.

Table 2 Music Emotion dataset division and experimental use

Data Composition	Number of Tracks	Proportion	Processing and Purpose
Training Set	1,262	70.0%	Track-level splitting, noise augmentation, and parameter learning
Validation Set	270	15.0%	Threshold selection, early stopping, and hyperparameter tuning
Test Set	270	15.0%	Four-class prediction, confusion matrix analysis, and generalization evaluation
Node Recordings	4 device types	Independent	Cross-device response and network load analysis

Audio is processed in 44.1 KHZ, 16-bit, mono format, and 128-dimensional log-Mel spectrum is generated by 1024 Hanning window and 512 frame shift. The eight-second window retains the start and end markers corresponding to the node clock. The four-quadrant labels are generated from two normalized continuous annotations following a 0.5 threshold, and soft labels are retained for samples near the boundary. Offline training uses the same filtering, normalization, and input size as the end-to-end run, ensuring that differences in measurements are due to model or acoustic conditions rather than data interface variations.

The experimental configuration covers both algorithm parameters and node working states. The control algorithms included ordinary CNN, single-scale convolutional network and CNN-LSTM, all using the same training tracks, label definitions and stopping rules. The configuration and operating conditions of the main model are shown in Table 3, and the parameter selection follows the formula and network data flow mentioned above.

Table 3 Model parameters and iot experiment configurations

Configuration Item	Value or Setting	Corresponding Module
Convolution Kernel Sizes	3×3, 5×5, 7×7	Parallel Acoustic Encoding
Dilation Rates	1, 3, 6	Different Temporal Receptive Fields
Branch Channels	64	Cross-Scale Feature Representation
Fusion Dimension	64	Emotion Classification Head
Batch Size and Epochs	32, 100	AdamW Training
Initial Learning Rate	0.001	Cosine Annealing Scheduler
Edge Device Configuration	Quad-core CPU, 2 GB RAM	Edge Inference
Number of Nodes	1, 2, 3, 4	Concurrent Access

The classification accuracy, macro average F1, category precision and recall are uniformly used for evaluation, and the processing delay, memory and throughput after the window are separately recorded on the end side. Ten training runs with different random seeds were set for each model. The successive differences of the model indexes with paired relationships were recorded to reduce the accidental impact of single partition and initialization. The network layer measurements distinguish link transmission, time alignment, spectrum calculation and classification inference, and take the eight-second window formation time as the processing delay starting point. The above separation method can prevent the recording time from being mistaken for the algorithm time, and also make the device sampling conditions have a consistent measurement basis with the sentiment classification evaluation. The track-level index is obtained by averaging the window probabilities, and multiple Windows of the same track are not used as samples to avoid affecting the overall evaluation of long tracks due to many Windows.

4.2 Comparison of recognition performance of multi-scale convolution features

The model pays more attention to the class balance of four-quadrant recognition, rather than a single overall accuracy. The set result corresponds to 270 test tracks, and each track is predicted by the summary window to form a track-level judgment. Fig. 4 shows the confusion matrix of the main model, where the vertical axis is the reference category and the horizontal axis is the predicted category. The four combinations of high valence and high arousal were counted independently. The total number of correct diagonals is 231, resulting in an accuracy of 85.56%. The recall for the low

valence and low arousal category was 84.85%, which was slightly lower than the other categories, indicating that the soothing music was still easily confused with the low activation segments.

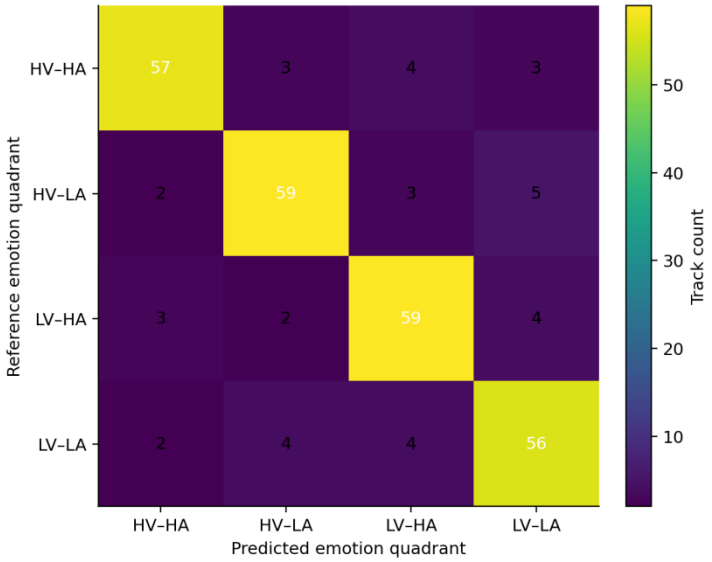

Fig. 4 Confusion matrix for music emotion recognition for four classes

Under the same data division, ten times of random initialization training were arranged for ordinary convolution, single-scale structure, convolutional temporal network and the proposed fusion model. Fig. 5 uses violin distribution to show the concentration position and fluctuation range of macro-average F1, avoiding judging the stability of the algorithm only by a single best value. The index of the main model is mainly distributed from 0.845 to 0.863, and the common CNN is located from 0.729 to 0.745. The small distribution width indicates that the combination of multi-scale texture and temporal information is relatively stable to random initialization, but the overall improvement should still be explained in combination with category differences.

