Journal of Network Intelligence ©2026 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **11**, Number **1**, February **2026**

Research on Music Intelligent Annotation Technology Based on Optimized Recurrent Neural Network

Lei Gao[1,*]

[1]Department of Music,
Nanchang Institute of Technology, Nanchang 330000, P. R. China
18879111970@163.com

Yu-Han Wang[2]

[2]International College,
Krirk University, Bangkok 10220, Thailand
zitong0825@gmail.com

*Corresponding author: Lei Gao

Received July 10, 2024, revised October 27, 2024, accepted January 20, 2025.

Abstract. *As the music social media rapidly developing, online music resources are increasing rapidly. Rich music annotation information becomes an important part of online music services as an effective means to organize massive music data. However, there is a large amount of noise in the real music annotation dataset, which leads to poor intelligent annotation. Relied on this, this article investigates the music intelligent annotation technique based on optimized Recurrent Neural Network (RNN). Intending to the issue that traditional RNN is easy to lead to the loss of effective information, the interpolation method is used to optimize the RNN, so as to achieve the accurate estimation of missing values. Secondly, the ERB spectrum and LDA model are adopted to extract features from audio signals and semantic features, respectively, and the attention mechanism is adopted to aggregate the two kinds of features as an enhanced feature representation. Since the feature sequence of music is highly susceptible to noise, this article transforms the aggregated feature vector sequence into a 3D tensor, which is used as the input to the ERNN, and fuses the tensor interpolation results with the prediction outcome of the ERNN to output the final predicted music annotation sequence, avoiding the accumulation of prediction errors due to successive misses. The experimental results imply that the suggested technique is more effective in intelligent annotation, with an average F1 value of 0.8469.*

Keywords: recurrent neural network; music intelligent annotation; LDA model; interpolation algorithm; attention mechanism

1. **Introduction.** Recently, as the digital music technology developing, the music carrier has changed, from physical tapes and records to digital audio files and saved in computer databases. The emergence of digital music has effectively broadened the ways in which music can be distributed, allowing it to be spread wider and faster [1]. Against this background, music streaming providers around the world are continuously developing and expanding their products and services to provide consumers with richer music services. In the face of increasingly rich music application scenarios, traditional organization methods based on music metadata, such as music annotation, can no longer meet all the demands [2]. As a way to enrich music information, music annotation can describe the semantic content of music by generating tags, and it can effectively organize music resources to

achieve fast retrieval, personalized recommendation and other functions. At present, manual annotation and social annotation are still commonly used music annotation methods, which not only face cost problems but also have variable quality of annotation results [3, 4]. Thus, improving the performance of music intelligent annotation system is an urgent problem to be solved nowadays.

1.1. **Related work.** Traditional music annotation mainly includes pitch estimation, rhythm recognition, etc. Bello et al. [5] offered a high-precision music detection algorithm based on the principle of speech signal generation, combining linear predictive coding and average amplitude difference function method. In addition, rhythm is also an important part of music, Kuo and Chuang [6] proposed an algorithm to annotate music rhythm, which can annotate specific types of note structures, and Muller et al. [7] proposed a music annotation method based on tempo and beat analysis of music signals, and Er [8] combined acoustic features and label information, and used LDA for corpus modelling to classify audio. modelling for audio classification. In recent years, song titles and lyrics, as an important part of music content, have also come to the attention of scholars. Rospocher [9] used a bag-of-words model with TF-IDF weighting for music tag annotation based on lyrics. Bolla et al. [10] explored the effect of different word weighting methods on the results of music sentiment classification. Miotto and Lanckriet [11] proposed a sentiment vector space model as a document representation model, using lyrics text to generate sentiment units and using SVM for label classification.

