Construction of Intelligent Evaluation Model for Music Aesthetic Education Effect in Colleges and Universities—Application Exploration Based on Informer Time Series Algorithm

Ying Qi
Department of Early Childhood Education, Shaanxi Xueqian Normal University, Xi'an 710100, Shaanxi, China
Corresponding author: Ying Qi (e-mail: tschi111yq@163.com)

ABSTRACT Accurate evaluation of music aesthetic education requires effective modeling of long-term temporal dependencies and short-term dynamic variations in heterogeneous educational data. This study proposes an intelligent evaluation framework based on a parallel Informer and BiGRU-Global Attention architecture to analyze multi-source time series collected from questionnaires, classroom interactions, and practical assessments. The Informer module captures global evolutionary patterns through ProbSparse attention and multi-scale decomposition, while the BiGRU-Global Attention branch enhances sensitivity to local fluctuations and key educational events. Experimental validation on 40-week records from 1,984 students demonstrates that the proposed model achieves superior predictive performance with an MAE of 2.3, RMSE of 3.0, and MAPE of 3.5%, while maintaining an inference time of only 8.4 ms. Comparative experiments and statistical significance tests confirm its robustness and effectiveness in dynamic educational assessment. Beyond music aesthetic education, the proposed multi-source temporal fusion strategy provides a transferable data-driven methodology for intelligent signal analysis and sequential decision-making tasks, offering potential reference for information processing frameworks in electromagnetic sensing and human-centered intelligent systems.

INDEX TERMS music aesthetic education, informer, BiGRU-Global attention, multi-source time series, intelligent signal processing

I. INTRODUCTION

WITH the deepening of higher education reform, music aesthetic education has become increasingly important in enhancing college students' aesthetic abilities, cultural literacy, and overall quality, a holistic approach that is vital for fostering creativity and technical excellence in specialized fields [1,2]. Existing evaluations mostly rely on static indicators such as final exam scores and questionnaires, ignoring students' dynamic learning process during the semester, and cannot meet the needs of integrated and real-time management of process and results [3]. Some studies have conducted discrete dynamic modeling of music teaching based on the theory of multiple intelligences and found that capturing up to 65 interaction frequencies within 10 seconds is helpful in understanding students' interest fluctuations [4]. However, this method lacks continuous tracking of global trends across semesters, which can easily lead to misjudgment of long-term learning outcomes. At the same time, it fails to take into account the agile response to short-term fluctuations, and teaching interventions are often delayed and have limited effects. It is urgent to build an intelligent evaluation model that can capture ultra-long time series dependencies in real time and take into account local short-term fluctuations, so as to promote the development of music aesthetic education in colleges and universities in a precise and scientific direction.

This paper aims to develop an intelligent evaluation model that considers both long-term trends and local key fluctuations, thereby addressing the static lag and dynamic blind spot issues of traditional evaluation, with a scope that extends to specialized creative field. This paper takes global trend-node feedback-real-time response as the core causal chain, first obtains the global evolution trend across 40 weeks of the semester through the Informer module to

avoid the failure of teaching intervention due to information lag; then focuses on short-term key fluctuations through the BiGRU-Global Attention branch to ensure agile response to sudden evaluation and performance links; after the two-way information fusion, a comprehensive score is generated in real time to assist in the rolling optimization of teaching strategies. The model training adopts AdamW optimization combined with gradient clipping and a cosine annealing strategy. The experimental results show that the model is significantly better than the six mainstream baselines in terms of overall accuracy (MAE 2.3 points, RMSE 3.0 points) and relative error (MAPE 3.5%), and the significant difference is verified by T test ($p<0.001$). The inference time is only 8.4 ms, and the average response time is 9.1 ms, which is better than the other comparison models, verifying the adaptability and practicality of the method to the dynamic evaluation scenario of music aesthetic education in colleges and universities. The contributions of this paper are as follows:

(1) Innovatively constructing a parallel model based on Informer + BiGRU-Global Attention to capture the global trend of ultra-long time series and local key dynamic fluctuations, thereby improving the accuracy and robustness of aesthetic education evaluation.

(2) Applying multi-scale convolution and multi-source data trend-season decomposition in the encoding layer, combining with Global Attention to focus on key nodes, and effectively enhancing the sensitivity and interpretability of emergencies.

(3) Designing a lightweight fully connected output layer and optimize the training strategy to achieve an inference time of only 8.4 ms, meeting the dynamic and real-time evaluation needs of college music aesthetic education on the basis of high precision.

II. RELATED WORK

Early music aesthetic education assessment mainly relied on static indicators and expert scoring, which belonged to the static assessment stage. Li Y et al. studied the impact of intelligent technology on music aesthetic education curriculum innovation. The results showed that it can significantly improve student performance. However, this study did not explore how to effectively model long-term data, and the evaluation model primarily relied on stage-by-stage results, overlooking dynamic changes [5]. Most existing studies are based on theory and lack in-depth modeling of time series data, which prevents them from effectively capturing changes in students' long-term trends.

Entering the stage of single-modal time series modeling, researches began to attempt to deeply model a single time series signal. In the evaluation of music effectiveness in colleges and universities, Li P et al. studied the application of artificial intelligence and found that it can improve students' grades, but did not explore in depth how to dynamically evaluate students' progress, nor did they solve the problem of long-term time series dependency [6]. Kusumawardani S S, Wang H et al. applied the Transformer model to music education and adaptive learning systems. Their proposed learning analytics and adaptive learning systems based on the Transformer model performed well in short-term prediction. However, the high computational complexity of the Transformer model makes it difficult to cope with the frequent emergencies in music practice and is still insufficient to capture the long-term dependencies and changes in time series. In addition, the high computational complexity also makes it difficult to perform real-time updates [7,8]. Most existing studies focus on music evaluation in colleges and universities, lack the evaluation of the effect of music aesthetic education, and fail to effectively solve the problems of long-term dependence and short-term fluctuations. Real-time and accuracy still face challenges.

In recent years, research has shifted to the stage of multimodal dynamic fusion, attempting to unify the modeling of different data sources and multi-scale information. Currently, many scholars have proposed various models to address the challenge of modeling long-term time series data. The violin emotion recognition model combining CNN-BiGRU and attention mechanism proposed by Ma S et al. significantly improved the accuracy of emotion recognition, but focused more on local emotion features and failed to fully handle the temporal dependencies in long sequences [9]. The Music Informer model of Sun H et al. used the informer technology to generate music and reduce computing consumption, but it mainly focused on the generation task and is not designed for the dynamic evolution of artistic expression in music aesthetic education evaluation [10]. Existing research has achieved certain results in time series modeling and computing efficiency, but lacks an end-to-end evaluation framework that couples global trends, local fluctuations, and multi-source data. In particular, there is a research gap in the field of music aesthetic education effect evaluation.

III. CONSTRUCTION OF INTELLIGENT EVALUATION MODEL FOR THE EFFECT OF MUSIC AESTHETIC EDUCATION IN COLLEGES AND UNIVERSITIES

Currently, commonly used models in social science time series research, such as ARIMA, simple RNN, or low-order LSTM, have demonstrated good performance in scenarios involving short sequences, low noise, or single statistical indicators. However, the evaluation of the effectiveness of music and aesthetic education in universities presents significantly different complexities in terms of data format and modeling objectives. This study involves time series data (questionnaires, classroom behavior, and periodic assessments) from multiple sources and asynchronous sampling across semesters (40 weeks). Long-term trends and short-term emergencies (such as concentrated presentations and periodic assessments) coexist, making it difficult for traditional models to simultaneously address both long-term dependencies and local key nodes within a unified framework. Moreover, the evolution of aesthetic education effectiveness

is not a stationary linear process but exhibits obvious staged, periodic, and structurally abrupt changes. Simple models are prone to trend lag or insufficient response to key events in this context. Therefore, Informer's ProbSparse attention mechanism can selectively model the most informative long-range dependencies while maintaining controllable computational complexity, avoiding interference from redundant historical information in long-term predictions. Furthermore, the introduction of a BiGRU branch with global attention compensates for the insufficient sensitivity of pure long-sequence models to short-term high-impact events. The parallel design of the two is not simply to pursue model complexity, but a structured response to the inherent mechanism of the task, which is driven by "long-term trends + short-term key nodes". Therefore, compared with traditional low-overhead models, it has clear necessity and methodological rationality in this specific aesthetic education evaluation scenario.

A. INFORMER TIME SERIES ALGORITHM MODEL

Timing embedding.