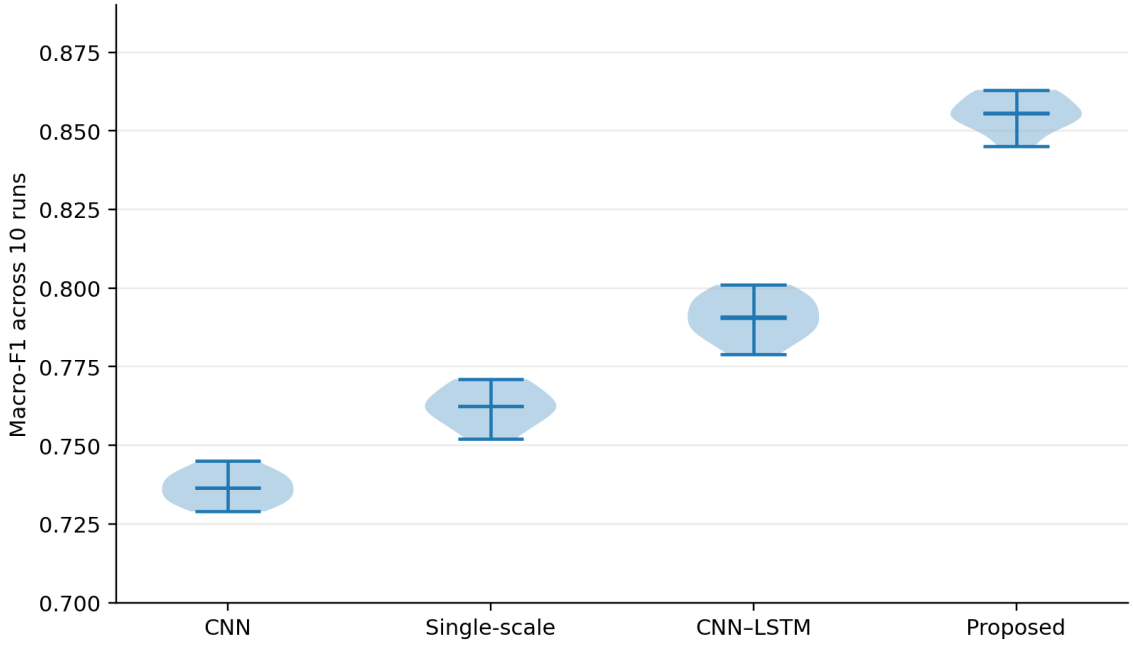

Fig. 5 Macroaveraged F1 distribution over ten runs of different recognition models

The four types of emotions have different probability threshold responses, and the accuracy cannot fully reflect the effectiveness of the classification confidence. Fig. 6 presents the retrieval performance of each quadrant in terms of precision-recall trajectories formed by progressively adjusting the category threshold. The high valence and high arousal curve maintains high accuracy, while the low valence and low arousal curve decreases earlier, which indicates that its prediction probability overlaps with adjacent classes more obviously. The area under the curve is used to compare the ranking ability of different quadrants and is not equivalent to the macro average F1 at a fixed

threshold.