As deep learning has made significant achievements in various fields such as classification, scholars have also tried to apply deep neural networks to the field of intelligent music labelling [12]. Yu et al. [13] proposed a method based on self-organizing neural networks for automatic labelling of multi-timbral harmonies. Singh [14] used deep neural networks to learn the mapping between the statistical information of each frame of a spectrogram or the overall statistical information and the music labels. Costa et al. [15] used Mel's spectral input to obtain time and frequency dimensional features, and then used a Convolutional Neural Network (CNN) to predict music labels, but this method suffers from mislabelling. Furner et al. [16] used a conditional random field to take into account the correlation between textual labels and the sparsity of the music features corresponding to individual labels in the computation of the weights of each feature. Tang et al. [17] input the aggregated audio features and tag vector features into a deep belief network for music tag prediction.

Music has the characteristics of relevance and continuity, which requires serialisation of audio data, and Recurrent Neural Networks (RNNs) can ensure that the information from one moment can be transferred to the next moment, so as to better deal with time-series data [18]. Rajesh and Nalini [19] firstly extracted the global features of the audio and fed them into RNN for label prediction. Lin and Chen [20] fused GRU, a variant of RNN, with CNN, and used the original waveform and Meier spectrogram as model inputs to predict music labels using CNN aggregated feature vectors of music segments. GRU, a variant of RNN, with CNN, fused the original waveform with Mel spectrogram as input to the model, and used CNN to aggregate music segment feature vectors to predict music labels.

1.2. **Contribution.** Deep learning algorithms have achieved better results in music intelligent annotation tasks. However, the real music annotation dataset inevitably has noise, which adversely affects the performance of the model in the music automatic annotation task. Aiming at the above problems, this article suggests a more effective music intelligent annotation technique based on optimized RNN. The main contributions are as follows.

(1) Aiming at the problem that traditional RNN is prone to cause the loss of effective information, the RNN is optimized by using interpolation method, and the accurate estimation of missing values is achieved through an efficient hierarchical learning network.

(2) The audio signal is preprocessed and feature extracted, the logarithmic energy of the ERB spectrum is taken as the audio feature, and the semantic features of the tags are captured using the LDA model, and finally the two features are aggregated using the attention mechanism as the enhanced feature representation.

(3) Since the feature sequence of music is highly susceptible to noise, this paper transforms the aggregated feature vector sequence into a 3D tensor, which is used as the input to the optimized RNN, and fuses the tensor interpolation results with the prediction results to output the final predicted music tag labelling sequence.

2. Theoretical analysis.

2.1. Recurrent neural network.

The traditional neural network is a forward-oriented network, which can only propagate backward from the current neuron, so it is not suitable for tasks containing contextual information in some fields such as natural language processing, speech recognition, etc. RNN is a kind of neural network model with the ability of long-time memory, which can effectively deal with long sequences to obtain their contextual information, and is an improvement and extension of the traditional neural network [21]. The network structure of RNN is implied in Figure 1.

