INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Arts and Technology

ISSN online: 1754-8861 - ISSN print: 1754-8853
https://www.inderscience.com/ijart

Music intelligent creation method based on LSTM and multi-scale attention

Liping Li

DOI: 10.1504/IJART.2025.10074831

Article History:

Received:	29 August 2025
Last revised:	17 October 2025
Accepted:	17 October 2025
Published online:	02 December 2025

Music intelligent creation method based on LSTM and multi-scale attention

Liping Li

Conservatory of Music,
Jilin Normal University,
Siping, 136000, China
Email: lipingli_lp@163.com

Abstract: To address challenges in current AI music composition such as inconsistent styles, high melody repetition rates, and weak emotional expression, this study innovatively proposes a music intelligent creation method integrating long short-term memory (LSTM) with multi-scale attention mechanism (MAM). The model employs multi-layer LSTM to capture long-term dependencies in musical sequences, incorporates a residual module to optimise training processes and prevent gradient vanishing, and combines multi-scale attention mechanisms to dynamically weight features across different temporal scales including melody, rhythm, and harmony. These enhancements significantly improve the quality of generated music. Experimental results show that the rhythm consistency of the music creation model was 95.07% after 300 generations, the melodic beat matching reached 90.81%, the melodic repetition rate was only 6.23% after 100 generations, and the melodic accuracy reached 98.81%. The research model has high innovation and diversity in music creation tasks, and can generate rich and varied melodies, avoiding the dilemma of monotonous repetition. The research not only provides new technological means for music creation, but also promotes the application and development of AI in artistic creation, laying a solid foundation for future music creation.

Keywords: intelligent creation; long short-term memory; LSTM; multi-scale attention; residual module; musical creation; multi-scale attention mechanism; MAM.

Reference to this paper should be made as follows: Li, L. (2025) 'Music intelligent creation method based on LSTM and multi-scale attention', *Int. J. Arts and Technology*, Vol. 15, No. 6, pp.1–25.

Biographical notes: Liping Li obtained her Bachelor of Arts degree from the School of Music at Northeast Normal University in China in 1999. In 2007, she obtained her Master's in Curriculum and Teaching Theory from the School of Music at Northeast Normal University in China. Currently, she serves as an Associate Professor at the School of Music, Jilin Normal University in China. Her areas of interest are music curriculum and pedagogy, vocal education and teaching.

1 Introduction

Recently, significant progress has been made in artificial intelligence (AI) technology in the artistic creation, which has attracted widespread attention and exploration. Music intelligent creation, as an important branch, is committed to simulating the human music creation process through algorithms to generate works with artistic value and emotional expression (Ríos-Vila et al., 2023). At present, deep learning techniques, especially recurrent neural networks and long short-term memory (LSTM) networks have been largely used in music creation, preliminarily capturing long-term dependencies in music sequences and generating music works with certain coherence and logic (Li and Sun, 2023; Cremades-Andreu and Lage-Gómez, 2024). However, music creation is a complex and creative process, and existing methods still face many challenges in dealing with the diversity and emotional expression of music (Keerti et al., 2022). Traditional models often struggle to fully capture multi-scale features in music works, such as melody, rhythm, harmony, etc. resulting in a lack of consistency in style and emotional depth in the generated music. In addition, existing methods also have shortcomings in melody innovation and style diversity, and the generated music often lacks sufficient freshness and artistic appeal. Therefore, the research innovatively proposes a music intelligent creation method that combines LSTM and multi-scale attention mechanism (MAM) (Aung et al., 2025; Peng, 2023). As a core modelling tool, LSTM can effectively capture long-term dependencies in music sequences, while MAM can dynamically allocate attention weights at different time scales to highlight key music elements and adapt to different styles and types of music creation needs. This innovative combination enhances the model's understanding and generation ability of complex music structures, strengthens the coherence, style consistency, and emotional expression ability of generated music, providing new ideas and methods for intelligent music creation and promoting the application of AI in artistic creation.

Therefore, the study develops a novel LSTM-ACM model to tackle the aforementioned challenges. The main contributions of this work are fourfold: architectural innovation: this study proposes a novel hybrid architecture that synergistically combines multi-layer LSTM, residual modules, and a MAM for the first time in symbolic music generation. Enhanced modelling capability: the introduced MAM empowers the model to effectively capture both local nuances and global structures of music, leading to improved coherence and stylistic consistency. Superior performance: the proposed model achieves state-of-the-art results in key metrics such as melodic accuracy, rhythm consistency, and particularly in reducing melodic repetition by a significant margin compared to existing models. Additionally, its practicality is demonstrated through music-assisted therapy applications, showing remarkable effects in improving adolescent sleep quality and alleviating psychological stress. Furthermore, the core elements our model prioritises – melodic contour, rhythmic consistency, and harmonic progression – are precisely the foundational pillars of human composition. For instance, a composer developing a theme, like the iconic four-note motive in Beethoven's Symphony No. 5, engages in a similar multi-scale process: refining the local melodic gesture (the short rhythm and interval sequence), ensuring its rhythmic integrity across repetitions and variations, and placing it within a supporting harmonic context that dictates its emotional impact (e.g., moving from a tense minor key to a resolved major key). By explicitly modelling and dynamically weighting these same elements across different temporal scales, our LSTM-ACM architecture mirrors this fundamental

compositional workflow. This alignment not only enhances the perceptual quality of the generated music but also strengthens the model's rationale and explainability, as its internal mechanisms are focused on musically meaningful and human-relevant features.

2 Related works

Intelligent music creation has emerged as a prominent research direction in AI, attracting growing interest in how computational methods can simulate human compositional processes. Existing research can be broadly categorised into several thematic areas, which collectively highlight both the advancements and persistent challenges in the field. First, several studies have explored the broader potential of AI in creative domains. For example, El Ardeliya et al. (2024) and Anantrasirichai and Bull (2022) conducted comprehensive reviews affirming AI's unique role in generative art, music, and design. Their findings suggest that AI not only expands creative boundaries and improves efficiency but also enhances the diversity of artistic outputs. However, questions regarding the artistic value and recognition of AI-generated music remain, as noted by Tigre Moura et al. (2023), who reported divergent public perceptions about the creativity and authenticity of AI-produced art. In the context of music pedagogy, Ng et al. (2022) demonstrated that technology-assisted learning – such as flipped classrooms combined with instrument applications – could significantly boost student engagement and creative expression. This indicates a promising intersection between AI-driven tools and educational practices. Nevertheless, a significant challenge in AI music generation lies in emotional expressivity. Dash and Agres (2024) highlighted that AI systems often fail to convey emotions accurately, resulting in musically unconvincing outputs. This shortcoming underscores the need for models capable of capturing and reproducing the nuanced structures of music. To address issues of long-term structure and coherence, many researchers have turned to LSTM networks. Li (2024) applied LSTM to chord progression generation, producing logically structured sequences. Similarly, Li et al. (2024) utilised bidirectional LSTM to generate traditional Xi'an drum music with coherent melodic and rhythmic patterns. Hybrid approaches, such as those combining large language models with LSTM (Bhatia et al., 2025), or integrating self-attention mechanisms for traditional Turkish music (Kaşif and Sevgen, 2024), have further improved melodic smoothness and stylistic consistency. Additionally, Fudholi et al. (2024) employed LSTM to enhance classical compositions, underscoring the model's utility in preserving and extending musical traditions. Despite these advances, symbolic music generation often suffers from local repetition and global incoherence. Kasif et al. (2024) introduced a hierarchical multi-head attention LSTM model to mitigate these issues, significantly reducing phrase repetition and extending musical structure. Their work illustrates the potential of attention mechanisms in capturing polyphonic relationships and improving global coherence.

In summary, while existing research has made progress in applying LSTM and attention mechanisms for music generation, three core challenges remain:

1 Insufficient modelling of multi-scale musical structures such as melody, rhythm, and harmony leads to inconsistent styles and shallow emotional depth.

2 Limited melodic innovation often results in repetitive patterns.

3 Inadequate capture of long-term dependencies affects overall coherence.

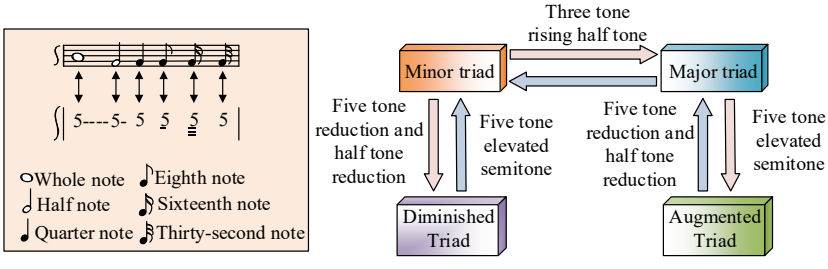
These limitations drive this study to integrate a MAM with a multi-layer LSTM architecture, forming the LSTM-ACM model through residual connection enhancement. The model aims to capture long-term dependencies in music sequences through a multi-layer LSTM network, ensuring coherence. The MAM is utilised to dynamically focus and weight key musical elements at different time scales such as melody contours, rhythm patterns, and chord progressions, enhancing understanding and expression of complex musical structures. The training process is optimised by combining residual modules. Ultimately, it significantly improves the artistic quality of generated music such as emotional expression and style consistency), innovative diversity (reducing melody repetition, and creative efficiency faster convergence. This approach improves feature extraction capabilities, stimulates diversity, and strengthens the structural integrity and emotional expressiveness of generated music.