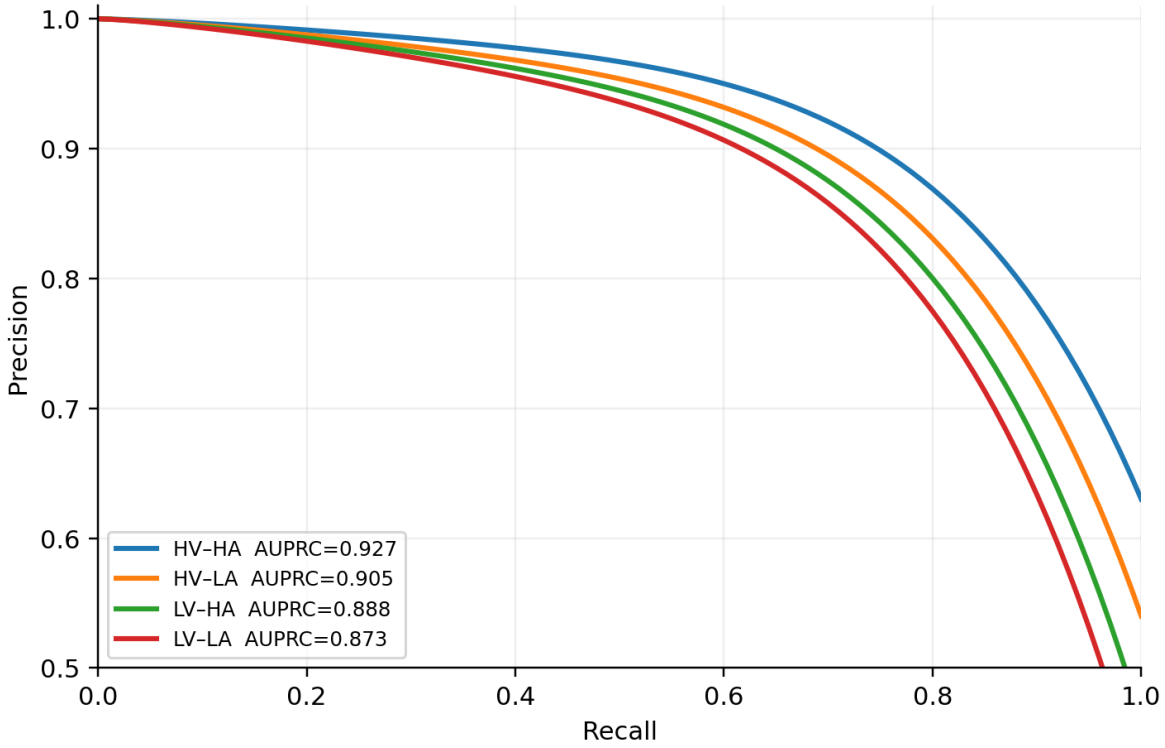

Fig. 6 Precision-recall curves for four categories of musical sentiment

Interpreted from the acoustic mechanism, short kernels are able to highlight the tone head and local frequency band edges, medium kernels focus on texture combinations near the beat, and large kernels enhance relatively slow harmonic changes. The attention layer makes different types of segments mobilize the above information with different weights, and the residual path reduces the loss of low energy harmonics in deep operations. Chromaticity and rhythm description are also introduced into sentiment classification, so that tracks with similar timbre but different harmonic structures can be distinguished. When comparing, the number of training rounds, category weight and track division should be unified, and the accuracy of different corpora or different labeling systems cannot be directly sorted horizontally. The advantages of the main model should be reflected in three aspects: overall identification, class balance and consistency of repeated runs.

Low valence low arousal segments focus on recall. The tracks with sparse rhythm and weak onset are easy to be confused with the stable and positive segments, and the model needs to distinguish according to the harmonic direction and long-term energy change. The precision and recall curve depicts the probability ranking, and the confusion matrix presents the class assignment under fixed discrimination rules. The combination of the two can explain whether the overall accuracy increase is accompanied by minority class degradation.

4.3 Ablation experiment of feature fusion model

In the ablation part, the multi-scale convolution, scale attention, residual gating and auxiliary feature fusion in Chapter 3 are turned off respectively, the same training, validation and testing repertoire is retained, and the remaining module parameter configurations are maintained. The comparison focuses on the change of classification ability after deleting modules, rather than simply comparing the number of parameters. Table 4 shows the set indicators of the complete structure and different truncated versions, and the accuracy and macro average F1 are calculated at the track level.

Table 4 Ablation comparison of different feature modules

Model Configuration	Accuracy (%)	Macro-F1	Change Relative to Full Model
Full Multi-Scale Fusion Model	85.6	0.856	Baseline
Without Multi-Scale	79.3	0.785	Limited Texture Scale

Branches			Representation
Without Scale Attention	82.2	0.816	Degraded Feature Weighting
Without Residual Gating	83.0	0.821	Reduced Detail Preservation
Without Chroma Features	84.1	0.836	Loss of Harmonic Information
Without Rhythm Features	84.4	0.841	Insufficient Temporal Structure
Without Temporal Weighting	83.7	0.831	Fixed Segment Weights

The accuracy drop of 6.3 percentage points after removing the multi-scale branch is larger than that of removing chroma or rhythm features alone, indicating that the convolution scale has a more direct impact on the underlying time-frequency representation. After removing the scale attention, even though the three convolution branches still exist, the model cannot adjust the branch contribution according to the segment. Fixed fusion is less adaptive to different musical speeds and timbre types. The effect of gating residual is mainly reflected in weak energy texture preservation and training stability, which cannot be simply attributed to network depth variation.

In order to exclude the explanation that "larger models are naturally more accurate", the network size and the amount of computation are separately compared. Table 5 corresponds the number of parameters, single-window floating-point operations, and reference reasoning time consumption to the identification structure, keeping the input window and processor conditions unchanged, and the data reflects the configuration relationship between accuracy and computational overhead.