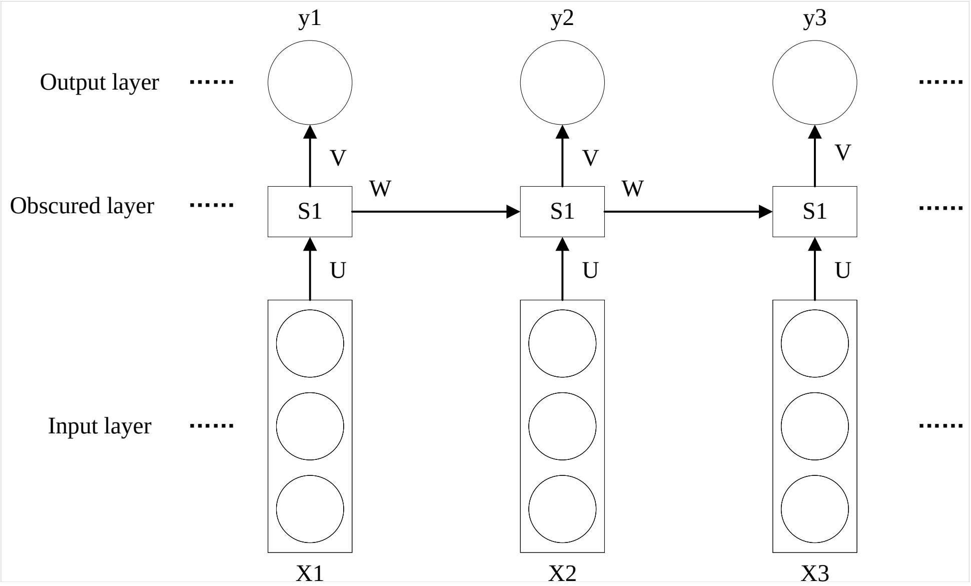

Figure 1. The network structure of RNN

Each neuron on the RNN consists of three layers, namely the input level X, the obscured level L and the output level O. For a sequence $x = (x_1, \dots, x_T)$, the input of the obscured level l is $l = (l_1, \dots, l_T)$, the output of the obscured level is $g = (g_1, g_2, \dots, g_T)$, and the output of the output level is $o = (o_1, \dots, o_T)$. At time t, the computation on the RNN is as follows:

$$\begin{cases} l_t = Ux_t + Vl_{t-1} + b \\ g_t = \vartheta(l_t) + b \\ o_t = \delta(Wg_t) + b \end{cases} \tag{1}$$

where U, V, W are the weight matrix parameters, which are dot-multiplied with the input vector, the obscured state in the previous moment, and the output vector, respectively, b is the bias constant, and ϑ and δ are the activation functions, which are usually adopted as sigmoid, tanh, and so on.

2.2. **LDA thematic model.** The LDA topic model is a hybrid probabilistic model based on the text generation assumption [22], which can reduce the dimensionality of large-scale corpus. By applying the LDA model to analyze a document collection, we can discover the potential topics of each document in the collection, which can express the main content of each document. Documents annotated with potential themes are not only conducive to the classification and organization of the document collection, but also facilitate the retrieval of documents [23].

Assuming that a document contains a total of M words, its subject mixing ratio is λ, and the corresponding subject set of each word in the document is y, then the joint probability distribution of the set of words w under the known conditions of μ and ρ is implied below.

$$p(\lambda, y, w \mid \mu, \rho) = p(\lambda \mid \mu) \prod_{m=1}^{M} p(y_m \mid \lambda) p(w_m \mid y_m, \rho) \tag{2}$$

Next, the edge distribution of the document is obtained by integrating λ and summing over the set of topics y. The formula is as below.

$$p(w \mid \mu, \rho) = \int p(\lambda \mid \mu) \left(\prod_{m=1}^{M} \sum_{y_m} p(y_m \mid \lambda_m) p(w_m \mid y_m, \rho) \right) d\lambda \tag{3}$$

Finally, the probability distribution of the entire document set is obtained by integrating the edge distribution of each document in the document set as bellow.

$$p(D \mid \mu, \rho) = \prod_{d=1}^{N} \int p(\lambda_d \mid \mu) \left(\prod_{m=1}^{M_d} \sum_{y_{dm}} p(y_{dm} \mid \lambda_d) p(w_{dm} \mid y_{dm}, \rho) \right) d\lambda_d \tag{4}$$

3. **Optimization of RNNs based on interpolation methods.** Traditional RNNs are prone to lose effective information when dealing with noisy data, which affects the prediction accuracy. To address this problem, this paper uses the interpolation method [24] to optimize the RNN (ERNN), in which the interpolation module is to capture the temporal correlation between the data within the data streams. The interpolation module is based on the fully-connected neuron layer to capture the temporal correlation between the data values between different data streams, to avoid the accumulation of prediction errors due to the consecutive misses.

The data stream d has a missing value at timestamp t, i.e., $x_t^d = *$; the predicted value at timestamp t, $\hat{x}_t^d$. Traditional RNNs use the actual value $x_t^{d'}$ in the data stream d to estimate the missing value, which leads to computational overload and overfitting problems [25]. In this paper, we design an RNN optimization model based on the interpolation method, which captures the data correlation within and between streams through an efficient hierarchical learning network to achieve accurate estimation of missing values. At the same time, the model limits the data to be learnt to a certain time range to avoid overfitting. The improved recurrent neural network model is implied in Figure 2.