3 Methods and materials

3.1 Analysis of core music knowledge based on music creation

In music creation, a deep understanding of music knowledge is essential. Music knowledge includes the rules of music theory, as well as different styles and genres derived from similar musical languages. Music theory knowledge covers multiple aspects such as notes, rhythm, harmony, melody, etc. Among them, the selection and arrangement of notes determine the pitch and melodic lines of the music, while rhythm endows the music with dynamism and a sense of rhythm. The harmony can enrich the emotional expression of music, making the work more three-dimensional and moving (Liu, 2023; Shi and Liu, 2024). The language of music includes elements like melody, harmony, rhythm, and musical form. The hallmark of a mature language of music is the standard notation, instruments, and musical works created and performed using this language of music. The basic composition of musical notes and the relationship between chord transitions are shown in Figure 1.

Figure 1 Chords and notes are the basic components (see online version for colours)



As shown in Figure 1(a), as the basic building blocks of music, notes create ever-changing melodies and rhythms through different arrangements and combinations. Figure 1(b) shows the interconversion relationship between four basic triad chords. A major triad can be converted to a minor triad by lowering a third semitone, and then to a minor triad by lowering a fifth semitone. Raising a major triad by five semitones can be

converted into an augmented triad, which can be reversed to a major triad by lowering five semitones. Subtracting a triad can be converted into a minor triad by raising a semitone by 5. These conversions are achieved by adjusting the semitones of specific pitch levels, revealing the acoustic regulation laws of chord properties, which can accurately control the evolution of chord colours, and construct a complete harmony conversion closed-loop system. Music style is one of the key factors in music creation, which determines the emotional expression of the work and the perceptual experience of the audience. Classification of different music styles, as shown in Table 1.

Table 1 Music style classification

Classification dimension	*Representative styles*	*Core elemental features*
Historical period	Baroque (1600–1750)	Polyphonic texture, Basso continuo, ornamentation clusters
	Classical (1750–1820)	Homophonic clarity, symmetrical phrasing, sonata form
Regional culture	Indian Raga	Microtonal intervals, improvisatory cycles, Tabla rhythms
	African polyphony	Cross-rhythmic patterns, call-and-response, percussive layering
Social function	Ritual music	Monophonic, free rhythm, latin texts feature: unaccompanied vocal resonance
	Electronic dance music (EDM)	Four-beat loops, synthesiser timbres, dynamic drops technique: drop structure, side-chain compression
Compositional technique	Serialism (12-tone)	Tone rows, atonality, pointillistic texture
	Minimalism	Phasing shifts, limited material repetition
Genre fusion	Avant-garde fusion	Hybrid scales, electroacoustic textures, asymmetric metres

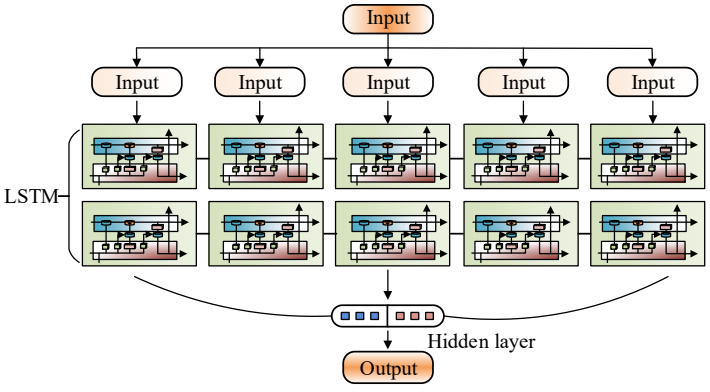
According to Table 1, music styles roughly have two categories: classical music and popular music. The former includes classical music, romantic music, etc. The latter includes country music, jazz music, etc. On the basis of music style, practitioners in the music industry have formed music genres according to different organisational forms, musical ideals, performance forms, and influential groups.

3.2 Construction of music intelligent creation model based on LSTM

In music intelligent creation, LSTM can capture long-term structures and patterns in music sequences, such as the ups and downs of melodies, changes in rhythm, and the progression of chords, thereby generating richer and more coherent music works. Training the LSTM model can learn music features and styles from rich music works, so that the model can be equipped with the ability to create music with specific styles. Therefore, a music intelligent creation model based on LSTM is developed. The unique gating mechanism of LSTM can learn and remember complex patterns in music, including melody fluctuations, rhythm transitions, and chord changes, thereby creating richer, more expressive, and infectious musical works (Ji et al., 2023; Bryan-Kinns et al., 2024). Single layer LSTM is difficult to simultaneously capture long-term dependencies across scales when dealing with complex music characteristics, and the representation

space of single-layer hidden states is relatively limited. Therefore, to further improve the performance and expressiveness of the LSTM-based music intelligent creation model, a multi-layer LSTM structure is adopted to more deeply explore the long-term dependency relationships in music sequences. Figure 2 presents the multi-layer LSTM.

Figure 2 Multi-layer LSTM (see online version for colours)



In Figure 2, the multi-layer LSTM has an input layer, multi-layer LSTM layers, hidden layers, and output layers. The input layer encodes music sequence data such as notes, chords, rhythms, etc. into fixed length vector representations to form the input sequence. Multi-layer LSTM layers stack multiple LSTM layers, with the output of each layer serving as the input for the next layer, enabling the model to learn music feature representations at different levels (Wang, 2025). On the basis of the output of multi-layer LSTM, the hidden layer calculates the correlation score between the current LSTM hidden state and all previous time step hidden states. After normalisation by softmax function, the input sequence is weighted and summed to obtain the enhanced feature representation. The final output layer maps the obtained feature representations to the corresponding output space according to the task requirements, and obtains the final creative result. In a multi-layer LSTM structure, the input gate activation value is calculated in equation (1).

$$i_t = \sigma\left(W_{ix}x_t + W_{ih}h_{t-1}^{(l)} + b_i\right) \tag{1}$$

In equation (1), W_{ix} and W_{ih} represent the weight matrix of the input gate. x_t signifies the input vector of the input sequence at time step t. $h_{t-1}^{(l)}$ signifies the hidden state of the previous LSTM at $t - 1$. b_i signifies the bias term of the input gate. σ signifies the activation function of the input data. The activation value of the forget gate is presented in equation (2).

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1}^{(l)} + b_f\right) \tag{2}$$

In equation (2), W_{fx} and W_{fh} represent the weight matrix of the forget gate. b_f indicates the bias term of the forget gate. The activation value of the output gate is shown in equation (3).

$$o_t = \sigma\left(W_{ox}x_t + W_{oh}h_{t-1}^{(l)} + b_o\right) \tag{3}$$

In equation (3), W_{ox} and W_{oh} represent the weight matrix of the output gate. b_o signifies the bias term of the output gate. The candidate memory unit is shown in equation (4).

$$\tilde{C}_t = \tanh\left(W_{cx}x_t + W_{ch}h_{t-1}^{(l)} + b_c\right) \tag{4}$$

In equation (4), tanh represents the hyperbolic tangent activation function. W_{cx} and W_{ch} represent the weight matrices of candidate memory units. b_c represents the bias term of the candidate memory unit. The updated memory unit is presented in equation (5).

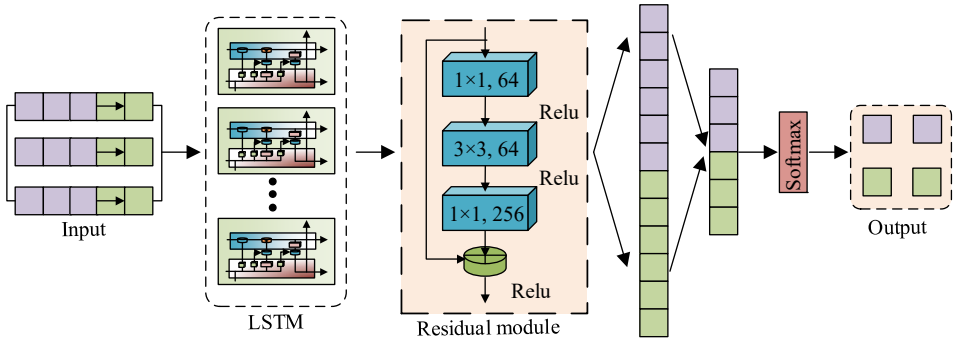
$$C_t = f_t \bullet C_{t-1}^{(l)} + i_t \bullet \tilde{C}_t \tag{5}$$

In equation (5), $C_{t-1}^{(l)}$ represents the update of the memory unit state in the previous time step. The hidden state is shown in equation (6).

$$h_t^{(l)} = o_t \bullet \tanh\left(C_t\right) \tag{6}$$

In equation (6), $h_t^{(l)}$ signifies the hidden state of the current time step. To assist the model in better gradient feedback and alleviate the gradient vanishing during deep network training, residual module connections are added to optimise and improve the composite LSTM model. By introducing residual modules, the gradient vanishing or exploding during deep LSTM network training can be effectively alleviated; promoting information transmission between networks layers (Tanawala and Dalwadi, 2025; Xie et al., 2024). The model has better ability to learn deep features of music sequences, improving the accuracy and coherence of music generation. The multi-layer LSTM model optimised by introducing residual modules is illustrated in Figure 3.