In a parallel hybrid framework, this paper applies the Informer time series algorithm model [11,12] to efficiently model the ultra-long time series dependency and periodic trend of multi-source music aesthetic education data across semesters.

The study first maps the original multi-dimensional time series input to a feature space of the same dimension and adds position and time information. The input mapping formula E_t is shown in (1).

$$E_t = W_e x_t + b_e \tag{1}$$

In formula (1), b_e represents the bias, W_e represents the linear mapping matrix, and x_t represents the original multidimensional time series.

After mapping, a fixed position encoding is added to each time step t, and the position encoding formula is shown in (2). Discrete time features such as date and week are mapped to T_t through trainable embedding.

$$P_t[i] = \begin{cases} \sin\left(\frac{t}{10000^{2i/d}}\right), \text{if } i \text{ is even} \\ \cos\left(\frac{t}{10000^{2i/d}}\right), \text{if } i \text{ is odd} \end{cases} \tag{2}$$

In formula (2), $P_t[i]$ represents the positional encoding.

The total embedding representation is shown in formula (3).

$$Z_t^{(0)} = E_t + P_t + T_t \tag{3}$$

In formula (3), $Z_t^{(0)}$ represents the total embedding representation.

Trend/Seasonal decomposition.

To improve the ability to capture cyclical fluctuations, the sliding average is used to decompose the input sequence into trend and seasonal terms.

$$T_t = \frac{1}{2k+1} \sum_{j=-k}^{k} Z_{t+j}^{(0)}, \qquad S_t = Z_t^{(0)} - T_t \tag{4}$$

In formula (4), $2k+1$ represents the sliding window size, and mirror padding is used at the endpoints. T_t represents the trend term, and S_t represents the season term.

Sparse self-attention encoder.

Now linearly map the trend term and the season term respectively to generate the query (Q), key (K), and value (V) of multi-head attention, as shown in formula (5).

$$\begin{aligned} Q^{(h)} &= Z^{(l-1)} W_Q^{(h)}, \\ K^{(h)} &= Z^{(l-1)} W_K^{(h)}, \\ V^{(h)} &= Z^{(l-1)} W_V^{(h)} \end{aligned} \tag{5}$$

In formula (5), $W_Q^{(h)}$, $W_K^{(h)}$, and $W_V^{(h)}$ represent the projection matrix of the h-th head, and l represents the number of encoder layers.

The traditional self-attention complexity of $O(L^2)$ is difficult to extend to very long sequences. The Informer time series algorithm model applies ProbSparse Attention [13], which only retains a small number of u queries with the most information. The selection method is based on the sparseness of the attention distribution between each query and all keys, expressed as $\mathrm{KL}(\mathrm{softmax}(Q_t K^\top) \| \,\mathrm{Uniform})$. The Informer time series algorithm model retains the attention calculation for the first u queries with the highest KL value, and directly simplifies the calculation of the remaining queries with the average key, reducing the complexity to $O(L \log L)$. The attention calculation is shown in formula (6).

$$A(Q,K,V) = \mathrm{softmax}\left(\frac{\tilde{Q}K^T}{\sqrt{d_h}}\right)V + \tilde{Q}\left(\frac{\sum_{j=1}^{L} K_j}{L}\right)^T \tag{6}$$

In formula (6), $\widetilde{Q}$ means that only the selected u rows are included.

The outputs of each head are concatenated and projected back to the original dimension, as shown in formula (7).

$$\begin{aligned} H^{(l)} &= \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_H) W^O, \\ \mathrm{head}_h &= A^{(h)}(Q^{(h)}, K^{(h)}, V^{(h)}) \end{aligned} \tag{7}$$

In formula (7), $H^{(l)}$ represents the multi-head sparse attention output, and W^O represents the projection matrix.

The residual connection and normalization processing are shown in formula (8).

$$Z^{(l)} = \mathrm{LayerNorm}\left(Z^{(l-1)} + \mathrm{Dropout}(H^{(l)})\right) \tag{8}$$

In formula (8), LayerNorm represents the layer normalization process, and $Z^{(l)}$ represents the residual connection and normalization process result.

Multi-scale convolution fusion.

On the last layer of the encoder, one-dimensional convolution with three different convolution kernel widths {3, 5, 7} is used for processing [14], as shown in formula (9).

$$C_k = \text{Conv1D}_{\text{kernel}=k}(Z^{(L)}), \qquad k \in \{3, 5, 7\} \tag{9}$$

In formula (9), k represents the width of the convolution kernel, and Conv1D represents 1D convolution. After convolution, $\{C_3, C_5, C_7\}$ is concatenated in the feature dimension and then reduced to d through 1×1 convolution to fuse the local information of different receptive fields, as shown in formula (10).

$$Z_{\text{enc}} = \text{Conv1D}_{1\times1}([C_3; C_5; C_7]) \tag{10}$$

In formula (10), Z_{enc} represents the final output of the Encoder.

Decoder prediction.

In the decoder prediction, the steps are as follows:

(1) After embedding the first O outputs predicted in the previous step in a sliding window manner, add position and time encoding in the same way as above to form $D^{(o)}$.

(2) In the l-th layer of the Decoder, first self-attention is performed on $D^{(l-1)}$, and then cross-attention is performed with the Encoder output Z_{enc}. The calculation method is the same as the encoder, but the Key/Value uses Z_{enc}.

(3) Each layer uses residual connection and LayerNorm, and the last layer of the Decoder outputs Z_{dec}. After linear mapping, the formula is shown in (11).

$$\widehat{Y} = Z_{\text{dec}} W_p + b_p \tag{11}$$

In formula (11), $\widehat{Y}$ represents the predicted characteristics of aesthetic education effects at o moments in the future.

B. BIGRU MODEL

In the parallel hybrid structure, the BiGRU module is responsible for fine-grained modeling of short-term temporal fluctuations of the input [15,16]. The BiGRU model extracts history→future and future→history information in parallel through bidirectional GRU layers, and applies normalization and Dropout between layers to generate a representation sequence that can reflect local dynamic changes.

For the input at time step t and the hidden state at the previous moment, the calculation formulas for the update gate, reset gate, and candidate hidden state are still as shown in equations (12) to (15) [17].

$$\vec{z}_t^{(l)} = \sigma\left(W_z^{(l)} Z_t^{(l-1)} + U_z^{(l)} \vec{h}_{t-1}^{(l)} + b_z^{(l)}\right) \tag{12}$$

$$\vec{r}_t^{(l)} = \sigma\left(W_r^{(l)} Z_t^{(l-1)} + U_r^{(l)} \vec{h}_{t-1}^{(l)} + b_r^{(l)}\right) \tag{13}$$

$$\tilde{\vec{h}}_t^{(l)} = \tanh\left(W_h^{(l)} Z_t^{(l-1)} + U_h^{(l)}\left(\vec{r}_t^{(l)} \odot \vec{h}_{t-1}^{(l)}\right) + b_h^{(l)}\right) \tag{14}$$

$$\vec{h}_t^{(l)} = \left(1 - \vec{z}_t^{(l)}\right) \odot \vec{h}_{t-1}^{(l)} + \vec{z}_t^{(l)} \odot \tilde{\vec{h}}_t^{(l)} \tag{15}$$

In formulas (12) to (15), σ represents Sigmoid activation, $\tanh$ represents hyperbolic tangent activation, $W_z^{(l)}$, $W_r^{(l)}$, and $W_h^{(l)}$ represent input mapping matrices, $U_z^{(l)}$, $U_r^{(l)}$, and $U_h^{(l)}$ represent hidden state mapping matrices, $b_z^{(l)}$, $b_r^{(l)}$, and $b_h^{(l)}$ represent bias vectors. $\odot$ represents element-wise product. In reverse time, the same calculation is performed on the input sequence in reverse order to obtain $\overleftarrow{h}_t^{(l)}$. Now the forward and reverse hidden states are concatenated in the feature dimension to form a bidirectional representation, as shown in (16).

$$h_t^{(l)} = \text{Concat}\left(\vec{h}_t^{(l)}, \overleftarrow{h}_t^{(l)}\right) \tag{16}$$

In formula (16), $h_t^{(l)}$ represents the hidden state representation after concatenation.

To stabilize training, accelerate convergence, and prevent overfitting, layer normalization and Dropout are applied in sequence after each layer output, as shown in formula (17).

$$\begin{aligned} u_t^{(l)} &= \text{LayerNorm}\left(h_t^{(l)} + Z_t^{(l-1)}\right), \\ Z_t^{(l)} &= \text{Dropout}\left(u_t^{(l)}\right) \end{aligned} \tag{17}$$

In formula (17), $Z_t^{(l)}$ represents the representation after layer normalization and Dropout processing.