(1) **Interpolation module.** The interpolation module constructs an interpolation function Φ that operates on a given data stream. The estimate $\tilde{x}_t^d$ of x_t^d depends on the data stream from which x_t^d is removed and is denoted as $\tilde{x}_t^d = \Phi(D - x_t^d)$. This

estimate uses only data from the data stream d and not from other data streams. A bi-directional RNN is used to construct Φ. Unlike a traditional bi-directional RNN, the inputs to the obscured level are forward-lagged and backward-advanced. At time point t, the inputs for the forward-hidden state come from $t-1$, and the inputs for the backward-hidden state come from $t+1$. This process ensures that the actual value x_t^d is not automatically used in the estimation of $\tilde{x}_t^d$.

$$\tilde{x}_t^d = g(U^d[\overrightarrow{g}_t^d; \overleftarrow{g}_t^d] + c_o^d) = g(\overrightarrow{U}^d\overrightarrow{g}_t^d + \overleftarrow{U}^d\overleftarrow{g}_t^d + c_o^d) \tag{5}$$

$$\overrightarrow{g}_t^d = f(\overrightarrow{V}^d\overrightarrow{g}_{t-1}^d + \overrightarrow{V}^d y_{t-1}^d + \overrightarrow{c}^d) \tag{6}$$

$$\overleftarrow{g}_t^d = f(\overleftarrow{V}^d\overleftarrow{g}_{t-1}^d + \overleftarrow{V}^d y_{t-1}^d + \overleftarrow{c}^d) \tag{7}$$

where f, g are ReLu activation functions; arrows indicate forward or backward direction. In the interpolation module, only the temporal correlation in each data stream is captured. The parameters of each data stream are learnt individually, and the number of parameters to be learnt is linearly related to the number of data streams. V is the weight matrix, U and W are diagonal matrices.

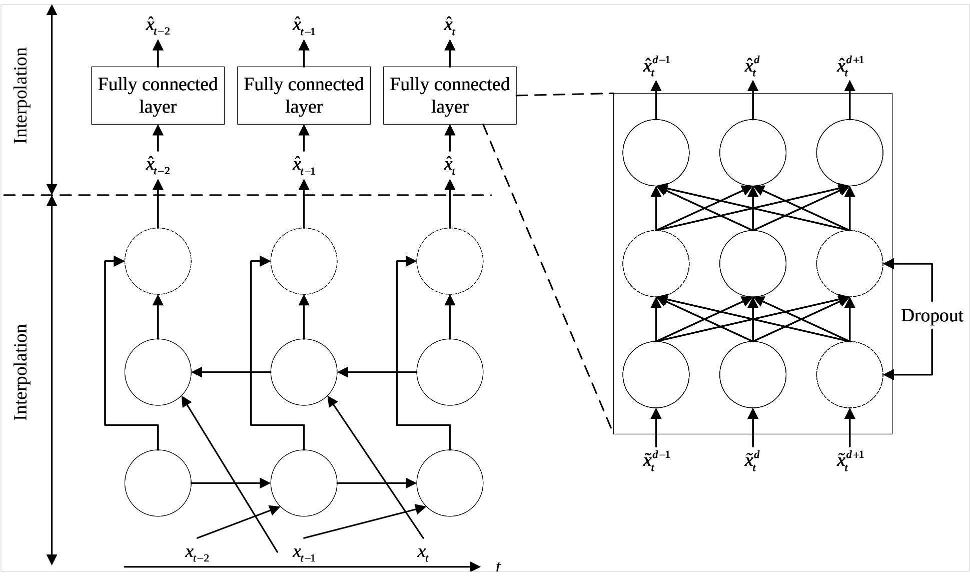

Figure 2. The framework of the suggested ERNN

(2) **Interpolation module.** The interpolation module constructs an interpolation function Ψ that operates across data streams. The estimate $\hat{x}_t^d$ of x_t^d depends on the data stream and is denoted as $\hat{x}_t^d = \Psi(D - x_t^d)$; only data $\hat{x}_t^d$ with timestamp s_t is used. The constructor Ψ is constructed to be independent of t and therefore uses a fully connected level.

$$\hat{x}_t = \delta(Vg_t + \alpha) \tag{8}$$

$$g_t = \vartheta(Ux_t + Wy_t + \beta) \tag{9}$$

where δ and ϑ are activation functions; the diagonal entries of matrix U are zero, α and β are constants greater than 0, respectively.