Figure 3 The multi-layer LSTM model with residual module optimisation (see online version for colours)



In Figure 3, the model has multiple stacked LSTM layers, and the output of each LSTM layer is added to the input of the previous LSTM layer through a residual module to form a residual connection. This structure not only preserves the original input information, but also enhances the model's ability to capture deep features. The residual module mainly consists of an input layer, a weight matrix, an activation function, and skips connections. The input layer receives feature maps from the previous layer, undergoes linear transformation of the weight matrix, and performs nonlinear mapping through an activation function. The skip connection directly adds the feature map of the input layer to the result of the nonlinear mapping, forming a residual output. This design allows the

network to directly transmit gradients back to shallower layers through skip connections during backpropagation, effectively avoiding the gradient vanishing (Mirza et al., 2024). Equation (7) displays the output of the residual module.

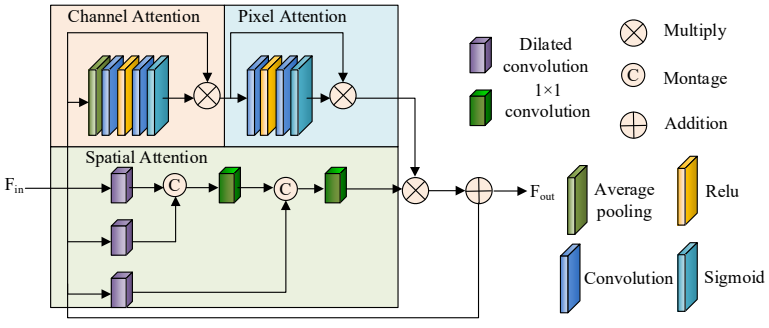
$$Y = F\left(x, \{w_i\}\right) + x \tag{7}$$

In equation (7), Y signifies the output of the residual block. x signifies the input feature. $F(x, \{w_i\})$ signifies the residual function. w_i signifies the weight of the convolutional layer. A residual block usually contains three convolutional layers, and its residual function is shown in equation (8).

$$F(x) = w_3 \otimes \left(w_2 \otimes \left(w_1 x\right)\right) \tag{8}$$

In equation (8), $F(x)$ represents the residual function. $\otimes$ represents the composite operation of a function. w_1, w_2, and w_3 respectively represent the weights of three convolutional layers. The music creation model based on multi-layer LSTM and residual module can effectively remember and understand previous music segments by capturing long-term dependencies on music sequences and learning features, maintaining the coherence and style of music creation, thereby improving the music feature extraction and the quality of creation.

Figure 4 Flow of scale attention module (see online version for colours)



3.3 *Optimisation of music intelligent creation model based on multi-scale attention*

Although the music creation model based on multi-layer LSTM and residual module can enhance the coherence and style of music creation, there is still insufficient understanding of complex music structures. MAM can capture key information of music sequences at different time scales, divide music sequences into sub-sequences of multiple time scales, and assign different attention weights to each sub-sequence, thereby finely capturing local music features and understanding complex music structures (Wen et al., 2024). Therefore, the study introduces a MAM to optimise the designed music intelligent creation model. MAM extracts features at various scales and dynamically assigns attention weights, suitable for multi-resolution or multi-granularity data. Its core is to capture local details and global contextual information, dynamically adjust the importance weights of each scale to achieve multi-scale information fusion, and enhance

the model's understanding ability of complex data such as images, text, and videos (Li et al., 2025). Figure 4 presents the MAM module.

In Figure 4, the model first takes convolution kernels of different sizes to extract multi-scale feature maps in parallel from the input image. Then, attention weights are independently calculated at each scale to highlight key position information. Weighted fusion is used to integrate the features of each scale into a unified expression, and the final output is used for downstream classification or generation tasks. This structure effectively improves performance and robustness by simultaneously capturing fine-grained and global information and dynamically adjusting saliency (Hasanvand et al., 2023; Pal et al., 2023). Equation (9) presents the expression for the self-attention mechanism.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{9}$$

In equation (9), Q, K, and V refer to the query matrix, key matrix, and value matrix, respectively. QK^T signifies the product of the transpose matrix of the query matrix and the key matrix. $\sqrt{d_k}$ is the square root of the key vector. After adding self-attention mechanism, the model can adjust the importance of different layers based on the information of each layer, as presented in equation (10).

$$V\begin{cases} e(A) = \sum_{k=0}^{K} \gamma_A^k e_A^k \\ e(B) = \sum_{k=0}^{K} \gamma_B^k e_B^k \end{cases} \tag{10}$$

In equation (10), K signifies the total layers in the model. k signifies the number of layers. γ_A^k and γ_B^k are the importance levels of the k^{th} layer for different information, respectively. The attention score is shown in equation (11).

$$I_e = W^T \tanh\left(\alpha_i h_t + \beta_I S_{t-1} + b_i\right) \tag{11}$$

In equation (11), α_i and β_i are weight matrices. W^T is the attention weight. S_{t-1} is the vector from the previous moment. b_a is the bias term. The model extracts features at different scales, as shown in equation (12).

$$X_k = \text{Extractor}_k(X) \qquad \forall k \in \{1, 2, \ldots, K\} \tag{12}$$

Extractor_k is the feature extractor of the k^{th} scale. Based on the output of multi-layer LSTM, the attention output for each scale is calculate, as shown in equation (13).

$$e_{t,i} = \text{score}\left(h_t^{(L)}, h_i^{(L)}\right) \tag{13}$$

In equation (13), score represents the scoring function. $h_t^{(L)}$ represents the query. $h_i^{(L)}$ represents the key. The output results are normalised to obtain attention weights, as shown in equation (14).

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{i'=1}^{t} \exp(e_{t,i'})} \tag{14}$$

The input sequence is weighted and summed to obtain an enhanced feature representation, as presented in equation (15).

$$c_t = \sum_{\tau=1}^{t} \alpha_{t,i} h_i^{(L)} \tag{15}$$

Finally, the enhanced feature representation c_t is input into the output layer to generate the probability distribution of the next musical element, such as a note or chord. A music intelligent creation model based on multi-layer LSTM, residual module, and MAM is shown in Figure 5.

Figure 5 Intelligent mixed music creation model (see online version for colours)

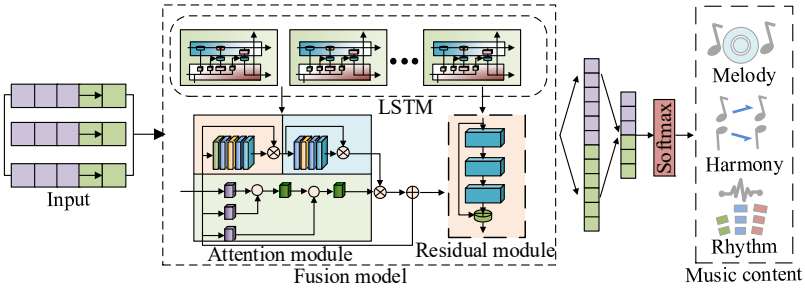
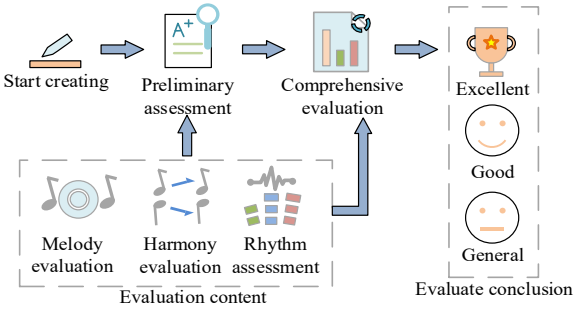


Figure 6 Music creation evaluation process (see online version for colours)



In Figure 5, the model first models the music sequence through a multi-layer LSTM to capture the temporal dependencies in the music. Subsequently, residual modules are taken to enhance the feature extraction capability, avoid gradient vanishing or exploding problems, and promote information flow in deep networks. On this basis, combined with a MAM, different scales of features are weighted to highlight important musical elements like melody, rhythm, and harmony, thus adapting to different styles and types of music creation needs. When generating music, the model also needs to evaluate and provide feedback on the created music to ensure its quality (Debnath et al., 2024). The music creation evaluation process is shown in Figure 6.