Finally, the above-mentioned bidirectional GRU and normalization, Dropout combination are stacked with multiple layers L_{gru}, and the top-level output sequence $Z_t^{(L_{\text{gru}})}$ is taken as the final short-term dynamic feature representation of the BiGRU module.

The overall framework diagram of the intelligent evaluation model of music aesthetic education effect in colleges and universities is shown in Figure 1.

In Figure 1, it can be seen that the framework takes multi-source time series data as input and models them separately through two parallel feature extraction branches. One branch is based on the Informer time series algorithm model, using sparse self-attention mechanism and multi-scale convolution to efficiently capture ultra-long time series trend characteristics. The other branch is based on BiGRU combined with the Global Attention mechanism, focusing on depicting local short-term dynamics and key moments. The two-way features are spliced in the fusion layer and passed to the fully connected layer to complete the regression prediction. The overall structure effectively balances the modeling requirements of long-term time series dependency and short-term fluctuations.

C. GLOBAL ATTENTION MECHANISM

To highlight the key nodes in local short-term fluctuations, this paper applies the Global Attention module. Through the trainable global query vector, the importance weight of

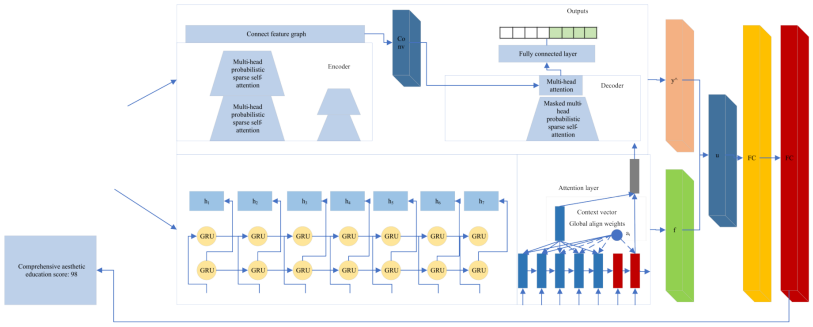

FIGURE 1. Overall framework of the intelligent evaluation model for the effect of music aesthetic education in colleges and universities

each moment in the time series is calculated by the additive attention mechanism, and a fixed-dimensional context representation is generated [18,19].

For each time step, the intermediate representation is calculated, as shown in formula (18).

$$m_t = \tanh(W_z Z_t + W_q q + b) \tag{18}$$

In formula (18), m_t represents the intermediate representation.

It is mapped to the scalar score e_t along the dimension, as shown in (19).

$$e_t = v^\top m_t \tag{19}$$

In formula (19), $v^\top$ represents a row vector.

For e_t, all e_t are passed through the Softmax function to obtain the normalized attention weight, and the calculation formula is shown in formula (20).

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \qquad \sum_{t=1}^{T} \alpha_t = 1, \quad \alpha_t \geq 0 \tag{20}$$

In formula (20), α_t represents the normalized attention weight. The BiGRU output is weighted and summed according to the weight to generate the global context representation, as shown in (21).

$$c = \sum_{t=1}^{T} \alpha_t Z_t \tag{21}$$

In formula (21), c represents the corresponding global context representation.

To stabilize the training, residuals are added to the context vector, and layer normalization and dropout are performed to output the features that are finally output by Global Attention. The calculation is shown in formula (22).

$$\begin{cases} u = \text{LayerNorm}(c + \bar{Z}), \qquad \bar{Z} = \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} Z_t, \\ f = \text{Dropout}(u; p) \end{cases} \tag{22}$$

In formula (22), p represents the probability of zeroing, and f represents the feature output by Global Attention.

The attention visualization diagram is shown in Figure 2.

The Informer self-attention heat map in Figure 2(a) intuitively demonstrates the model's ability to capture global dependencies over very long time series. The horizontal and vertical axes represent the time step indexes of the Key and Query, respectively, and the color depth corresponds to the size of the sparse self-attention weight. As can be seen from Figure 2, the weights of the 10th, 20th, and 30th rows are obviously biased towards the later nodes (such as week 40), indicating that at these query moments, Informer can effectively focus on the key course evaluation and work display weeks, thereby capturing semester cycle fluctuations and long-term trends; while the attention of the middle and non-key rows is relatively scattered, verifying that the ProbSparse Attention mechanism can selectively retain the most informative attention connections, significantly reducing the computational complexity without sacrificing the modeling accuracy of long-term dependencies.

The Global Attention weight bar chart in Figure 2(b) further focuses on the key nodes of the BiGRU output features, with the horizontal axis representing the week and the vertical axis representing the normalized attention weight. It can be seen that the weight has multiple obvious peaks in the 10th, 20th, 30th, and 40th weeks, which is highly consistent with the important events such as classroom performance scoring and evaluation corresponding to each stage evaluation week in the experimental data.

D. FEATURE FUSION AND OUTPUT LAYER DESIGN

Informer output feature aggregation.

This paper performs pooling aggregation on the multi-step outputs of the Informer time series algorithm model to generate a fixed-dimensional representation. The calculation formula is shown in (23).

$$\bar{y} = \frac{1}{O} \sum_{i=1}^{O} \hat{y}^{(i)} \tag{23}$$

In formula (23), O represents the number of future prediction steps, $\hat{y}^{(i)}$ represents the predicted feature vector of the i-th step, and $\bar{y}$ represents the averaged informer feature representation.

Parallel feature concatenation.

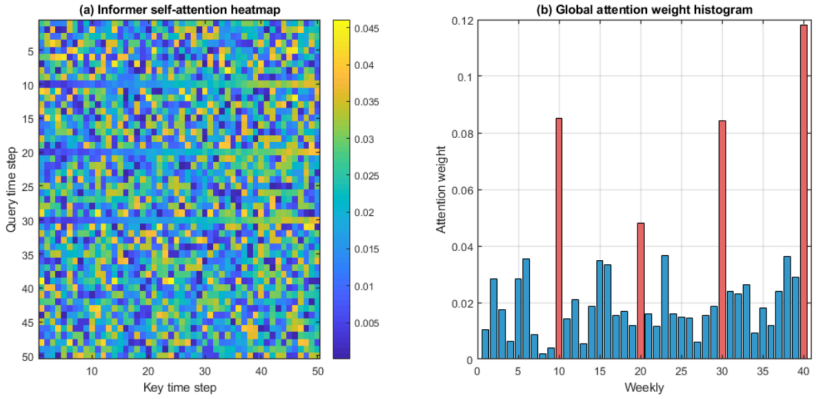

FIGURE 2. (a) Informer self-attention heat map; (b) Global Attention weight bar graph

Now, directly concatenating the average-pooled Informer features and the Global Attention context vector in the feature dimension to obtain the fused vector, as shown in (24).

$$u = \begin{bmatrix} \hat{y} \\ f \end{bmatrix} \tag{24}$$

In formula (24), u represents the overall representation of the fusion of Informer trend features and BiGRU– Attention local features.

Two-layer fully connected network design.

This paper uses a two-layer Fully Connected (FC) network to perform a nonlinear transformation on the fusion vector to enhance the model's expressiveness. The first layer mapping and activation operations are shown in formula (25).

$$h = \mathrm{ReLU}(W_1 u + b_1) \tag{25}$$

In formula (25), W_1 represents the weight matrix of the first layer, b_1 represents the bias vector of the first layer, and h represents the hidden representation of the first layer.

After the first layer mapping and activation, Dropout regularization is applied and the output is $\widetilde{h}$. The second layer mapping and output are shown in formula (26).

$$\widehat{y} = W_2 \widetilde{h} + b_2 \tag{26}$$

In formula (26), W_2 represents the second-layer weight row vector, b_2 represents the second-layer scalar bias, and $\widehat{y}$ represents the regression output result, which corresponds to the predicted music aesthetic education effect score.

E. TRAINING AND OPTIMIZATION OF PARALLEL HYBRID MODELS

Definition of loss function.

The expression of the loss function $\mathcal{L}(\theta)$ is shown in formula (27).

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \widehat{y}_i(\theta) \right)^2 + \frac{\lambda}{2} \|\theta\|_2^2 \tag{27}$$

In formula (27), $\widehat{y}_i$ represents the predicted value of the fusion layer output, y_i represents the actual aesthetic education score, and θ represents the total parameter vector of the model.