(3) **Time-based dynamic weighting.** If there are consecutive missing values, the later predicted values will accumulate the errors generated by the previous interpolation, which will have a cumulative effect as the time of consecutive missing becomes larger. To solve this problem, the weight of the predicted value $\hat{x}_t^d$ of the RNN should be as small as possible. Therefore, ERNN introduces time-based dynamic weights.

$$V_t^d = \exp\{-\max(0, \gamma_V \delta_t^d + a_V)\} \tag{10}$$

where γ_V denotes the weight vector and a_V denotes the bias vector.

4. Research on music intelligent annotation technology based on optimized RNN.

4.1. **Feature extraction of music audio signals.** Based on the ERNN model designed in the previous section, this article investigates the ERNN-based music intelligent annotation technique, the overall framework of which is implied in Figure 3 for audio feature extraction, semantic feature extraction, feature aggregation and final label prediction for music. The technique learns rich audio features from raw audio signals and combines them with music semantic vectors to get aggregated features, which are fed into the final ERNN model for label prediction.

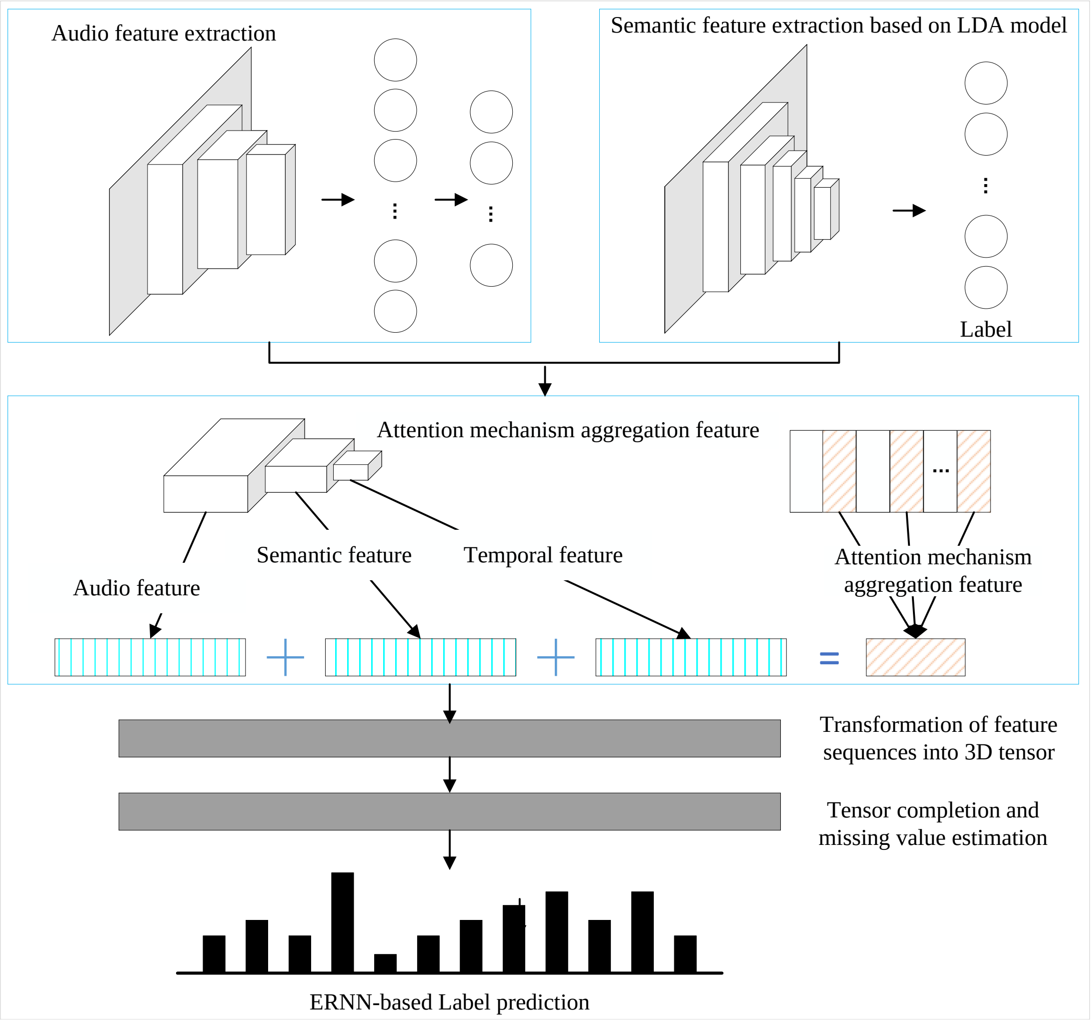
Audio feature extraction
Semantic feature extraction based on LDA model
Label
Attention mechanism aggregation feature
Audio feature
Semantic feature
Temporal feature
Attention mechanism aggregation feature
Transformation of feature sequences into 3D tensor
Tensor completion and missing value estimation
ERNN-based Label prediction