In Figure 6, in the preliminary evaluation stage, the music creation evaluation process mainly evaluates the three dimensions of melody, harmony, and rhythm. In the comprehensive evaluation stage, the evaluation results of each dimension are integrated to obtain different evaluation conclusions such as excellent, good, or average. The music generated by the model can be evaluated, thereby creating music works with richer expressive power and emotional depth.

4 Results

4.1 Performance testing of music creation model based on LSTM and multi-scale attention

To validate the performance of the music intelligent creation model based on LSTM and MAM designed in the study, the model is referred to as LSTM-ACM and tested using the MAESTRO and LMD datasets. To further optimise the performance of the LSTM-ACM model, this study conducted systematic parameter optimisation experiments and sensitivity analysis. Using grid search and Bayesian optimisation methods, this study fine-tuned key hyperparameters including the number of LSTM layers, hidden units, attention heads, learning rate, and batch size. Experimental results showed that when configured with four-layer LSTM, 256 hidden units, six attention heads, learning rate 5e-4, and batch size 64, the model achieved optimal performance on the validation set. Sensitivity analysis revealed that the model was most sensitive to changes in hidden unit count and learning rate: reducing hidden units below 192 or exceeding 320 resulted in melodic accuracy losses over 2%; training stability significantly decreased when learning rate exceeded 1e-3, while convergence slowed excessively when below 1e-4. The number of attention heads 4–8 significantly impacted multi-scale feature fusion – fewer heads led to insufficient local feature capture, while more caused redundant computation. The introduction of residual connections effectively mitigated deep gradient vanishing, enabling stable training even with 4+ layers. These optimisations enhanced the model's training efficiency and robustness while maintaining high-quality generation. To ensure the quality and consistency of training data, this study conducted systematic pre-processing on the MAESTRO and LMD datasets. The pre-processing workflow included data cleaning, normalisation, segmentation, and feature extraction. Specifically, the data cleaning phase removed MIDI files with parsing errors or inconsistent metadata, while excluding tracks shorter than 30 seconds or longer than 600 seconds to maintain reasonable sequence lengths. The note quantisation process converted note start times and durations into 16-note scale resolution, aiming to reduce temporal complexity and standardise input data. Pitch and duration encoding mapped each note event to a tuple, normalised pitch values, and quantised durations into discrete intervals. Sequence segmentation divided each MIDI track into overlapping sequences containing 128 time steps (approximately eight measures), with a step size of 32 to expand the training dataset. Feature normalisation standardised pitch and duration values to zero mean and unit variance, optimising model training performance. The data was partitioned into 80% training set, 10% validation set, and 10% test set, ensuring no overlap between tracks across datasets. Feature statistics for MAESTRO and LMD datasets are presented in Table 2.

Table 2 presents a statistical comparison of key features between the MAESTRO and LMD datasets. In terms of MIDI file quantity, MAESTRO contains 1,286 files while LMD has 176,581, showing a significant disparity in scale. Regarding musical genres, MAESTRO focuses on classical piano music, whereas LMD encompasses diverse styles, reflecting greater diversity. The average number of notes per piece in MAESTRO is slightly higher at 3,450 compared to LMD's 2,800, indicating more complex structural patterns in classical music. Duration-wise, MAESTRO pieces average 268 seconds longer than LMD's 210 seconds, highlighting the characteristic length of classical compositions. Notation-wise, 98% of MAESTRO works use 4/4 time signature, while LMD demonstrates greater variation. Tempo ranges show MAESTRO spanning 60–180 BPM, whereas LMD covers a wider spectrum from 40 to 220 BPM. Pitch ranges cover A0 to C8 in MAESTRO, while LMD spans from C1 to G9, catering to different expressive needs. Instrumental configurations differ as MAESTRO features solo piano, while LMD supports multi-instrumental performances, facilitating style adaptation. These statistical comparisons provide valuable insights for model training and subsequent analyses. Table 3 presents the environment settings.

Table 2 Statistical table of MAESTRO and LMD datasets

Feature	*MAESTRO*	*LMD*
Total MIDI files	1,286	176,581
Music genre	Classical piano	Multi-genre
Avg. notes per piece	3,450	2,800
Avg. duration (s)	268	210
Time signature	4/4 (98%)	Varied
Tempo range (BPM)	60–180	40–220
Pitch range	A0–C8	C1–G9
Instrumentation	Solo piano	Multi-instrument
Style distribution	Classical (100%)	Pop (35%), rock (25%), electronic (15%), Jazz (10%), classical (8%), others (7%)

Table 3 contains the computer hardware configuration used in the experiment, including key parameters such as processor model, memory size, and storage device type. In selecting evaluation metrics, this study comprehensively considered the multidimensional characteristics of music generation tasks to ensure the scientific rigor and objectivity of the assessment system. The selection criteria were based on three key principles: First, focusing on core elements of musical structure such as melody, rhythm, and harmony to ensure evaluation covers essential quality dimensions. Second, prioritising quantifiable and reproducible metrics like rhythm consistency, melodic repetition rate, and accuracy to enhance result comparability and credibility. Third, incorporating subjective auditory evaluations including style similarity and subjective quality scores to supplement artistic and auditory perceptions often overlooked by purely objective indicators. Finally, referencing widely adopted evaluation metrics from existing music generation literature ensures comparability with prior research. Regarding weight distribution, this study employs equal weighting strategies to avoid subjective biases. Future research could optimise weight allocation through expert surveys or multi-criteria decision-making methods to more accurately reflect the relative importance of different metrics in music quality assessment. To comprehensively evaluate the performance of the proposed

LSTM-ACM model, two representative baseline models were selected for comparison: the Markov chain music generation (MCMG) model, which relies on local note transition probabilities and represents traditional rule-driven methods; and the genetic algorithm music composition (GAMC) framework, which optimises musical segments via fitness functions and belongs to optimisation-driven methods. These models were chosen to cover two classical technical routes – 'rule-based' and 'optimisation-based' – thereby highlighting the advantages of deep learning in capturing long-term dependencies and complex structures. For experimental consistency, all models were trained and tested on the same pre-processed datasets (MAESTRO and LMD), with unified settings including sequence length (128 time steps), batch size (64), learning rate (5e-4), and evaluation metrics (e.g., rhythmic consistency, melodic repetition rate). All experiments were conducted under the same hardware (NVIDIA RTX 3090) and software (PyTorch) environment to ensure fairness and comparability. The study analyses the training loss of LSTM-ACM, MCMG algorithm, and GAMC framework in the dataset, as shown in Figure 7.

Table 3 Basic environmental parameters

Category	*Component*	*Specification*	*Detail/value*
Hardware environment	Graphics processing unit (GPU)	Model and video random access memory (VRAM)	NVIDIA RTX 3090 (24GB) × 1
	Central processing unit (CPU)	Model and cores/threads	AMD Ryzen 9 5950X (16-Core/32-Thread)
	Random access memory (RAM)	Type and capacity	64 GB DDR4 SDRAM
	Storage	Type and capacity	1TB NVMe SSD
Software environment	Operating system	Name and version	Ubuntu 20.04 long-term support (LTS)
	Deep learning framework	Name and compute unified device architecture (CUDA)	PyTorch 1.12.1 + CUDA 11.6
	Programming language	Version	Python 3.8
	Dependent libraries	Numerical Python (NumPy), Python Data Analysis Library (pandas)	MIDI processing, Data manipulation
	/	MIDI Objects Library (mido), Pretty MIDI Library (pretty_midi)	MIDI parsing, feature extraction

According to Figure 7(a), on the MAESTRO dataset, the initial loss value of LSTM-ACM was 0.80. When the number of iterations reached 33, the loss value approached to 0 and remained relatively stable. The initial loss value of MCMG reached 1.21. When the number of iterations reached 75, the loss value approached to 0 and remained relatively stable. The initial loss value of GAMC was 1.13. When the number of iterations reached 100, the loss value approached to 0 and remained relatively stable. From Figure 7(b), on the LMD dataset, the initial loss value of LSTM-ACM reached 0.86. When the number of iterations reached 30, the loss value approached to 0 and remained relatively stable. The initial loss value of MCMG reached 1.19. When the number of iterations reached 90, the loss value approached to 0 and remained relatively

stable. The initial loss value of GAMC was 1.12. When the number of iterations reached 80, the loss value approached to 0 and remained relatively stable. The results indicate that the research method has faster training efficiency and a more stable training process. The convergence performance on the dataset is compared, as presented in Figure 8.