Optimization algorithm and learning rate scheduling.

The experiment uses the AdamW optimizer [20-22] to iteratively update the parameter θ, combined with cosine annealing learning rate scheduling to avoid falling into the local optimum. The formula for AdamW parameter update is shown in (28).

$$\theta_t = \theta_{t-1} - \alpha_t \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} - \alpha_t \lambda \theta_{t-1} \tag{28}$$

In formula (28), α_t represents the learning rate, ϵ represents the numerical stability constant, and $\alpha_t \lambda \theta_{t-1}$ represents the weight decay.

In the t-th round, the formula of the cosine annealing learning rate schedule [23] is shown in (29).

$$\alpha_t = \alpha_{\min} + \frac{1}{2}(\alpha_{\max} - \alpha_{\min}) \left(1 + \cos\left(\frac{\pi t}{T} \right) \right) \tag{29}$$

In formula (29), $\alpha_{\max}$ is set as the initial learning rate, and $\alpha_{\min}$ represents the minimum learning rate. The hyperparameter settings are shown in Table 1.

TABLE 1. Hyperparameters

Parameters	Value
Batch Size	32
Learning Rate (initial)	1×10^{-3}
Learning Rate (minimum)	1×10^{-5}
Number of training rounds (maximum)	100
Dropout rate	0.2
Weight decay coefficient	1×10^{-4}
Gradient clipping threshold	1
Informer encoder layers	3
Informer attention heads	8
Decomposer window size	25
BiGRU layers	2
BiGRU hidden state dimension (unidirectional)	128
Input window length	12
Prediction step length	4

Gradient clipping and parallel synchronization.

The model contains ultra-long sequence attention and a multi-layer RNN (Recurrent Neural Network), and the gradient may explode. The experiment performs global norm clipping on the gradient of the entire model and uses synchronous SGD (Stochastic Gradient Descent) in a multi-GPU (Graphics Processing Unit) environment. Each device calculates the local gradient and summarizes it through All-Reduce, and then clips and updates it to ensure the consistency of model parameters.

IV. EXPERIMENT ON INTELLIGENT EVALUATION OF MUSIC AESTHETIC EDUCATION EFFECT IN COLLEGES AND UNIVERSITIES

A. EXPERIMENTAL DATA

The data used in this study comes from the teaching and evaluation records of undergraduate students majoring in music at a key university's art college in the 2021-2024 academic year. It contains three types of multi-source time series data, a total of 1,984 students, and 40 weeks. The student questionnaire emotional attitude data is collected in the form of online questionnaires 6-7 key weeks per semester, covering 5 dimensions such as learning interest, classroom satisfaction, and self-efficacy. For example, data are collected in weeks 3, 5, 8, 10, 12, 14, and 16 of the 2021-2022 fall semester, and in weeks 3, 6, 9, 12, 15, and 18 of the 2022-2023 spring semester. Class participation and interaction log data automatically record students' "online sign-in times", "number of speeches", "homework submission timeliness", and other indicators in each class based on the teaching management system; they are summarized once a week, with a total of 40 weekly records. Practice works and assessment scores include classroom performance scores, comprehensive assessment scores, and homework work scores, once a week, of which the 10th, 20th, 30th, and 40th weeks correspond to stage assessments.

The training labels in this paper are not derived from a single subjective evaluator, but are constructed through a multi-source, quantifiable composite labeling process. The labeling is mainly composed of (1) blind evaluation of students' final presentations and works by three independent experts with teaching experience based on a unified scoring scale (five-dimensional scoring items: expressiveness, skill, creativity, material application, and overall aesthetics; the expert scores are averaged with a weight of 0.60), (2) standardized practical assessment scores at the end of the semester (min-max normalized with a weight of 0.30), and (3) a weighted linear combination of the sum of students' self-evaluation and peer evaluation at the end of the semester (weight of 0.10) to obtain the final "aesthetic education effect score" as a monitoring signal. To ensure the reliability of the labeling, the experiment calculated the intragroup consistency (ICC(2,1)=0.82) and Cronbach's α=0.86 for the expert scores, and the Pearson correlation coefficient between the composite score and the existing final comprehensive scores of the department was r=0.88 (p<0.001), indicating good construct validity. Furthermore, sensitivity analysis was conducted for different weight configurations (changing the expert scoring weight by $\pm$10%), and the MAE of the model performance fluctuated only slightly ($\pm$0.08), further demonstrating that the composite label used not only overcomes the subjectivity problem of single-point scoring, but also has sufficient stability and interpretability to serve as a supervision target for time series prediction.

At the data aggregation level, the raw information was scaled (min–max normalization) from three sources (expert blind review of five-dimensional sub-items, standardized practice assessment, and self-assessment/ peer assessment) and its composability was tested using exploratory factor analysis (EFA) and principal component analysis (PCA): the first principal component explained approximately 61% of the total variance, indicating the existence of a dominant common factor that can represent the overall aesthetic performance; subsequently, confirmatory factor analysis (CFA) was used to evaluate the fit of the single-factor model (CFI$\approx$0.95, RMSEA$\approx$0.045), supporting the rationality of aggregating several observation dimensions into a single latent variable. Secondly, the design of the aggregation weights was guided by educational theory and experience, with expert scores having the highest weight to reflect professional judgment, followed by assessment scores and students' subjective feelings.

B. DATA PREPROCESSING

For the original multi-source time series data, missing values and outliers are processed first. For missing items in the student questionnaires due to unfilled or system anomalies, the interpolation algorithm based on K-nearest neighbor is used to fill them; at the same time, the IQR (interquartile range) method is used to identify outliers for all numerical features. If a sample feature value exceeds the interval [1-1.5IQR, 3+1.5IQR], the value is truncated to the nearest boundary point to ensure that the time series is smooth and without extreme interference.

After filling and truncation, each feature is normalized and time-series segmented, and continuous variables such as class participation and assessment scores are mapped to [0,1] using Min–Max normalization, as shown in formula (30).

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{30}$$

In formula (30), $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature in the training set, respectively.

The discrete option dimensions of the questionnaire are mapped to atomic features through one-hot encoding. Finally, the continuous time series is segmented into sliding windows with a fixed window length of 12 weeks and a step length of 1 week to generate input-label pairs, so that the model can learn local short-term dynamics and maintain sequence alignment.

This study strictly adheres to ethical norms and the principle of minimum necessity in educational data research. All raw data underwent de-identification and anonymization before entering the modeling process, including removing any information that could directly or indirectly identify individuals, such as student names, student IDs, account numbers, and precise timestamps, and replacing individual identifiers with random codes. Classroom interaction data was used only in the form of statistical frequency and time aggregation features, without involving specific speech content or behavioral text, ensuring that the data only reflects the intensity of the learning process rather than individual behavioral details. The collection and use of data were approved by the teaching management departments of the respective institutions, and informed consent was obtained from students based on teaching improvement and research purposes. The data was not used for individual evaluation, rewards and punishments, or administrative decisions. To reduce the risk of potential algorithmic bias, this paper did not introduce sensitive attributes such as gender, ethnicity, and family background during the modeling process. At the same time, time-series cross-validation, multi-source feature fusion, and regularization constraints were used to reduce the model's overfitting to a single data source or a specific student group. In the experimental analysis phase, the distribution of model error across different time periods and data subsets did not show a systematic shift, indicating that the model focuses more on the dynamic changes in the learning process rather than inherent individual characteristics. It is important to emphasize that the model in this paper is positioned as a trend assessment and teaching feedback support tool at the group level, rather than an automated value judgment or labeling of individual students. Its design intention and application boundaries are in line with the ethical principle of current educational artificial intelligence: "to assist decision-making rather than replace evaluation".

C. EVALUATION METRICS

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{31}$$

Here y_i represents the true value, and $\widehat{y}_i$ represents the predicted value.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{32}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{33}$$

D. EXPERIMENTAL PLAN

To fully verify the evaluation performance of the parallel hybrid model (Informer + BiGRU-Global Attention), the experimental scheme is designed with different modeling strategies as comparison baselines, including FFA-BiGRU based on attention mechanism and gated recurrent unit, LSTNet (Long- and Short-term Time-series Network) combining convolution and temporal recurrence, Transformer with classic self-attention structure, CNN-LSTM integrating convolution and LSTM, and Informer-BiLSTM (Informer-Bidirectional Long Short-Term Memory) based on long sequence sparse attention and recursive mixture. All models are run under the same training set and test set division to ensure the comparability of the results. The experiment uses ten-fold cross validation to divide the data and keeps the student data divided in chronological order to avoid information leakage.