Figure 3. General framework of ERNN-based music intelligent annotation technology

Through a certain method, we can obtain the features that can express the essence of music, and these features can distinguish different music, which are audio features. This article will focus on analyzing how to extract audio features that contain music information, so as to improve the effect of music intelligent annotation technology.

(1) **Preprocessing.** Firstly, the original audio signal is processed by pre-emphasis, windowing and framing to obtain the time-domain signal $x(n)$ of each data frame.

(2) The time-domain signal $x(n)$ obtained in step (1) is subjected to a fast Fourier transform to obtain its frequency spectrum $X(k)$, and the conversion formula is as follows.

$$X(k) = \sum_{m=0}^{M-1} x(m) e^{-j2\pi mk/M}, \quad 0 \leq m, k \leq M-1 \tag{11}$$

(3) The linear spectrum $X(k)$ is passed through the GC filter bank to obtain the ERB spectrum, where the GC filter bank is set up with several bandpass filters $H_n(K)$, $0 \leq n < N$, and N representing the number of filters within the spectrum of the signal.

(4) Taking the logarithmic energy of the ERB spectrum obtained in step (3) yields the logarithmic spectrum $S(n)$, with the following transfer function, where $0 \leq n < N$.

$$S(n) = \ln\left(\sum_{k=1}^{M-1} |X(k)|^2 H_m(k)\right) \tag{12}$$

(5) $S(n)$ is transformed to the cepstrum domain by the discrete cosine transform and the audio features D are extracted using differential cepstrum [26], where c and d denote the speech parameters of a frame and k is a constant.

$$D = d(n) = \left(1/\sqrt{\sum_{i=-k}^{k} i^2}\right) \sum_{i=-k}^{k} i * c(n+i) \tag{13}$$

4.2. **LDA-based semantic feature extraction for music.** In the previous section, the focus is on the extraction process of audio features, and the following focuses on the semantic modelling of the music context space. As it is difficult for traditional semantic models to provide any interpretation of the underlying semantics, this paper utilizes the LDA topic model for semantic modelling of music tags.

The goal of this paper is to approximate the original distribution using the variational distribution described above, in the hope that the KL distance between the LDA topic set w^* and document set probability C^* distributions is as small as possible.

$$(w^*, C^*) = \arg\min_{w,C} D(q(\lambda, y \mid w, \varphi) \| p(\lambda, y \mid y, \mu, \rho)) \tag{14}$$

where λ is the subject mixing ratio, y is the set of subjects corresponding to each word, μ, ρ and φ are the LDA parameters, and $D()$ denotes the Dirichlet distribution.