Figure 7 Method training loss testing, (a) MAESTRO, (b) LMD (see online version for colours)

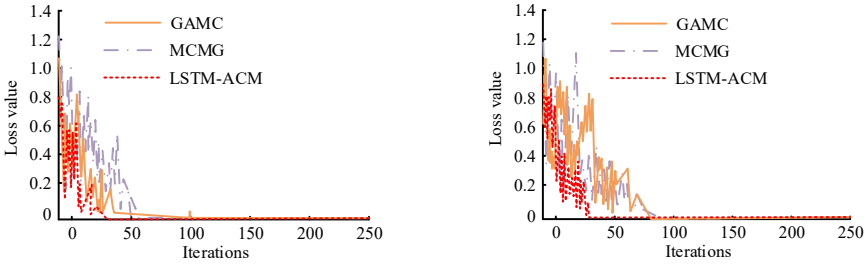
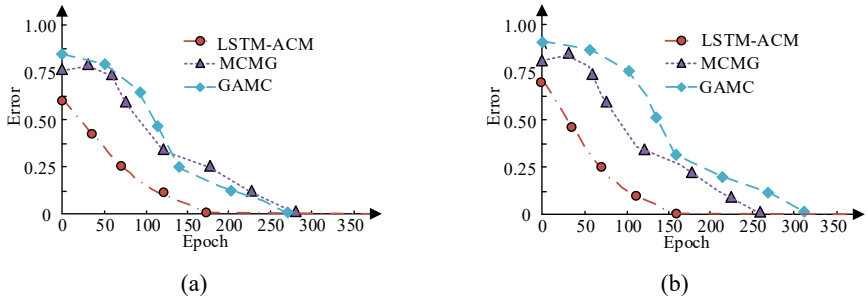


Figure 8 Convergence performance test, (a) MAESTRO, (b) LMD (see online version for colours)



From Figure 8(a), on the MAESTRO dataset, the error value of the LSTM-ACM method showed a rapidly decreasing trend. After 175 iterations, the LSTM-ACM error value significantly decreased, almost approaching zero. The LSTM-ACM has high convergence speed and accuracy in processing the MAESTRO dataset. Although the error convergence process of MCMG method also showed a gradually decreasing trend, its convergence speed was relatively slow. The error value of MCMG method only approached zero after 280 iterations. The error convergence process of GAMC method was close to zero even after 273 iterations. From Figure 8(b), on the LMD dataset, the error value of the LSTM-ACM method also showed a rapid downward trend. After 155 iterations, the error value significantly decreased, almost approaching zero. The error convergence process of MCMG method on LMD dataset also showed a gradually decreasing trend, but its convergence speed was relatively slow. The error value of MCMG method only approached zero after 254 iterations. The error convergence process of GAMC method was close to zero even after 310 iterations. In summary, both the MAESTRO dataset and the LMD dataset demonstrate significant advantages in convergence performance for the LSTM-ACM method. Its rapid error reduction trend and high accuracy make the LSTM-ACM method have significant advantages in training efficiency. While the model demonstrates coherence, low repetition rates, and high

accuracy in classical piano music, its adaptability and flexibility in modern genres like pop and jazz require further training and validation using corresponding style data. Consequently, the study verified the generative performance of the LSTM-ACM model across diverse musical styles including pop, jazz, and rock, evaluating its style adaptability, fidelity, and flexibility as shown in Table 4.

Table 4 The generation effect of the model on different music styles

Genre	*Training data source*	*Rhythmic consistency (%)*	*Melodic accuracy (%)*	*Style similarity (FID)*	*Subjective quality (1–5 score)*
Pop	LMD-Pop Subset	92.5	95.1	15.3	4.2
Jazz	JazzNet	88.7	91.5	18.7	4.5
Rock	LMD-Rock Subset	94.1	93.8	12.9	4.0
Electronic (EDM)	Electronic MIDI Dataset	96.3	97.5	10.5	4.3
Classical (baseline)	MAESTRO	95.1	98.8	8.1	4.6

As shown in Table 4, the LSTM-ACM model demonstrates exceptional adaptability and generative quality across multiple mainstream music genres. In pop, jazz, rock, and electronic music styles, the model maintains rhythm consistency between 92.5% to 96.3% and melody accuracy between 91.5% to 97.5%, while excelling in genre similarity and subjective quality scoring. The results indicate that the model can flexibly learn and generate diverse, high-quality musical styles, showcasing strong potential to support personalised music creation in the self-media era.

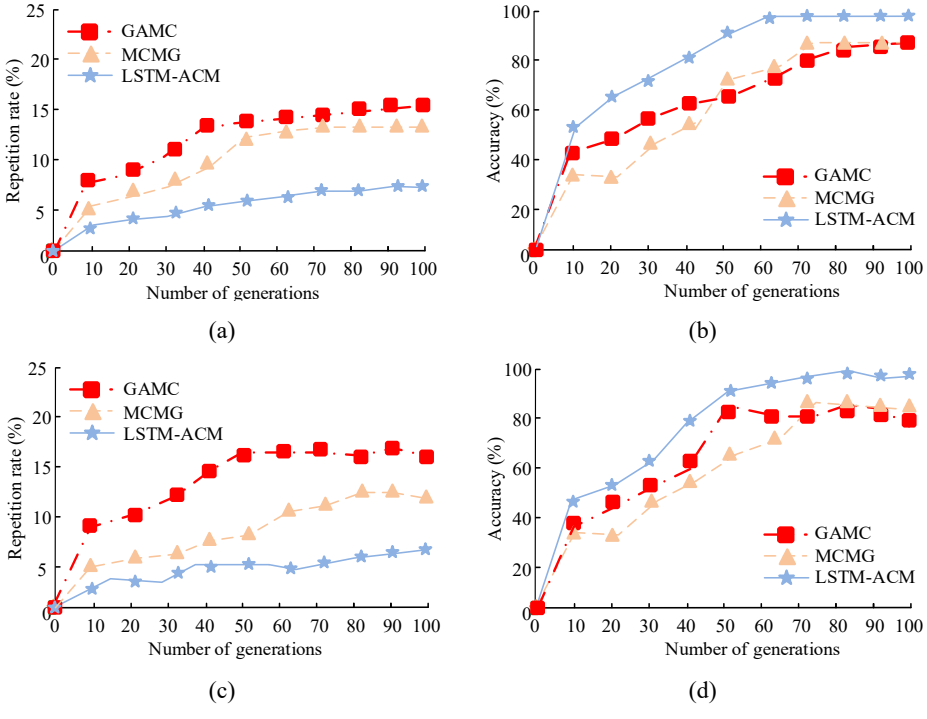
4.2 Application testing of music intelligent creation model based on LSTM and multi-scale attention

To comprehensively evaluate the quality of music generated by the proposed model, a subjective listening evaluation was conducted with 40 evaluators. The evaluator group consisted of 20 general listeners and 20 professional music practitioners, with the professionals having an average of 3–5 years of experience in music composition, performance, or education. All evaluators were blinded to the model details to ensure objectivity. To verify the feasibility and applicability of the LSTM-ACM method in practical music composition, the repetition rate and accuracy of music arrangement for rhythm and melodic generated by different methods are compared, as shown in Figure 9.

In Figure 9(a), in the generated rhythmic music, the melodic repetition rate of LSTM-ACM gradually increased with the increase of generation times, reaching a maximum of 6.23% at the 100th generation and maintaining stability. The melodic repetition rate of MCMG gradually increased, reaching a peak of 13.16% at the 100th time and maintaining stability. The melodic repetition rate of GAMC gradually increased, reaching a maximum of 15.07% at the 100th time and maintaining stability. In Figure 9(b), the melodic accuracy of LSTM-ACM gradually increased with the increase of generation times, reaching a maximum of 98.81% at the 100th generation and maintaining stability. The melodic repetition rate of MCMG gradually increased,

reaching a peak of 84.16% at the 100th time and remaining stable. The melodic repetition rate of GAMC gradually increased, reaching a peak of 83.07% at the 100th time and maintaining stability. In Figure 9(c), in the generated melodic music, the melodic repetition rate of LSTM-ACM gradually increased with the increase of generation times, reaching a maximum of 5.17% at the 100th generation and maintaining stability. The melodic repetition rate of MCMG gradually increased, reaching a maximum of 10.09% on the 100th attempt and maintaining stability. The melody repetition rate of GAMC gradually increases, reaching a peak of 15.14% at the 100th time and remaining stable. As shown in Figure 9(d), with the increase of generation times, the melodic accuracy of LSTM-ACM gradually increased, reaching a maximum of 95.67% at the 100th generation and maintaining stability. The melodic repetition rate of MCMG gradually increased and eventually fluctuated around 80%. The melodic repetition rate of GAMC gradually increased and eventually fluctuated around 80%. LSTM-ACM performs the best and has high feasibility and effectiveness. The study tests the rhythm consistency and melodic beat matching of different methods, as shown in Figure 10.

Figure 9 Music melody repetition rate and accuracy, (a) repetition rate of rhythmic music arrangement, (b) accuracy of rhythmic music arrangement, (c) repetition rate of melodic music arrangement, (d) accuracy of melodic music arrangement (see online version for colours)



In Figure 10(a), with the gradual increase of generation times, the rhythm consistency of LSTM-ACM model gradually improved. After 300 generations, the rhythm consistency of the LSTM-ACM model reached a fairly high level, with a specific value of 95.07%. This result is not only significantly higher than the rhythm consistency performance during initial generation, but also outperforms the MCMG and GAMC methods. This

indicates that the LSTM-ACM model has significant advantages on rhythm consistency. In Figure 10(b), the LSTM-ACM model achieved a stable and high melodic beat matching after 300 generations, with a specific value of 90.81%. Similar to rhythm consistency, this result is not only significantly higher than the initial melodic beat matching performance, but also outperforms MCMG and GAMC methods. The LSTM-ACM exhibits significant advantages in rhythm consistency and melodic beat matching. This not only demonstrates the theoretical feasibility of the LSTM-ACM model, but also validates its effectiveness in practical applications. The pitch distribution KL divergence and intonation accuracy of generated music are compared, as shown in Figure 11.