In the experiment, all models use the same input features, the same sliding window length (12 weeks) and prediction step (4 weeks), the optimizer used AdamW, the initial learning rate is set to 0.001, the maximum number of learning rounds is 100, and the early stopping strategy is used on the validation set. All experiments use MAE, MAPE, RMSE and other indicators to comprehensively evaluate the prediction accuracy, supplemented by the contribution of multi-source data, and the visualization of the attention mechanism to further analyze the model differences.

V. MUSIC AESTHETIC EDUCATION EFFECT EVALUATION RESULTS

A. OVERALL PREDICTION ACCURACY ANALYSIS

The overall prediction accuracy evaluation uses three indicators, MAE, RMSE, and MAPE, to measure the regression performance of each model on the test set. The overall prediction accuracy of the six comparison models is shown in Figure 3.

In Figure 3, it can be see that the Informer+BiGRU-Global Attention model performs best in terms of MAE and RMSE, with scores of 2.3 and 3.0, respectively, which is significantly better than other comparison models. For the Informer-BiLSTM model, its MAE is 3.0 and RMSE is 3.8, which is better than the traditional Transformer (MAE 3.5, RMSE 4.2) and CNN-LSTM (MAE 3.9, RMSE 4.7). However, FFA-BiGRU and LSTNet perform relatively poorly, with MAE exceeding 4 and RMSE exceeding 5, indicating large errors.

This paper predicts a continuous scale for "Music Aesthetics Education Effectiveness Score" ranging from 0 to 100, consistent with the percentage-based grading commonly used in higher education teaching management. 60 points represents the basic passing grade, 70-80 points indicate a good level, and above 80 points indicate high aesthetic literacy and learning outcomes. Based on this scoring range, the model's MAE on the test set is 2.30, equivalent to an average prediction error of only 2.3% of the full score. This is significantly smaller than the width of a typical grade range (usually 5-10 points), meaning that the model's predictions will generally not lead to substantial changes in student evaluation levels or teaching decisions in actual teaching evaluations. From an educational application perspective, this error level is close to the subjective

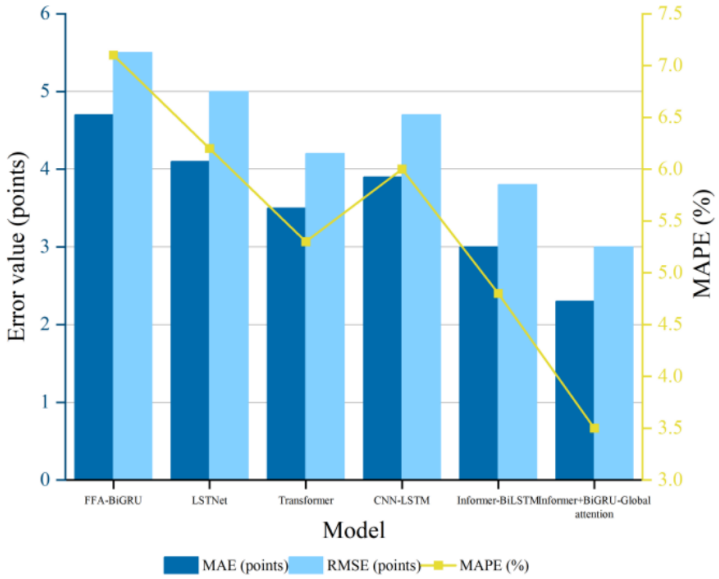

FIGURE 3. Overall prediction accuracy

fluctuation range between expert ratings (ICC=0.82 in the consensus analysis of blind expert reviews in this paper), thus possessing clear practical acceptability and interpretative significance. In contrast, when the MAE of a simple baseline model exceeds 5 points, it often corresponds to a cross-grade deviation between "passable" and "good," making it difficult to directly use for teaching diagnosis and intervention decisions. In summary, under a clearly defined 0–100 score evaluation system, the error index reported in this paper is not only statistically significant, but also has clear, understandable, and operable practical value in educational evaluation practice. To verify whether the overall prediction accuracy difference is statistically significant, this paper conducts a paired sample t-test on the test set MAE error series of the Informer+BiGRU-Global Attention model and the other five baseline models. The steps are as follows:

(1) For each pair of models, the absolute error sequence of N samples in the test set is calculated and the difference sequence is constructed.

(2) The sample mean and sample standard deviation of the sequence are calculated, and the t statistic is calculated to compare the two-tailed p value under the degree of freedom N-1.

(3) The significance level is set to α=0.05. When p<α, the difference in prediction accuracy between the two models is considered statistically significant.

The t-test results are shown in Table 2.

TABLE 2. T-test results

Baseline Model	t Statistic	p Value	Significance ($\alpha = 0.05$)
FFA-BiGRU	8.52	<0.001	Significant
LSTNet	7.34	<0.001	Significant
Transformer	6.21	<0.001	Significant
CNN-LSTM	7.01	<0.001	Significant
Informer-BiLSTM	5.47	<0.001	Significant

In Table 2, the Informer+BiGRU-Global Attention model shows significant differences in overall prediction accuracy compared to the other five baseline models. In the paired sample t-test of the test set MAE, the t-statistic of the Informer+BiGRU-Global Attention model with FFA-BiGRU is as high as 8.52 ($p < 0.001$), the t-value with LSTNet is 7.34 ($p < 0.001$), the t-value with Transformer is 6.21 ($p < 0.001$), the t-value with CNN- LSTM is 7.01 ($p < 0.001$), and the t-value with Informer-BiLSTM is 5.47 ($p < 0.001$). All p values are much smaller than the significance level $\alpha = 0.05$, indicating that the Informer+BiGRU-Global Attention model has a statistically significant advantage over the comparison model in terms of MAE performance on the test set. This shows that the parallel hybrid model in this paper not only outperforms various baseline models in terms of mean error but also has significantly improved stability and generalization ability at the sample level.

B. ANALYSIS OF THE EFFECT OF LONG-TERM DEPENDENCE TREND MODELING

This paper uses time series comparison curves to intuitively demonstrate the ability of each model to capture the trend of aesthetic education scores under 20-week and 30-week predictions. The predicted values of six models, FFA-BiGRU,

LSTNet, Transformer, CNN-LSTM, Informer-BiLSTM, and Informer + BiGRU–Global Attention, are compared with the actual scores. The results are shown in Figure 4 (a) and Figure 4 (b).

In the 20-week forecast in Figure 4(a), the Informer+ BiGRU–Global Attention curve almost completely overlaps with the true value. For example, the true value in the 20th week is 72.66 points, and the predicted value is 72.71 points, with an error of only 0.05 points; the predictions of FFA-BiGRU (73.36 points) (error 0.70 points), Transformer (72.37 points) (error 0.29 points), and Informer-BiLSTM (72.15 points) (error 0.51 points) are all more deviated. The Informer+BiGRU–Global Attention model can also accurately follow the trend turning point (jumping from 69.98 to 70.69 in the 8th week) (predicting 70.71, a difference of 0.02 points), which fully demonstrates the high-precision fitting ability of the hybrid model for macro trends and key nodes within the cycle.

In the 30-week prediction in Figure 4(b), the Informer+BiGRU–Global Attention model predicts 73.43 points in the 30th week, the actual value is 73.31 points, and the error is 0.12 points; while FFA-BiGRU is 74.09 points (error 0.78 points), Transformer is 73.78 points (error 0.47 points), and CNN-LSTM is 73.77 points (error 0.46 points). In particular, during the two significant fluctuations in the 14th week (the actual value is 71.72 points) and the 27th week (the actual value is 73.47 points), the Informer+BiGRU–Global Attention hybrid model has only 0.10 points and 0.09 points of deviation, respectively, which is significantly better than other models.

C. ANALYSIS OF SENSITIVITY OF LOCAL SHORT-TERM DYNAMIC CHANGE RESPONSE

To evaluate the responsiveness of each model to short-term events such as sudden teaching reform, centralized assessment week and intensive training week, the experiment is tested in weeks 10-12 (teaching reform), 20-22 (centralized assessment) and 30-32 (intensive training) during the 40-week semester. The sensitivity curve of the prediction error of each model with the change of week is shown in Figure 5. In Figure 5, the yellow area represents teaching reform, the green area represents centralized assessment, and the blue area represents intensive training.