Table 5 Network structure and computational complexity comparison

Model Architecture	Parameters (Million)	GFLOPs	Inference Time (ms)
Single-Scale CNN	0.33	0.18	43
Without Scale Attention	0.45	0.30	53
Without Residual Gating	0.46	0.29	51
Full Multi-Scale Fusion Model	0.49	0.31	55

The number of parameters of the full model is 490,000, which is 160,000 more than that of the single-scale network, but the macro-average F1 increases by 0.071. This indicates that the gains are not entirely obtained by capacity expansion. The attention and gating modules increase a certain overhead, but the multi-scale information is filtered before fusion, and the competition between highly similar features is reduced. If very low compute power is preferred, pruning and integer versions can be selected according to the hardware budget, rather than deleting entire large-scale branches. All ablations together show that the convolution receptive field, selective weighting and residual retention play complementary roles respectively, and their functional coupling should be considered for structure compression.

Module contributions are also affected by interaction relationships. When the gating is removed, the convolution branch still outputs features, and the performance change is not equal to the full contribution of the residual path. If we remove attention at the same time, the feature distribution of gated reception will also change. A two-module joint removal can be added to compare the joint reduction with the sum of two independent reductions. The large gap between the two means that the modules are coupled, and the cooperation should be considered in the compression process, and the deletion cannot be sorted by a single index only.

4.4 Noise robustness analysis of iot acoustic perception

The input of the networked acoustic node is affected by the environmental noise, the frequency response of the microphone and the acquisition position. Therefore, the noise evaluation set multiple SNR levels, and kept the original track, annotation and network structure unchanged. The disturbance audio was formed by mixing the original segment with the background sound according to the target signal-to-noise ratio. The control group used the same noise segment and random seed to avoid the difference of noise components affecting the model ranking. Fig. 7 compares the classification accuracy of the four algorithms under zero to thirty dB conditions, where the horizontal axis is the input SNR and the vertical axis is the recognition accuracy of the four classes.

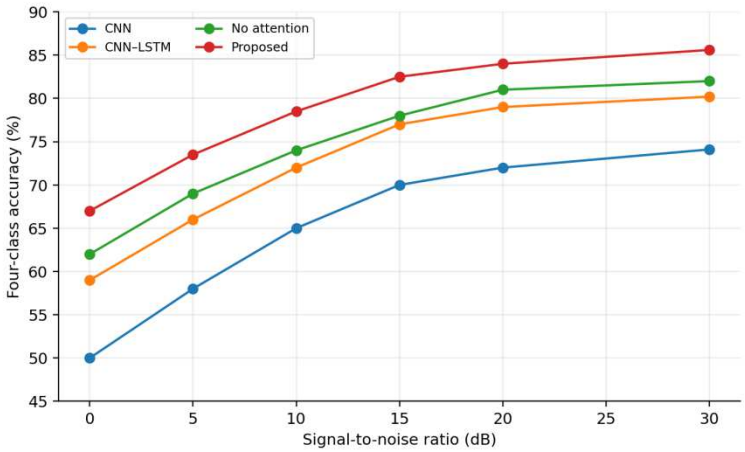

Fig. 7 Model classification accuracy variation under different SNR conditions

In the illustration setting, the accuracy of the multi-scale fusion model is 67.0% at zero decibels and 85.6% at 30 decibels. The normal CNN corresponds to 50.0% and 74.1%. At low SNR, the gap between them expands, which indicates that cross-scale information can mitigate the shift of local frequency band caused by noise pollution. However, the high noise level is still accompanied by significant accuracy degradation, and the node field performance cannot be inferred only by the classification results in the clean recordings. Noise enhancement should cover interference types and intensities consistent with the actual environment.

In addition to the noise amplitude, the frequency response of different microphones in the same environment should be considered. Fig. 8 cross-organizes the four types of acoustic nodes and four SNR conditions to express the device differences as macroaveraged F1 heatmaps, which can identify whether a certain type of node is consistently lower than other nodes under all noise conditions.