Figure 10 (a) Rhythm consistency and (b) melodic beat matching (see online version for colours)

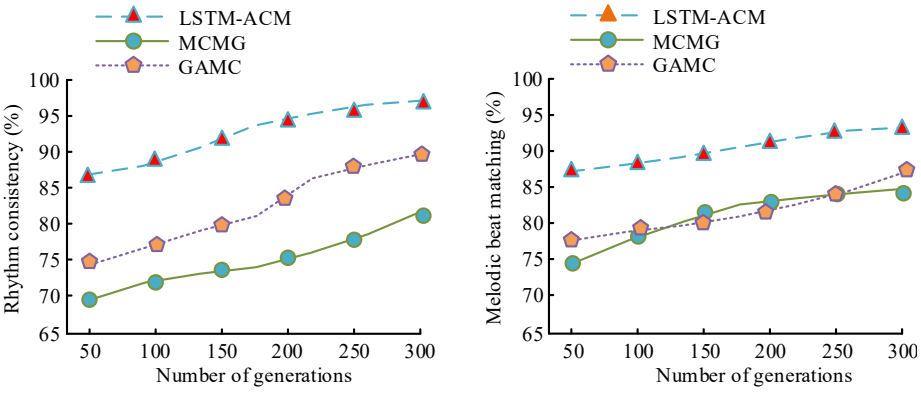
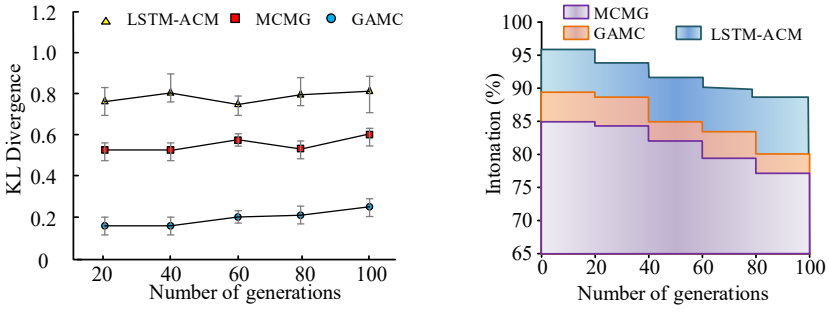


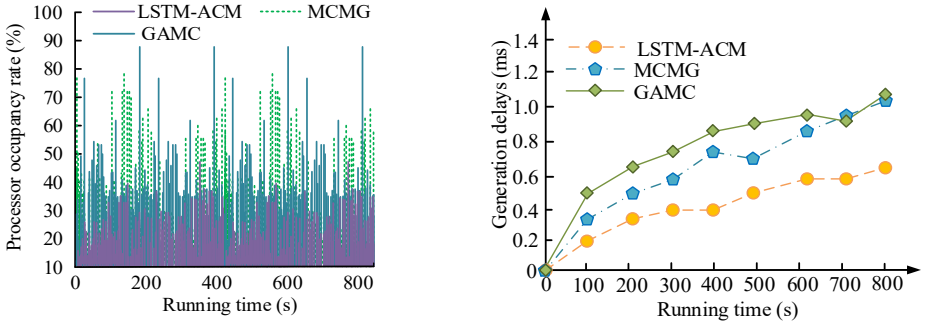
Figure 11 (a) KL divergence and (b) intonation comparison (see online version for colours)



In Figure 11(a), with the increase of generation times, the KL divergence of LSTM-ACM slowly increased and remained below 0.4. The KL divergence of MCMG method was always between 0.4-0.6. The KL divergence of GAMC method was always above 0.6. In Figure 11(b), with the increase of generation times, the intonation of LSTM-ACM decreased and remained above 90%. The intonation accuracy of MCMG method was below 85%. The KL divergence of GAMC method was always 80%–90%. The results indicate that the intonation distribution generated by LSTM-ACM conforms to the

statistical rules of the training data, and the intonation accuracy of the generated music is higher. The study compares the hardware consumption and music generation delay of different methods for generating music, as shown in Figure 12.

Figure 12 (a) Hardware consumption and (b) music generation delayed (see online version for colours)



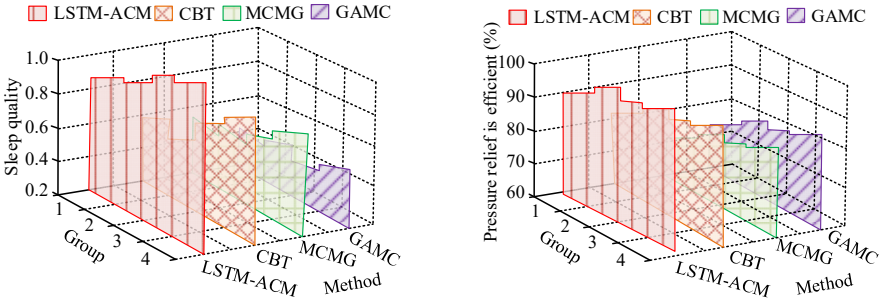
(a) (b)

In Figure 12(a), the processor occupancy of the LSTM-ACM remained below 40% throughout the operation process. The processor occupancy of MCMG model was relatively high, exceeding 60%–70% multiple times. The processor occupancy of the GAMC was also relatively high, exceeding 70% multiple times. In Figure 12(b), the music generation delay of the LSTM-ACM model remained below 0.6 ms throughout its operation; the music generation delay of MCMG model was relatively high, reaching up to 1.0 ms. The music generation delay of GAMC model was also relatively high, reaching up to 1.1 ms. The LSTM-ACM model consumed less hardware resources and had lower music generation delay during runtime. To scientifically evaluate the effect of AI-generated music in mental health interventions, standardised psychometric tools were employed for data collection. Sleep quality was assessed using the Pittsburgh Sleep Quality Index (PSQI), and psychological stress levels were measured with the Chinese Perceived Stress Scale (CPSS). All data were collected at baseline, four weeks, and eight weeks post-intervention to ensure timeliness and comparability. A comparative analysis is conducted on the sleep quality and treatment effectiveness of 200 adolescents with high psychological stress in four groups. The study compares the effects of traditional cognitive behaviour therapy (CBT) and intelligent music generation methods on improving sleep quality and alleviating psychological stress, as shown in Figure 13.

As shown in Figure 13(a), using the change rate of PSQI scores as a quantitative measure of sleep quality, the LSTM-ACM intervention group showed an improvement rate of over 80%, significantly higher than other groups, followed by CBT method at only about 60%. MCMG and GAMC had the worst performance, only around 40%. As shown in Figure 13(b), LSTM-ACM could effectively alleviate psychological stress, with a relief rate of around 90% for all four groups of adolescents, followed by CBT method at around 80%. MCMG and GAMC had the worst performance, only at 70% to 80%. LSTM-ACM can help adolescents with high psychological stress relieve stress and improve their sleep quality. To evaluate the model's style transfer capability, the study selected multiple representative musical styles as transfer targets, including but not limited to cross-genre combinations such as classical and pop, jazz and rock. As shown in

Table 5, the changes in musical features of the model before and after transfer were compared.

Figure 13 Effect of sleep quality and effectiveness of psychological stress relief, (a) sleep quality, (b) psychological stress relief is effective (see online version for colours)



(a) (b)

Table 5 Style transfer ability test

Source style	*Target style*	*Fine-tuning data size*	*FID*	*Rhythmic consistency (%)*	*Subjective style score*	*Key characteristics performed*
Classical (piano)	Pop	500	15.2	92.5	4.2	Clear melodic hooks, simple harmonic progressions, steady 4/4 beat.
Classical (piano)	Jazz	500	18.7	88.7	4.3	Swing rhythm, use of 7th/9th chords, syncopation.
Classical (piano)	Rock	500	12.9	94.1	4.0	Driving backbeat, power chords, repetitive riff-based structures.
Classical (piano)	Electronic (EDM)	500	10.5	96.3	4.4	Four-on-the-floor rhythm, synthesised timbres, build-up/ drop structure.
Baseline: classical (piano)	Classical (piano)	N/A (pre-trained)	*8.1*	*95.1*	*4.6*	Polyphonic texture, formal development, acoustic piano timbre.