For the goal of solving for the minimum value, each variational parameter is separately derived and the partial derivative is set to zero, so as to obtain the iterative expression, and the optimal variational parameter is obtained after convergence of several iterations.

$$\varphi_{ni} \propto \rho_{iw_n} \exp\{E_q[\log(D_i) \mid w]\} \tag{15}$$

$$E_q[\log(D_i) \mid w] = \varphi(w_i) - \varphi\left(\sum_{j=1}^{k} w_j\right) \tag{16}$$

By using Equation (16) to update the polynomial expectation in Equation (15), the optimal variational parameters are obtained, and the optimal sum of model parameters is finally obtained by maximizing Dirichlet function.

After modelling by LDA, two matrices $P(m \times k)$ and $Z(k \times n)$ will be obtained based on the music label matrix $A(m \times n, \text{n songs and m labels})$. Matrix P is the semantic topic of the labelling matrix, while Z is the topic importance indicator matrix. In the

k-dimensional semantic space, each column vector z_i of Z is a vector of topic distributions for a particular song i, which is also a feature representation of the music context.

4.3. **Feature aggregation based on attention mechanisms.** After the audio features and semantic features of the music are extracted, the two are aggregated using the attention mechanism [27] as a better feature representation of the music. For a music sample, an attention weight matrix $A = [\alpha_1, \alpha_2, \ldots, \alpha_r]$ is computed based on the audio feature sequence $D = \{d_1, d_2, \ldots, d_m\}$, as follows.

$$A = \mathrm{softmax}(V_1 \phi(V_1 X^T)) \tag{17}$$

where V_1 is the parameter matrix and r is a hyperparameter noting the number of weight vectors. Similarly, the attention weight matrix of semantic features can be derived $B = [\beta_1, \beta_2, \ldots, \beta_r]$.

On the basis of the attention weight matrix, this article further calculates a two-dimensional embedding matrix Q to aggregate the feature sequences of the music.

$$Q = (A + B)X \tag{18}$$

From the above, the i-th row vector q_i in Q is the weighted sum $Q_i = \sum_j (A_{i,j} + B_{i,j})x_j$ of the feature vectors corresponding to each moment in the feature sequence. Therefore, in the music embedding representation matrix Q, each individual vector actually focuses on a different part of the whole feature sequence, thus preserving the musical characteristics of many different aspects of the sequence features.

4.4. **Optimized RNN-based label prediction for music sequences.** After obtaining the aggregated feature Q, Q is mapped to a vector sequence $(q_1, q_2, \ldots, q_n)$ by feature mapping, and at the same time the temporal feature E is introduced as an additional feature, so that the input music feature representation is a splicing of the feature vectors of the aggregated feature Q and the temporal feature vector e representation as shown below, and the entire feature vector sequence of the music is $(h_1, h_2, \ldots, h_n)$.

$$h_t = [q_t, e_t], \quad t \in [1, n] \tag{19}$$

Input h_t into the ERNN model for music annotation prediction, because the feature sequence of music is very susceptible to noise, resulting in the loss of effective features, in this paper, the feature vector sequence is transformed into a three-dimensional tensor, denoted as $H \in R^{K \times D \times T}$, H is an incomplete tensor containing missing values, and the tensor is complemented by combining the interpolation method of Section III and the idea of low-rank approximation [28], through the kernel paradigm of convex relaxation for the minimization of the tensor H. In tensor operations, the tensor $H \in R^{K \times D \times T}$ can be expanded into three matrix forms $H \in R^{K \times (DT)}$, $H \in R^{D \times (KT)}$, and $H \in R^{T \times (KD)}$, denoted as $H_{(1)}, H_{(2)}, H_{(3)}$, respectively, and to obtain the interaction information of different sub-tensors, one tensor can be written as three tensor expansion matrices of weights and forms $H = \sum_{k=1,2,3} \mu_k H_{(k)}$, $\sum_{k=1,2,3} \mu_k = 1$. Therefore, the optimization problem of minimizing the tensor H is transformed into $\min_H \sum_{k=1,2,3} \mu_k \|H_{(k)}\|_*$.