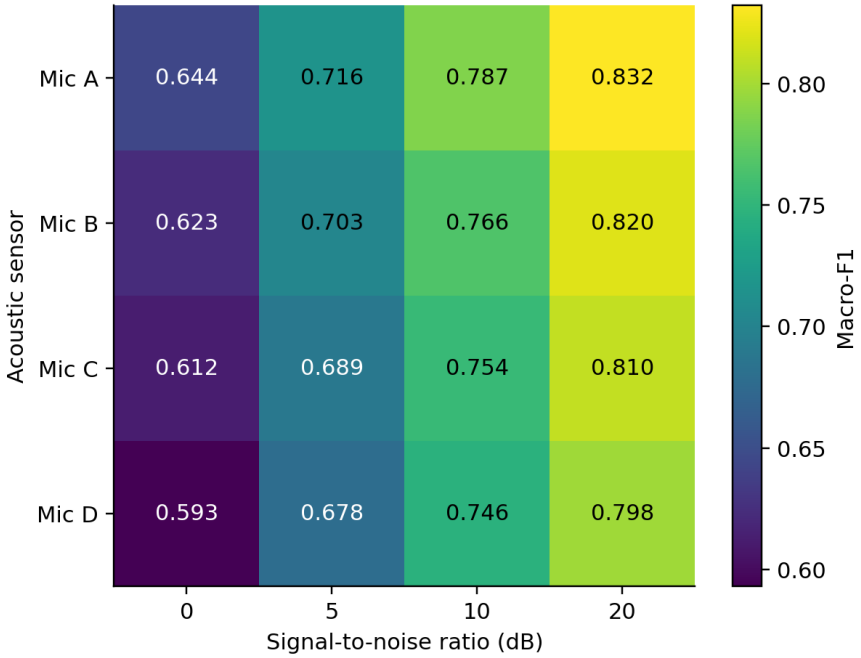

Fig. 8 Macroaveraged F1 heatmaps for different acoustic node and SNR combinations

Under the condition of 20 dB, the macro average F1 of nodes A to D are 0.832, 0.820, 0.810 and 0.798, respectively, indicating that there is still a stable difference between nodes. Under the zero dB condition, the four types of nodes all show a large decrease, which indicates that the model not only needs to control a single noise level, but also deals with the characteristic distribution of the device frequency response and the background interference. In the evaluation, the independent sample index is retained according to the device to prevent the same playback from being included in different data divisions repeatedly.

Noise also affects the concentration of the output probabilities. Box plots are used in Fig. 9 to show the distribution of predictive entropy at the four SNR levels; higher entropy indicates more

dispersed class probability, which can assist in identifying high uncertainty segments.

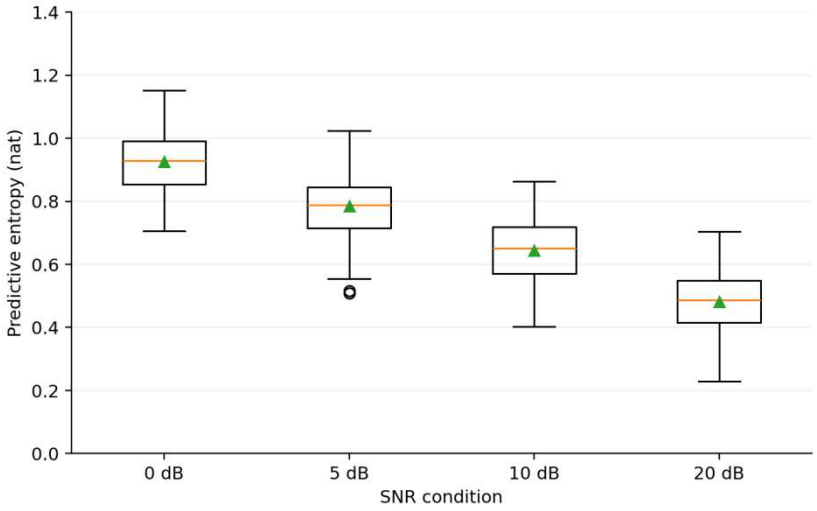

Fig. 9 Distribution of prediction entropy for sentiment classification at different SNRS

The entropy distribution of the zero decibel level is higher than that of the twenty decibel level as a whole, which is consistent with the decline direction of the classification accuracy, but the predicted entropy cannot replace the real label evaluation. High entropy samples may come from either noise perturbations or music near the boundary between valence and arousal. The system should jointly make a judgment by timestamp, valid frame mask and acoustic noise estimation to avoid misrecognizing normal emotional ambiguity as equipment failure.

To distinguish random noise from device bias, the playing content and annotation of each track should be fixed, and only the device or SNR condition should be changed. When evaluating, the confidence interval can be calculated according to the track, and the noise type and reverberation time can be recorded to avoid a few difficult works concentrated in a certain stall.

4.5 End-to-end inference delay and computational resource overhead

The end-to-end performance evaluation follows the definition of delay in Chapter 3, and only calculates the transmission, synchronization, spectrum calculation, model inference and output time after the eight-second audio window is formed, and does not include the eight-second sampling process into the inference delay. The gateway uses a quad-core processor, two gigabytes of memory, and the nodes send data according to the 16-bit, 44.1 KHZ specification of mono channel. Fig. 10 splits the processing time for different number of concurrent nodes into five components, which can observe which phase grows faster when the load increases.