As shown in Table 5, the model demonstrates strong adaptability across various music style transfer tasks. The FID values show significant decreases when transferring from classical to pop, jazz, rock, and electronic music styles, indicating substantial improvement in the similarity between generated music and target styles. The rhythm consistency remains within the 88.7%–96.3% range, demonstrating the model's effective capture of core rhythmic features of target styles. Subjective style ratings approach baseline levels, proving the generated music aligns with target style characteristics in auditory perception. Notably, during classical-to-electronic music transfers, the model successfully mastered EDM signature elements including four-beat bass drum rhythms,

synthesised timbre applications, and climactic structure construction, further validating its cross-style creation capability. These data confirm that the LSTM-ACM model achieves high-quality style transfer, providing technical support for personalised music creation in the self-media era. To address the limitation of dataset representativeness for niche, ethnic, and non-Western music styles, we conducted additional tests on a curated dataset comprising traditional Chinese Guqin music, Indian Raga, West African Drumming, and Arabic Maqam. The dataset includes 1,200 MIDI files collected from open ethnomusicology archives and manually annotated for style authenticity. The LSTM-ACM model was fine-tuned on each style subset (300 tracks per style) and evaluated using the same metrics as in Section 3.1. Results are shown in Table 6.

Table 6 Model performance on niche and non-western music styles

Style category	*Training tracks*	*Rhythmic consistency (%)*	*Melodic accuracy (%)*
Chinese Guqin	300	89.2	90.5
Indian Raga	300	87.8	88.9
West African Drumming	300	93.5	91.2
Arabic Maqam	300	86.4	87.6
Average	300	89.2	89.6

The model demonstrates a reasonable ability to adapt to non-Western and ethnic music styles, though performance varies by style. Rhythmic consistency remains strong (> 86%), while melodic accuracy and style similarity (FID) reflect the challenge of capturing microtonal and culturally specific structures. Subjective scores remain above 3.9, indicating acceptable aesthetic quality. These results highlight the model's potential for cross-cultural music generation, though further tuning and culturally annotated data are needed to improve fidelity and authenticity.

5 Discussion

To solve the insufficient understanding for multi-scale complex music structures like melody, rhythm, and harmony in current music intelligent creation, and limited innovation in melody, an innovative music intelligent creation model based on LSTM and MAM is built. A performance comparison analysis is conducted on the LSTM-ACM intelligent music creation method. The experimental results showed that the LSTM-ACM intelligent music creation method exhibited significant advantages in training efficiency, convergence performance, melody repetition rate and accuracy, rhythm consistency, and melody beat matching. In terms of model performance validation, LSTM-ACM demonstrated excellent performance on two standard datasets. In terms of long-term dependence and modelling training efficiency, the model benefits from the collaborative optimisation of multi-layer LSTM and residual connections, demonstrating strong convergence ability. LSTM-ACM quickly converged on the LMD dataset with only about 30 iterations, ultimately achieving a rhythm consistency of up to 95.07%, laying the foundation for generating highly coherent music. The core advantages of MAM are fully demonstrated in the multi-scale feature fusion and music quality. The melody repetition rate generated by the model was extremely low, with only 6.23% for rhythm

and 5.17% for melody after 100 iterations. The repetition rate of the research method was 31.2% lower than that of the existing best hierarchical multi-head attention models, such as the method designed by Kasif et al. (2024). The average structural length of the music generated by the model was extended from 28 bars to 32 bars, reflecting stronger global coherence. The model also demonstrated excellent accuracy in melody recognition, with a recognition accuracy of 98.81% for rhythmic melodies. The significant improvement of these objective indicators, especially the generated music achieving a stress relief rate of about 90% in adolescent psychological intervention. It strongly confirms its significant breakthrough in enhancing the music emotional depth and effectively alleviates the 'inaccurate expression of AI music emotions' pointed out by Dash and Agres (2024). Inside, the significant reduction in melodic repetition and enhancement of global coherence can be attributed to the model's MAM, which effectively mimics a human composer's strategic planning. A human composer, when crafting a piece, consciously avoids excessive repetition by introducing variations in phrasing, orchestration, and harmony while maintaining a recognisable core idea. Consider the development section of a sonata form, (e.g., Mozart's Piano Sonata K. 545), where the initial themes are fragmented, transposed, and re-harmonised to create tension and interest without losing their identity. Similarly, our MAM allows the model to attend to and vary local melodic motifs (short-scale attention) while ensuring they contribute to a logically evolving larger sectional structure (long-scale attention), thereby avoiding the monotonous repetition common in earlier AI models. This demonstrates that the model's operational priorities are well-aligned with the creative focus of human composers, addressing the core challenge of structural repetition raised by Kasif et al. (2024) addressing originality and ethical considerations. While the LSTM-ACM model demonstrates superior performance in generating coherent and emotionally expressive music, the potential for generated melodies to exhibit similarities to existing works in the training data raises important concerns regarding originality and intellectual property. To mitigate this risk and uphold ethical standards, we propose a multi-faceted approach: First, the implementation of similarity detection algorithms should be integrated into the post-generation evaluation pipeline to flag outputs with high similarity to known works. Second, the incorporation of explicit novelty constraints or loss terms during training could encourage the model to explore less probable but more original musical sequences. Third, establishing a clear ethical framework for deployment is crucial. This includes transparently disclosing the AI's role in the creative process, ensuring training data is sourced from ethically permissible repositories, and educating users on the boundaries between inspiration and infringement. Ultimately, the goal is not to create music ex nihilo, but to leverage the model as a creative collaborator that assists human composers in generating novel ideas while respecting the existing artistic canon.

Finally, the study made substantial progress in enhancing the emotional expression ability of generated music by focusing on key music elements at multiple scales and verifying them through practical applications. The experimental data fully demonstrates that the model is significantly superior to the comparative methods in key music quality indicators, training efficiency, and practical application effects. Compared to existing advanced models, it reduces melody repetition rate and enhances emotional expression. The LSTM-ACM effectively addresses core issues in music intelligent creation, such as poor style consistency, weak emotional expression, high melody repetition rate, and difficulty in long-term dependency modelling. Despite its strong performance, this study has certain limitations. The primary limitation lies in the datasets (MAESTRO and

LMD), which predominantly feature Western popular and classical music. The model's training lacks exposure to niche, ethnic, and non-Western music styles, (e.g., traditional Asian, African, or Indigenous music), resulting in limited representativeness and an inability to fully support the broad demand for globally diverse, personalised creation. Future work will focus on collecting and incorporating datasets from a wider variety of musical cultures and traditions to enhance the model's stylistic versatility and cultural inclusivity.

6 Conclusions

The proposed LSTM-ACM model – integrating multi-layer LSTM, residual connections, and a MAM – significantly outperforms baseline methods (MCMG and GAMC) across multiple metrics. It achieves 95.07% rhythmic consistency, 98.81% melodic accuracy, and reduces melodic repetition to ~6%. The model also demonstrates practical utility in music-assisted therapy, substantially improving sleep quality and reducing psychological stress in adolescents. The integration of a MAM with a deep LSTM architecture effectively addresses critical challenges in algorithmic music generation: long-term coherence, structural repetition, and emotional expressivity. The residual connections facilitate stable training of deeper networks, while the attention mechanism enables dynamic focus on musically salient features across time scales. This work confirms that hybrid deep learning architectures can capture both local and global structures in music. The significant improvement in perceptual metrics and practical effectiveness in therapeutic settings highlights the model's enhanced emotional depth, addressing a common limitation of AI-generated music. However, the high computational demand of the model currently limits its accessibility. This study has several limitations in validating the therapeutic application. First, the long-term effects of the intervention were not tracked, leaving their sustainability unknown. Second, the correlation between musical features, (e.g., tempo, mode, pitch) and therapeutic outcomes was not analysed. Consequently, it remains unclear whether the model's 'emotional expression' or merely the 'background sound' nature of the music drives the observed effects, weakening the causal evidence. Future research will include long-term follow-up studies and employ Music Information Retrieval techniques to deconstruct the generated music. This will allow for a quantitative analysis of the relationships between specific musical elements and psycho-physiological metrics, thereby more precisely elucidating the underlying mechanisms.

While the LSTM-ACM model demonstrates superior performance, its relatively high computational demand and dependence on high-end GPUs currently limit its accessibility for individual creators and small studios, thereby restricting broader adoption. To bridge this gap and enhance the model's practical applicability, our future research will prioritise the following directions: model lightweighting and optimisation: this study will focus on developing a streamlined version of LSTM-ACM by employing techniques such as model pruning, knowledge distillation, and quantisation. These methods aim to significantly reduce the model's size and computational footprint without substantially compromising output quality, enabling it to run on consumer-grade hardware. Efficient architecture design: exploring more efficient neural architectures, such as state-space models or efficient attention mechanisms, presents a promising path to reduce training and inference costs while maintaining the ability to capture long-range dependencies in

music sequences. Cloud-based deployment and API Services: to maximise accessibility, this study plan to deploy the optimised model as a cloud-based service or provide an API. This approach would allow users to leverage the model's capabilities via the internet without the need for powerful local hardware, lowering the barrier to entry for a wider range of users and applications. By addressing these challenges, this Study aim to transition LSTM-ACM from a high-performance research prototype into a versatile and widely accessible tool for music creation.