In Figure 5, the error peaks of each model increased significantly during the 10th to 12th week (teaching reform period). Taking the 10th week as an example, the error of FFA-BiGRU is as high as 0.97 points, LSTNet 1.09 points, Transformer only 0.12 points, CNN-LSTM 0.43 points, Informer-BiLSTM 0.01 points, and Informer+BiGRU–Global Attention only 0.00 points; the 11th and 12th weeks also show similar trends, with the errors of the Informer+BiGRU–Global Attention hybrid model being 0.28 points and 0.53 points, respectively. During the concentrated evaluation period (weeks 20–22), the error of the Informer+BiGRU–Global Attention hybrid model is 0.23 points in week 20, while FFA-BiGRU reaches 2.12 points, LSTNet 0.16 points, Transformer 0.84 points, CNN-LSTM 0.86 points, and Informer-BiLSTM 0.29 points; the errors of the hybrid model in weeks 21–22 are 0.08/0.16 points, which are significantly lower than the comparison models of 0.28/0.98 points (FFA-BiGRU) and 0.37/1.31 points (LSTNet). During the training period (weeks 30–32), the hybrid model error remains in the 0.04–0.55 range, while FFA-BiGRU fluctuates more (0.27 in week 30, 1.52 in week 31, and 0.55 in week 32). Overall, Informer+BiGRU–Global Attention exhibits the smallest error fluctuations in the three short-term event windows, demonstrating its high sensitivity and robustness to bursty and focused tasks.

D. IMPACT OF DIFFERENT DATA SOURCES ON EVALUATION RESULTS

To evaluate the contribution of each data source to the model's performance, the model is trained and tested using a single data source and its various combinations, and the MAE, RMSE, and Pearson correlation coefficient are calculated. At the same time, to quantitatively evaluate the relative impact of each data source on the prediction results of the parallel hybrid model, this paper calculated the average contribution ratio of the three feature groups of emotional attitude questionnaire, classroom participation log and practical evaluation score based on the SHAP (SHapley Additive exPlanations) method. The results of the impact of different data sources on the evaluation effect are shown in Figure 6. In Figure 6, the emotional attitude questionnaire is represented by TS1, the classroom participation log is represented by TS2, and the practical evaluation score is represented by TS3.

In Figure 6 (a), the MAE and RMSE of the model under each single data source vary significantly with the data type. When only the emotional attitude questionnaire (TS1) is used, the MAE is 5.1 points and the RMSE is 6.3 points; when the class participation log (TS2) is used, it drops to 4.4 points and 5.5 points respectively; when the practical assessment score (TS3) is used, it further drops to 3.9 points and 4.6 points. This shows that, compared with subjective questionnaires, objective scoring and participation records can more accurately reflect student dynamics. After combining TS1 and TS2, MAE/RMSE continue to drop to 3.6/4.2 points; TS2+TS3 is 3.3/4.0 points; TS1+TS3 is 3.1/3.8 points; when the three are used together, MAE/RMSE finally drop to 2.3/3.0 points, and the accuracy is significantly improved, indicating that multi-source fusion can complement each other. In the Pearson correlation coefficient comparison in Figure 6(b), the correlation between the model and the true score under single source is TS1: 0.64, TS2: 0.71, and TS3: 0.76, respectively, reflecting that the contribution of practical assessment scores to the evaluation results is higher than that of classroom participation and subjective attitude. After the dual-source combination, the correlation coefficients are increased to TS1+TS2: 0.80, TS2+TS3: 0.83, and TS1+TS3: 0.85; when the three-source data are used simultaneously,

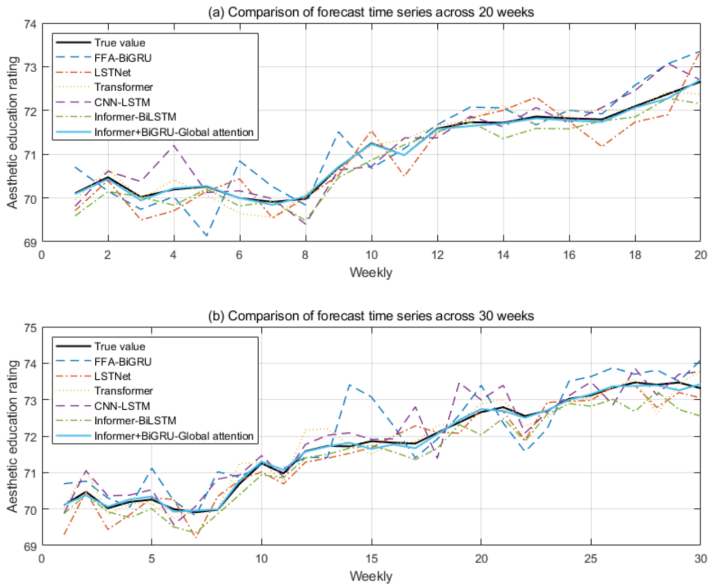

FIGURE 4. Time series comparison curve of actual and predicted values of aesthetic education scores of different models: (a) Time series comparison curve of actual and predicted values of aesthetic education scores of different models across 20 weeks; (b) Time series comparison curve of actual and predicted values of aesthetic education scores of different models across 30 weeks

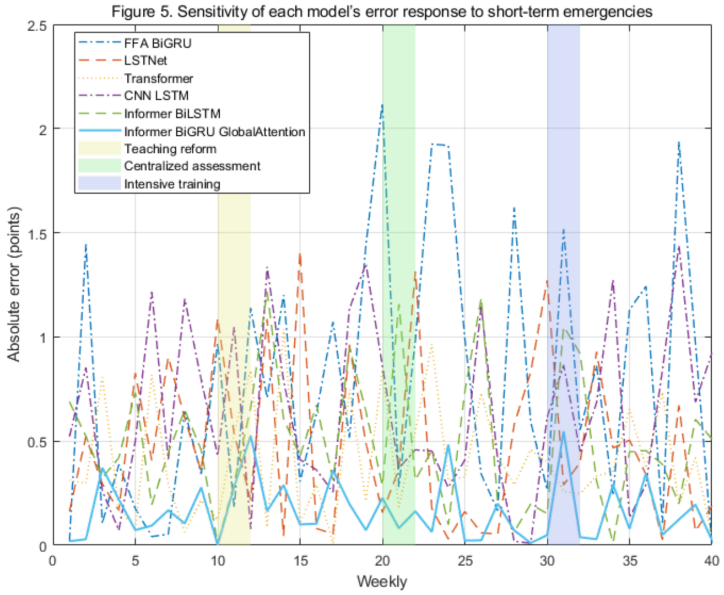

FIGURE 5. Sensitivity curves of prediction errors of each model as the week changes

the correlation coefficient reaches 0.91, which is close to a perfect linear relationship, further proving that multi-source information fusion can greatly improve the discrimination consistency and correlation of the model.

The average contribution ratio of SHAP in Figure 6(c) quantifies the impact of each data source on the prediction of the parallel hybrid model from another perspective. Class participation log (TS2) accounts for 37.5%, practical assessment score (TS3) accounts for 34.1%, and emotional attitude questionnaire (TS1) contributes 28.4%. Practical assessment is the most intuitive quantitative standard. Class participation log plays a greater role in the overall prediction by virtue of its high-frequency capture of the learning process. The subjective questionnaire is slightly inferior, but still indispensable.

E. ROBUSTNESS TEST

To verify the robustness of the parallel hybrid model under data disturbance conditions, this paper adds Gaussian noise of different intensities (standard deviation σ) to the input

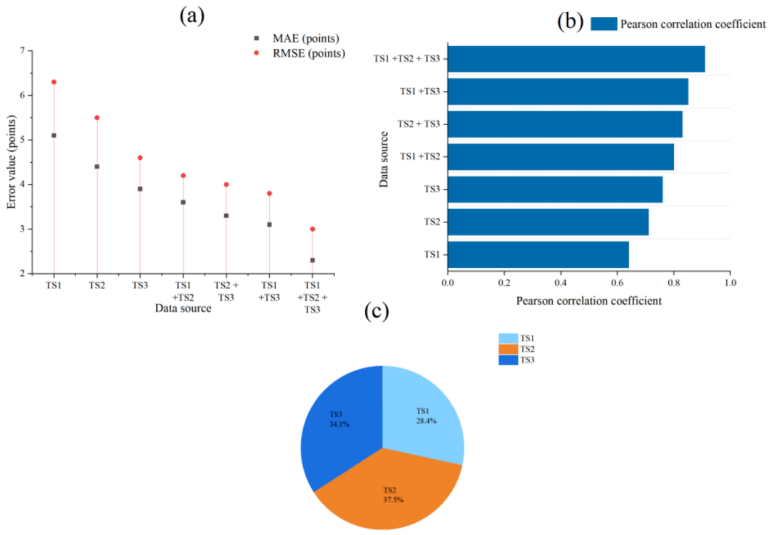

FIGURE 6. The impact of different data sources on evaluation results: (a) MAE and RMSE values of different data sources; (b) Pearson correlation coefficient of different data sources; (c) Contribution ratio of different data sources to model prediction

data, applies extreme outliers of different proportions, and evaluates the changes in the model in terms of MAE, RMSE, and MAPE indicators. The robustness test results are shown in Figure 7.