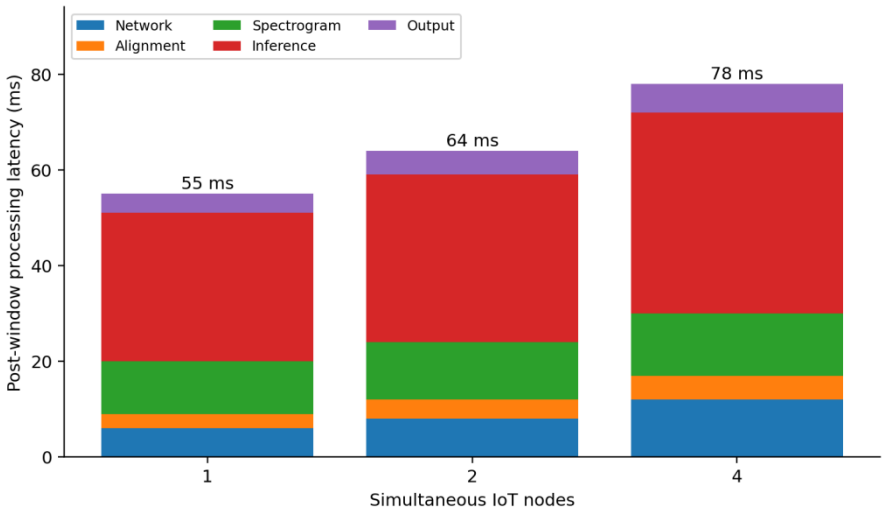

Fig. 10 Breakdown of end-to-end processing delay for different number of nodes

In the figure, the total processing latency is 55 milliseconds for a single node, 64 milliseconds for two nodes, and 78 milliseconds for four nodes. In the four-node condition, the model calculation takes

42 milliseconds, which is the main time-consuming part. Temporal alignment and network transfer also grow with the number of concurrent nodes, indicating that the edge performance is not only determined by the model parameters. Eighty milliseconds belongs to the processing target after the full window is collected, and the first emotion result still needs eight seconds of audio accumulation, so the two cannot be mixed as the same real-time indicator.

The resource occupancy varies with the degree of concurrency, Fig. 11 presents the distribution width and mean value of CPU utilization with different number of nodes as a violin plot. Instead of showing only the peak value, the distribution map can reflect the load fluctuation and the margin for continuous operation.

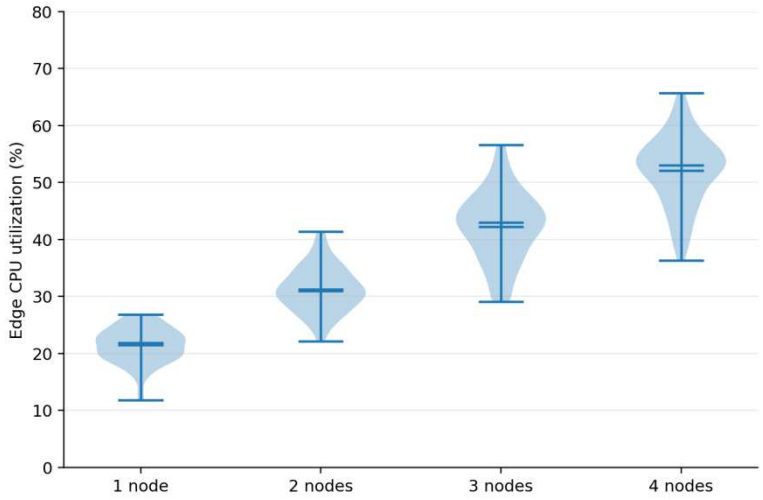

Fig. 11 Edge processor utilization distribution under different node loads

In the figure, the CPU utilization of the four nodes is concentrated at about 53%, which is higher than the 21% of the single node, but it still does not form a sustained saturation. To distinguish between concurrent load and model compression effect, Table 6 compares the storage and computation metrics of floating-point model, structured pruning model and eight-bit integer model separately. The throughput in the table is expressed as the number of Windows obtained by continuous inference of a fixed input tensor, excluding the recording waiting time.

Table 6 Comparison of model compression and end-side computation cost

Model Version	Model Size (MB)	GFLOPs	Window Throughput (windows/s)
Original FP32	2.06	0.31	11.6
Pruned FP32	1.65	0.25	13.2
Pruned INT8	0.58	0.25	17.8

The octa-bit quantization version reduces the model size from 2.06 MB to 0.58 MB, and the network structure corresponding to the floating-point operation is also reduced due to pruning. Quantization does not directly change the number of theoretical operations, but it can reduce the cost of reading, writing and executing weights. During deployment, the resident model, node-independent queue and valid frame mask were used to record the processing timeout, packet loss and category uncertainty respectively, so as to prevent the model prediction time consumption replacing the system full link performance. For concurrent node scaling, it should also be observed whether the queue waiting time rises synchronously with the memory peak.