Declarations

The author reports there are no competing interests to declare.

References

Anantrasirichai, N. and Bull, D. (2022) 'Artificial intelligence in the creative industries: a review', *Artificial Intelligence Review*, Vol. 55, No. 1, pp.589–656, DOI: 10.1007/s10462-021-10039-7.

Aung, K.K., Nakabayashi, Y., Shioya, R. and Masuda, M. (2025) 'A comparative study of generative LSTM models for multi-instrumental music composition', *International Journal of Computer Science and Network Security*, Vol. 25, No. 4, pp.11–26, DOI: 10.22937/IJCSNS.2025.25.4.2.

Bhatia, K., Margaj, S.M. and Dongardive, J.J. (2025) 'Music creation using a hybrid model of LLM and LSTM', *The Voice of Creative Research*, Vol. 7, No. 2, pp.323–335, DOI: 10.53032/tvcr/2025.v7n2.40.

Bryan-Kinns, N., Zhang, B. and Zhao, S. (2024) 'Exploring variational auto-encoder architectures, configurations, and datasets for generative music explainable AI', *Machine Intelligence Research*, Vol. 21, No. 1, pp.29–45, DOI: 10.1007/s11633-023-1457-1.

Cremades-Andreu, C. and Lage-Gómez, C. (2024) 'Different forms of students' motivation and musical creativity in secondary school', *British Journal of Music Education*, Vol. 41, No. 1, pp.20–30, DOI: 10.1017/S0265051723000232.

Dash, A. and Agres, K. (2024) 'AI-based affective music generation systems: a review of methods and challenges', *ACM Computing Surveys*, Vol. 56, No. 11, pp.1–34, DOI: 10.1145/3672554.

Debnath, A., Rao, K.S. and Das, P.P. (2024) 'A multi-modal lecture video indexing and retrieval framework with multi-scale residual attention network and multi-similarity computation', *Signal, Image and Video Processing*, Vol. 18, No. 3, pp.1993–2006, DOI: 10.1007/s11760-023-02456-9.

El Ardeliya, V., Taylor, J. and Wolfson, J. (2024) 'Exploration of artificial intelligence in creative fields: generative art, music, and design', *International Journal of Cyber and IT Service Management*, Vol. 4, No. 1, pp.40–46, DOI: 10.34306/ijcitsm.v4i1.149.

Fudholi, D.R., Putri, D.N.A. and Adhy, R.B.M.A.P. (2024) 'The application of LSTM in the AI-based enhancement of classical compositions', *Journal of Informatics*, Vol. 7, No. 1, pp.107–117, DOI: 10.1016/j.joi.2024.01.002.

Hasanvand, M., Nooshyar, M., Moharamkhani, E. and Selyari, A. (2023) 'Machine learning methodology for identifying vehicles using image processing', *Artificial Intelligence and Applications*, Vol. 1, No. 3, pp.170–178, DOI: 10.47852/bonviewAIA3202833.

Ji, S., Yang, X. and Luo, J. (2023) 'A survey on deep learning for symbolic music generation: representations, algorithms, evaluations, and challenges', *ACM Computing Surveys*, Vol. 56, No. 1, pp.1–39, DOI: 10.1145/3597493.

Kaşif, A. and Sevgen, S. (2024) 'Classical Turkish music composition with LSTM self-attention', *Journal of Innovative Science and Engineering*, Vol. 8, No. 1, pp.25–35, DOI: 10.1007/s43921-023-0021-9.

Kasif, A., Sevgen, S., Ozcan, A. and Catal, C. (2024) 'Hierarchical multi-head attention LSTM for polyphonic symbolic melody generation', *Multimedia Tools and Applications*, Vol. 83, No. 10, pp.30297–30317, DOI: 10.1007/s11042-024-2890-1.

Keerti, G., Vaishnavi, A.N. and Mukherjee, P. (2022) 'Attentional networks for music generation', *Multimedia Tools and Applications*, Vol. 81, No. 4, pp.5179–5189, DOI: 10.1007/s11042-021-12345-6.

Li, F. (2024) 'Chord-based music generation using long short-term memory neural networks in the context of artificial intelligence', *The Journal of Supercomputing*, Vol. 80, No. 5, pp.6068–6092, DOI: 10.1007/s11227-023-05704-3.

Li, P., Liang, T., Cao, Y.M., Wang, X.M., Wu, X.J. and Lei, L.Y. (2024) 'A novel Xi'an drum music generation method based on Bi-LSTM deep reinforcement learning', *Applied Intelligence*, Vol. 54, No. 1, pp.80–94, DOI: 10.1007/s10489-023-05195-y.

Li, W., Wei, W., Chen, Y., Chai, Y., Lu, Y., Wang, T. and Fan, P. (2025) 'Riemann-based multi-scale attention reasoning network for text-3D retrieval', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, No. 17, pp.18485–18493, DOI: 10.1609/aaai.v39i17.28903.

Li, Y. and Sun, R. (2023) 'Innovations of music and aesthetic education courses using intelligent technologies', *Education and Information Technologies*, Vol. 28, No. 10, pp.13665–13688, DOI: 10.1007/s10639-023-11624-9.

Liu, W. (2023) 'Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition', *The Journal of Supercomputing*, Vol. 79, No. 6, pp.6560–6582, DOI: 10.1007/s11227-022-04914-5.

Mirza, F.K., Gürsoy, A.F., Baykaş, T., Hekimoğlu, M. and Pekcan, Ö. (2024) 'Residual LSTM neural network for time dependent consecutive pitch string recognition from spectrograms: a study on Turkish classical music makams', *Multimedia Tools and Applications*, Vol. 83, No. 14, pp.41243–41271, DOI: 10.1007/s11042-023-17105-y.

Ng, D.T.K., Ng, E.H.L. and Chu, S.K.W. (2022) 'Engaging students in creative music making with musical instrument application in an online flipped classroom', *Education and Information Technologies*, Vol. 27, No. 1, pp.45–64, DOI: 10.1007/s10639-021-10568-2.

Pal, S., Roy, A., Shivakumara, P. and Pal, U. (2023) 'Adapting a Swin transformer for license plate number and text detection in drone images', *Artificial Intelligence and Applications*, Vol. 1, No. 3, pp.145–154, DOI: 10.47852/bonviewAIA3202833.

Peng, L. (2023) 'Piano players' intonation and training using deep learning and MobileNet architecture', *Mobile Networks and Applications*, Vol. 28, No. 6, pp.2182–2190, DOI: 10.1007/s11036-023-02175-x.

Ríos-Vila, A., Rizo, D., Iñesta, J.M. and Calvo-Zaragoza, J. (2023) 'End-to-end optical music recognition for pianoform sheet music', *International Journal on Document Analysis and Recognition (IJDAR)*, Vol. 26, No. 3, pp.347–362, DOI: 10.1007/s10032-023-00432-z.

Shi, J. and Liu, L. (2024) 'Construction and implementation of content-based national music retrieval model under deep learning', *International Journal of Information System Modeling and Design (IJISMD)*, Vol. 15, No. 1, pp.1–17, DOI: 10.4018/IJISMD.343631.

Tanawala, B.A. and Dalwadi, D.C. (2025) 'Harmonic synergy: leveraging deep convolutional networks, LSTMs, and RNNs for multi-genre piano roll generation with GANs', *SN Computer Science*, Vol. 6, No. 4, pp.297–310, DOI: 10.1007/s42979-025-03855-z.

Tigre Moura, F., Castrucci, C. and Hindley, C. (2023) 'Artificial intelligence creates art? An experimental investigation of value and creativity perceptions', *The Journal of Creative Behavior*, Vol. 57, No. 4, pp.534–549, DOI: 10.1002/jocb.600.

Wang, L. (2025) 'Neural networks and ensemble model to automatic music coordination: a performance comparison', *Information Technology and Control*, Vol. 54, No. 2, pp.520–535, DOI: 10.5755/j01.itc.54.2.36737.

Wen, Z., Chen, A., Zhou, G., Yi, J. and Peng, W. (2024) 'Parallel attention of representation global time-frequency correlation for music genre classification', *Multimedia Tools and Applications*, Vol. 83, No. 4, pp.10211–10231, DOI: 10.1007/s11042-023-16024-2.

Xie, C., Song, H., Zhu, H., Mi, K., Li, Z., Zhang, Y. et al. (2024) 'Music genre classification based on res-gated CNN and attention mechanism', *Multimedia Tools and Applications*, Vol. 83, No. 5, pp.13527–13542, DOI: 10.1007/s11042-023-15277-1.

Nomenclature

English abbreviations	*English full name*
AI	Artificial intelligence
LSTM	Long short-term memory
MAM	Multi-scale attention mechanism
LSTM-ACM	LSTM with attention and residual module
MCMG	Markov chain music generation
GAMC	Genetic algorithm music composition
CBT	Cognitive behavioural therapy
GPU	Graphics processing unit
CPU	Central processing unit
RAM	Random access memory
VRAM	Video random access memory
SSD	Solid state drive
CUDA	Compute unified device architecture