In the Gaussian noise test in Figure 7 (a), as the noise intensity increases from 0.01 to 0.20, the model MAE gradually increases from about 2.3 points to 3.8 points, while the RMSE increases from 3.0 points to 5.2 points. When σ is 0.10, the MAE increases to nearly 2.9 points, the RMSE is close to 4.0 points, and the performance degradation caused by noise remains in the medium range. In the extreme noise environment where σ reaches 0.20, the model still controls the MAE within 3.8 points, indicating that the parallel hybrid structure has a certain resistance to random small disturbances.

In the extreme outlier test in Figure 7 (b), when the proportion of outliers increases from 0 to 9%, the model error deteriorates more significantly. Only 1% of the extreme points cause the MAE to jump from about 2.3 points to 2.7 points, and the RMSE increases to about 3.4 points; when it reaches 5%, the MAE reaches 3.6 points, and the RMSE approaches 4.8 points; when 9% of the data is extremely abnormal, the MAE climbs to about 4.5 points, and the RMSE reaches 6.2 points, indicating that compared with uniformly distributed Gaussian noise, a small number of but large-amplitude outliers have a more severe impact on the regression results. In summary, the model in this paper has a certain degree of robustness to outliers.

F. COMPARISON WITH SIMPLE BASELINE MODELS

To demonstrate the rationale for using the complex and high-overhead Informer+BiGRU architecture, comparative experiments were conducted with three classic baselines (moving average, ARIMA, and a simple LSTM).

The experimental steps were as follows: Under the same data preprocessing, sliding window (12 weeks input $\rightarrow$ 4 weeks prediction), 10-fold time series cross-validation partitioning, and the same evaluation metrics (MAE, RMSE, MAPE) as described above, the following models were implemented and trained: (1) Simple moving average (window = 4); (2) ARIMA based on AIC automatic selection (p,d,q) (modeling each time series separately and merging predictions by week); (3) Single-layer LSTM (unidirectional, 128 hidden units, the same optimizer and early stopping strategy as in this paper). Training and inference of the deep learning models were measured on an NVIDIA V100 GPU (batch size = 32); ARIMA and moving average were measured on an Intel Xeon CPU on the same server to reflect common deployment environments. All results are the mean of 10-fold cross-validation, inference time is the average of 1000 iterations on the test set, and the number of parameters is an approximate count. The comparison with the simple baseline model is shown in Table 3.

In Table 3, the simple statistical baseline model showed significant limitations in predicting the effectiveness of music aesthetics education in higher education institutions. The MAE scores for moving average and ARIMA were 6.18 and 5.42, respectively, with MAPEs as high as 10.47% and 9.21%. Although it had extremely low inference overhead ($\leq$1.6 ms), it struggled to characterize the multi-scale and nonlinear evolutionary features inherent in the teaching evaluation sequence. After introducing deep learning, the MAE of a single-layer LSTM decreased to 3.62 and the RMSE to 4.58, but its inference time reached 14.8 ms, and its accuracy still significantly lagged behind self-attention-based models. The Informer-BiLSTM baseline further reduced the MAE and RMSE to approximately 3.0 and 3.8, respectively, validating the advantages of sparse self-attention in long

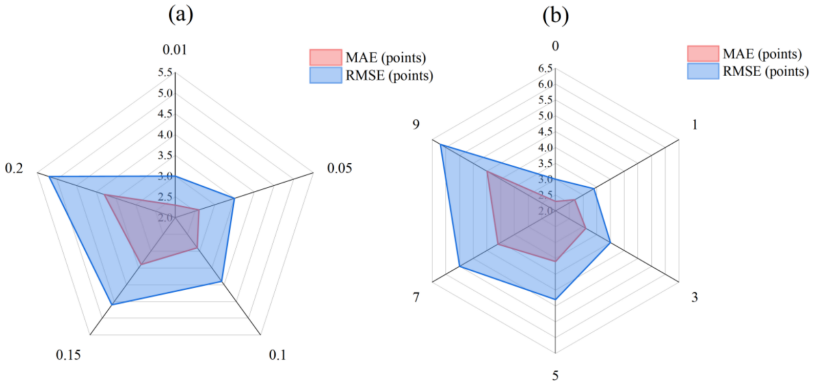

FIGURE 7. Robustness test results: (a) Results under Gaussian noise of different intensities; (b) Results with extreme outliers of different proportions

TABLE 3. Comparison with the Simple Baseline Model

Model (Configuration)	MAE (minutes)	RMSE (minutes)	MAPE (%)	Inference Time (ms)	Number of Parameters (k)
Moving Average (window=4, CPU)	6.18	7.93	10.47	0.4 (CPU)	∼0.1
ARIMA (auto-ARIMA per series, CPU)	5.42	6.88	9.21	1.6 (CPU)	∼0.1
Simple LSTM (1 layer, 128 units, GPU)	3.62	4.58	5.81	14.8 (GPU)	512
Informer-BiLSTM (baseline, GPU)	3	3.8	4.8	9.6 (GPU)	1250
Informer + BiGRU–Global Attention (This work)	2.30	3.00	3.50	8.4 (GPU)	1800

sequence modeling. In comparison, our Informer + BiGRU–Global Attention model achieves scores of 2.30, 3.00, and 3.50% on MAE, RMSE, and MAPE, respectively, showing significant error improvement compared to simple LSTM and Informer-BiLSTM. Furthermore, its inference time is only 8.4 ms, lower than LSTM and Informer-BiLSTM. These results demonstrate that, within an acceptable parameter size (approximately 1.8M) and real-time inference overhead, introducing parallel BiGRU and a global attention mechanism can achieve a better balance between accuracy and efficiency, proving the rationality and necessity of the complex architecture used in practical applications.

VI. EXPERIMENTAL DISCUSSION

In this paper, the Informer+BiGRU-Global Attention parallel hybrid model is significantly superior to the single structure and other comparison models in terms of overall accuracy (MAE/RMSE), long-term and short-term trend capture, local event response, and robustness. The fundamental reason is that this method has both the ability to model long-term dependencies at the macro-global level and the ability to extract dynamic details at the micro-local level at the algorithm level. The Informer module uses ProbSparse Attention and multiscale decomposition to efficiently capture global trends across semesters and modules at $O(L \log L)$ complexity, effectively reducing gradient attenuation and information loss caused by sequence length; the parallel BiGRU layer uses a bidirectional gating mechanism to synchronously learn short-term interactions and assessment fluctuations in both the history and future directions, ensuring keen capture of key moments such as “sudden teaching reforms” and “centralized assessment weeks”; Global Attention performs trainable global weighting on BiGRU outputs, automatically highlighting the window nodes with the highest contribution, balancing global trends between local anomalies and periodic changes, and significantly reducing prediction errors. Further analysis found that the synergy of each submodule of the model is also an important reason for the performance improvement. The parallel rather than serial architecture avoids the cumulative misleading of a single path. Even if one path (such as long-term trends) is slightly biased, the other side (such as short-term dynamics) can still compensate for the final output through the attention mechanism; and sparse self-attention reduces the computing and memory pressure, making the training and reasoning process more stable, and more robust under noise and outlier disturbances; the fully connected fusion layer and regularization strategies such as Dropout and LayerNorm jointly ensure the seamless connection of cross-module features and the stable convergence of the training process, avoiding overfitting and gradient problems; the overall parameter scale of the model is moderate, with both high expressiveness and computing efficiency, and it also shows significant advantages in throughput and MFLOPs, providing a solid foundation for actual deployment. This paper uses data from a single source, and the scope and extrapolation boundaries of the research conclusions are discussed explicitly here. It is important to emphasize that the Informer+BiGRU–Global Attention parallel hybrid model proposed in this paper does

not rely on the specific grading rules of a particular institution or course. Instead, it is designed to address the structural problem of "multi-source heterogeneous data, long-term evolution, and overlapping of phased events" prevalent in higher education music aesthetics. Its methodological contribution has cross-scenario significance. However, due to limitations in data acquisition, the current experimental data only comes from 40 weeks of records from a single university, a specific major, and a unified evaluation system. This inevitably embeds the teaching organization methods and evaluation culture of that institution into the model's numerical distribution, weight configuration, and feature importance ranking. Therefore, this paper does not advocate for "zero-cost direct transfer" of the model to other institutions. A more reasonable approach is to adapt it to different institutions, courses, or art categories by relabeling the weights, performing minor retraining, or fine-tuning the parameters, while maintaining the overall model architecture (long-sequence sparse attention + short-term dynamic modeling + global attention fusion). In fact, this paper has shown in the data source contribution analysis and weight sensitivity experiments that the model has good stability with reasonable weight adjustments (MAE fluctuation $\leq$ 0.08), which to some extent supports its feasibility for cross-scenario transfer. Future work will introduce data from multiple institutions, courses, and evaluation systems for joint modeling or transfer learning verification, in order to further systematically evaluate the model's generalization ability and applicability in a wider range of higher art education scenarios.