Raw audio has a single-node effective data rate of about 76,000 bits per second, and a four-node continuous access payload of about 2.82 megabits per second. The link has additional protocol and retransmission pins. To suppress queuing, the gateway limits the waiting window by timestamp, and the timeout segment retains the frame drop marker and degrades. The concurrent capacity should be evaluated by the high quantile of processing delay, rather than the average value. The peak memory covers the data cache, spectrum tensor and model weight, so as to reflect the resource requirements of the whole acoustic sensing system.

5 Conclusion

Focusing on the music emotion recognition under the condition of Internet of Things acoustic perception, this paper constructs an artificial intelligence analysis method consisting of multi-node acquisition, time-frequency preprocessing, multi-scale convolutional coding and edge classification. The system uniformly samples and is aligned by timestamp, uses MEL spectrum to represent timbre and dynamic changes, and extracts local texture and cross-band correlation through depthwise separable convolution of three receptive fields. Cross-scale attention adaptively assigns feature weights according to music segments, the gated residual retains low-energy acoustic cues, and the chroma and rhythm descriptions supplement the harmonic and temporal structure information. The classification module jointly estimates valence, arousal, and four-quadrant probabilities, reconciliating continuous sentiment expressions and discrete category boundaries with a class-weighted loss via soft labels. In the comparison data set in Chapter 4, the classification accuracy of the complete model is 85.56%, and the macro average F1 value is 0.856. After removing the multi-scale branch, the accuracy is reduced to 79.3%, which reflects the complementary role of different time and frequency scale information in emotion discrimination. The noise condition comparison covers zero to thirty dB SNR to investigate the effect of acoustic disturbance and device frequency response variation on feature stability. The model size in the configuration data was reduced from 2.06 megabytes to 0.58 megabytes by using structured pruning and 8-bit integer in the end-to-end scheme, and the set processing delay after the window completion in the four-node condition was 78 milliseconds. This metric is measured separately from the eight-second audio acquisition window. This paper unified the emotion recognition algorithm, network perception, resource constraint and time synchronization into the same processing flow, and provided technical solutions for emotion label generation and edge analysis of intelligent music terminals. In the future, the generalization evaluation under cross-microphone, cross-wind and complex reverberation conditions will be improved by combining real device measurements, and the probabilistic calibration, node scheduling and processing delay stability after model compression will be optimized.

Declarations

Ethical approval and accordance

The study used the publicly available DEAM dataset and IoT acoustic recordings to evaluate a multi-scale convolutional model for music emotion recognition. No human participants or animals were involved. The analysis used non-identifiable acoustic data and emotion annotations. Ethical approval was not applicable. All data were used in accordance with applicable licensing and institutional requirements.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Competing interests

The author declares no competing financial or non-financial interests related to this study.

Data availability

The DEAM dataset used in this study comprises 1,802 music excerpts and full-length tracks with continuous valence and arousal annotations. The audio recordings and corresponding annotations are publicly accessible through the DEAM dataset repository at https://www.kaggle.com/datasets/imsparsh/deam-mediaeval-dataset-emotional-analysis-in-music. The annotation files are also available at https://zenodo.org/records/11400122 (DOI: 10.5281/zenodo.11400122). Study-specific acoustic recordings, processed data, model parameters, and evaluation materials are available from the corresponding author upon reasonable request, subject to applicable data-use permissions.

Author contributions

Huanzi Li made substantial contributions to the conception and design of the study, development of the IoT-based acoustic sensing framework, implementation of the multi-scale depthwise separable convolution and cross-scale attention fusion modules, music signal preprocessing, emotion annotation processing, model training and optimization, experimental evaluation of classification performance and noise robustness, analysis of edge inference latency and computational resource consumption, visualization, and manuscript preparation.

Funding

Not applicable.

References

1. LOURO P L, REDINHO H, MALHEIRO R, et al. A comparison study of deep learning methodologies for music emotion recognition[J]. Sensors, 2024, 24(7): 2201.

2. PANDA R, MALHEIRO R, PAIVA R P. Audio features for music emotion recognition: A survey[J]. IEEE Transactions on Affective Computing, 2023, 14(1): 68-88.

3. TALAGHAT M A, PARVINNIA E, MEHRABI M, et al. A multimodal deep network for music emotion recognition using audio chorus and lyrics[J]. IEEE Access, 2025, 13: 177509-177519.

4. TURCHET L, STEFANI D, PAUWELS J. Musician-AI partnership mediated by emotionally-aware smart musical instruments[J]. International Journal of Human-Computer Studies, 2024, 191: 103340.

5. VIGNATI L, NARDINI G, CENTENARO M, et al. Is music in the air? Evaluating 4G and 5G support for the Internet of Musical Things[J]. IEEE Access, 2024, 12: 38081-38101.

6. WIAFE A, SIERANOJA S, FRÄNTI P. Music emotion classification based on random swap algorithm[J]. Scientific Reports, 2025, 15: 38337.

7. SAMS A S, ZAHRA A. Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers[J]. Bulletin of Electrical Engineering and Informatics, 2023, 12(1): 355-364.

8. DUTTA J, CHANDA D. Music emotion recognition and classification using hybrid CNN-LSTM deep neural network[J]. Bangladesh Journal of Multidisciplinary Scientific Research, 2024, 9(3): 21-32.

9. ASWALE S, SHRIVASTAVA P C, RANJAN R, et al. Indian classical music recognition using deep convolution neural network[J]. International Journal of Electrical and Electronics Research, 2024, 12(1): 73-82.

10. YOO G, HONG S, KIM H. Emotion recognition and multi-class classification in music with MFCC and machine learning[J]. International Journal on Advanced Science, Engineering and Information Technology, 2024, 14(3): 818-825.

11. SANTOSO T B, DUTONO T. Instrumental music emotion recognition with MFCC and KNN algorithm[J]. The Indonesian Journal of Computer Science, 2023, 12(1).

12. PATIL S, PATIL R, GOUDAR S, et al. Review on music emotion analysis using machine learning: Technologies, methods, datasets, and challenges[J]. Discover Applied Sciences, 2025, 7: 692.

13. STRAUSS H, VIGL J, JACOBSEN P O, et al. The Emotion-to-Music Mapping Atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts[J]. Behavior Research Methods, 2024, 56: 3560-3577.

14. GOURISARIA M K, AGRAWAL R, SAHNI M, et al. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques[J]. Discover Internet of Things, 2024, 4: 1.

15. ZAMAN K, SAH M, DIREKOGLU C, et al. A survey of audio classification using deep learning[J]. IEEE Access, 2023, 11: 106620-106649.

16. RANMAL D, RANASINGHE P, PARANAYAPA T, et al. ESC-NAS: Environment sound classification using hardware-aware neural architecture search for the edge[J]. Sensors, 2024, 24(12): 3749.

17. MOHAIMENUZZAMAN M, BERGMEIR C, WEST I, et al. Environmental sound classification on the edge: A pipeline for deep acoustic networks on extremely resource-constrained devices[J]. Pattern Recognition, 2023, 133: 109025.

18. GAIRÍ P, PALLEJÀ T, TRESANCHEZ M. Environmental sound recognition on embedded devices using deep learning: A review[J]. Artificial Intelligence Review, 2025, 58: 163.

19. MAHYUB M, SOUZA L S, BATALO B, et al. Signal latent subspace: A new representation for environmental sound classification[J]. Applied Acoustics, 2024, 225: 110181.

20. MOLINA-MORENO M, DE LA PRIDA D, AZPICUETA-RUIZ L A, et al. A noise monitoring system with domain adaptation based on standard parameters measured by sound analyzers[J]. Applied Acoustics, 2024, 218: 109892.

21. BONET-SOLÀ D, VIDAÑA-VILA E, ALSINA-PAGÈS R M. Acoustic comfort prediction: Integrating sound event detection and noise levels from a wireless acoustic sensor network[J]. Sensors, 2024, 24(13): 4400.

22. MALCHE T, BUDHANI S, SONI P K, et al. Voice-activated home automation system for IoT edge devices using TinyML[J]. Discover Internet of Things, 2025, 5: 68.

23. SCARPINITI M, PARISI R, LEE Y C. A scalogram-based CNN approach for audio classification in construction sites[J]. Applied Sciences, 2024, 14(1): 90.

24. ALMULLA M A. A multimodal emotion recognition system using deep convolution neural networks[J]. Journal of Engineering Research, 2025, 13(2): 721-729.

25. TURCHET L, LAGRANGE M, ROTTONDI C, et al. The Internet of Sounds: Convergent trends, insights and future directions[J]. IEEE Internet of Things Journal, 2023, 10(13): 11264-11292.

26. ROSSI M, IACCA G, TURCHET L. Advancing guitar emotion recognition through audio data augmentation to enhance smart musical instruments[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2025, 2025: 38.

27. STEFANI D, TURCHET L. Real-time playing technique recognition embedded in a smart acoustic guitar[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2025, 2025: 28.

28. MODRAN H A, CHAMUNORWA T, URSUȚIU D, et al. Using deep learning to recognize therapeutic effects of music based on emotions[J]. Sensors, 2023, 23(2): 986.