VII. CONCLUSION

This paper adopts an intelligent evaluation model based on the parallel hybrid architecture of Informer + BiGRU-Global Attention, designed to effectively assess aesthetic outcomes and the technical skill development crucial innovation. In the dual-path modeling of trend and local dynamic features, the sparse self-attention mechanism and multi-scale decomposition of Informer are first used to efficiently capture the global trend across 40 weeks of long time series, and then the BiGRU-Global Attention branch is used to focus on key short-term fluctuations, and finally the fusion output is used to predict the aesthetic education score. Experimental results show that this method is significantly better than traditional Transformer, LSTM and other baseline models in terms of MAE (2.3 points), RMSE (3.0 points), MAPE (3.5%), and the inference time is only 8.4 ms, which meets the real-time requirements. This paper has made some achievements, but the model still has room for improvement in handling abnormal missing data and multi-school heterogeneous data adaptability, and the number of parameters can be further optimized. In the future, we will explore methods for enhancing generalization and interpretability on cross-regional diversified datasets, further exploring lightweight strategies while maintaining accuracy to support a wider range of application scenarios, particularly in assessing specialized creativity and technical execution within smart and material design.

AUTHOR CONTRIBUTIONS

Ying Qi designed, collected and analyzed the data, and drafted the manuscript. Ying Qi conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Ying Qi participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, Research on the Practice and Innovation of TPACK Model in the Ideological and Political Teaching of 'Music Skills I (Music Theory and Solfeggio)' from the Perspective of Digital Transformation.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] D. Hu, "Research on the Impact of Music Aesthetic Education on the Cultivation of Students' Aesthetic Appreciation," Journal of Art, Culture and Philosophical Studies, vol. 1, no. 1, pp. 1-7, 2024, doi: 10.70767/jacps.v1i1.17.

[2] P. Xiao, "Research on the teaching reform of college music aesthetic education from the perspective of core accomplishment," Curriculum and Teaching Methodology, vol. 6, no. 18, pp. 27-34, 2023, doi: 10.23977/curtm.2023.061805.

[3] J. Zheng, Y. Zhang, and S. Zhang, "Audio-visual aesthetic teaching methods in college students' vocal music teaching by deep learning," Scientific Reports, vol. 14, no. 1, pp. 1-18, 2024, doi: 10.1038/s41598-024-80640-7.

[4] C. Yi, "Innovation and Discrete Dynamic Modeling of College Music Teaching Model Based on Multiple Intelligences Theory," Journal of Sensors, vol. 2022, no. 1, pp. 1-8, 2022, doi: 10.1155/2022/8613485.

[5] Y. Li and R. Sun, "Innovations of music and aesthetic education courses using intelligent technologies," Education and Information Technologies, vol. 28, no. 10, pp. 13665-13688, 2023, doi: 10.1007/s10639-023-11624-9.

[6] P. Li and B. Wang, "Artificial intelligence in music education," International Journal of Human–Computer Interaction, vol. 40, no. 16, pp. 4183-4192, 2024, doi: 10.1080/10447318.2023.2209984.

[7] S. Kusumawardani S and I. Alfarozi S A, "Transformer encoder model for sequential prediction of student performance based on their log activities," Ieee Access, vol. 11, no. 1, pp. 18960-18971, 2023, doi: 10.1109/ACCESS.2023.3246122.

[8] H. Wang, "Transformer-based deep learning for adaptive pedagogy under uncertain student preferences," Scientific Reports, vol. 15, no. 1, pp. 1-23, 2025, doi: 10.1038/s41598-025-25996-0.

[9] S. Ma and R. Zhou, "Violin Music Emotion Recognition with Fusion of CNN–BiGRU and Attention Mechanism," Information, vol. 15, no. 4, pp. 224-239, 2024, doi: 10.3390/info15040224.

[10] H. Sun, X. Wang, Y. Wang, and P. Lu, "Music informer as an efficient model for music generation," Scientific Reports, vol. 15, no. 1, pp. 1-14, 2025, doi: 10.1038/s41598-025-02792-4.

[11] Q. Zhu, J. Han, K. Chai, and C. Zhao, "Time series analysis based on informer algorithms: A survey," Symmetry, vol. 15, no. 4, pp. 951-995, 2023, doi: 10.3390/sym15040951.

[12] X. Wang, M. Xia, and W. Deng, "MSRN-informer: Time series prediction model based on multi-scale residual network," IEEE Access, vol. 11, no. 1, pp. 65059-65065, 2023, doi: 10.1109/ACCESS.2023.3289824.

[13] B. Liu, Z. Li, Z. Li, and C. Chen, "CL-Informer: Long time series prediction model based on continuous wavelet transform," PloS one, vol. 19, no. 9, pp. 1-18, 2024, doi: 10.1371/journal.pone.0303990.

[14] Y. Liusong and D. Hui, "Voice quality evaluation of singing art based on 1DCNN model," Mathematical Problems in Engineering, vol. 2022, no. 1, pp. 1-9, 2022, doi: 10.1155/2022/2074844.

[15] S. Ghosh and F. Riad O, "Attention-based cnn-bigru for Bengali music emotion classification," International Journal of Computer Science & Network Security, vol. 23, no. 9, pp. 47-54, 2023, doi: 10.22937/IJCSNS.2023.23.9.6.

[16] W. Chaoguo, Z. Liang, and Y. Wei, "Relation Extraction Based on BERT and BGRU in the Chinese Music Scene," Procedia Computer Science, vol. 225, no. 1, pp. 2429-2438, 2023, doi: 10.1016/j.procs.2023.10.234.

[17] M. Ashraf, F. Abid, U. Din I, J. Rasheed, M. Yesiltepe, F. Yeo S, et al., "A hybrid cnn and rnn variant model for music classification," Applied Sciences, vol. 13, no. 3, pp. 1476-1486, 2023, doi: 10.3390/app13031476.

[18] N. Le, K. Nguyen, A. Nguyen, and B. Le, "Global-local attention for emotion recognition," Neural Computing and Applications, vol. 34, no. 24, pp. 21625-21639, 2022, doi: 10.1007/s00521-021-06778-x.

[19] Y. Cheng, K. Qian, and F. Min, "Global and local attention-based multi-label learning with missing labels," Information Sciences, vol. 594, no. 1, pp. 20-42, 2022, doi: 10.1016/j.ins.2022.02.022.

[20] S. Sun and H. Gao, "Meta-AdaM: An meta-learned adaptive optimizer with momentum for few-shot learning," Advances in Neural Information Processing Systems, vol. 36, no. 1, pp. 65441-65455. https://doi.org/10.52202/075280-2855.

[21] M. Reyad, M. Sarhan A, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," Neural Computing and Applications, vol. 35, no. 23, pp. 17095-17112, 2023, doi: 10.1007/s00521-023-08568-z.

[22] C. Zhang, Y. Shao, H. Sun, L. Xing, Q. Zhao, and L. Zhang, "The WuC-Adam algorithm based on joint improvement of Warmup and cosine annealing algorithms," Math. Biosci. Eng, vol. 21, no. 1, pp. 1270-1285, 2024, doi: 10.3934/mbe.2024054.

[23] V. Johnson O, C. Xinying, W. Khaw K, and H. Lee M, "ps-CALR: periodic-shift cosine annealing learning rate for deep neural networks," IEEE Access, vol. 11, no. 1, pp. 139171-139186, 2023, doi: 10.1109/ACCESS.2023.3340719.