After the minimization tensor H is obtained, the interpolation value $\hat{x}_t^d$ is obtained by using the estimated value $\tilde{x}_t^d$ and the predicted value of the previous time step of the ERNN, which are computed jointly. The initial hidden state g_0 is initialized to a zero vector and then the model is updated by the following equation.

$$c_t^d = m_t^d \times x_t^d + (1 - m_t^d) \times I_t^d \tag{20}$$

$$g_t = ERNN(g_{t-1}, c_t^d) \tag{21}$$

$$\hat{x}_{t+1}^d = V_x g_t + b_x \tag{22}$$

$$l = \sum_t \sum_d m_t^d (x_t^d - I_t^d)^2 \tag{23}$$

where m denotes the equilibrium parameter, c_t^d is the input value, with the true value used directly as input for the non-missing values, and the missing values replaced by a weighted sum of the interpolated values from tensor complementation and the estimated value of ERNN. g_t denotes the hidden state, $\hat{x}_{t+1}^d$ denotes the predicted value at time step t, and l denotes the loss function for a time series using squared error.

Based on the above calculation, the score of each tag is calculated as $o_t = V_o g_t + b_o$ based on the complementary music features H_t. Define a transfer matrix A_{ij} to represent the scores from tag i to tag j. Then the score of the whole sequence $(h_1, h_2, \ldots, h_n)$ is implied in Equation (24), y_0 and y_{n+1} are the start and end tags of the music.

$$S(x, y) = \sum_{t=0}^{n} A_{y_t, y_{t+1}} + \sum_{t=1}^{n} o_{t, y_t} \tag{24}$$

The probability of the entire sequence is implied below.

$$P(y \mid x) = \frac{e^{S(x,y)}}{\sum_{\bar{y} \in Y_x} e^{S(x,\bar{y})}} \tag{25}$$

The logarithmic probability of maximizing the correct label sequence during training is as in Equation (26).

$$\log(P(y \mid x)) = S(x, y) - \log\left(\sum_{\bar{y} \in Y_x} e^{S(x,\bar{y})}\right) \tag{26}$$

where Y_x denotes all possible tag sequences, the most probable tag sequence is predicted to be output at the time of decoding by the following equation, i.e., the final predicted music labelling sequence.

$$y^* = \operatorname{argmax}_{\bar{y} \in Y_x} S(X, \bar{y}) \tag{27}$$

5. **Performance testing and analysis.**

5.1. **Comparison and evaluation of experimental results.** In this section, the performance of the proposed smart music annotation technique is experimentally evaluated by training and testing the models on a CentOS 7 server with a 24-core 2.10GHz CPU, 128G RAM, 500GB mechanical hard drive capacity, and a GeForce GTX 1080 Ti GPU with 12G graphics memory. The models are implemented using Keras and TensorFlow deep learning frameworks. The number of training iterations was set to 100 and the network was trained using the Adam optimization algorithm to update the network parameters. The EarlyStopping method is introduced in the experiments to reduce the overfitting of the network to the training data.

The performance of the music annotation technique is evaluated using the popular multi-million music dataset MTAT [29], which contains 29,358 music clips, each of which is 28.15 seconds long and stored in mp3 format. There are 183 music tags, which can be classified into four types according to their attributes: music genre (61), instrument (59), singer information (35), and music mood (28). In the implementation of this study, the training set, validation set and test set are distributed according to 12:1:3.

In this paper, average accuracy (AvgAcc), average precision (AvgPre), average recall (AvgRec), average F1 (AvgF1), mean average precision MAP, and area under the ROC curve enclosed with the axes AUC are used to evaluate the performance of the SMA-SVM method [11], MIR-SOM method [13], MCS-CNN method [15], MAT-GRU method [20] and MIA-ERNN method performance as implied in Figure 4.

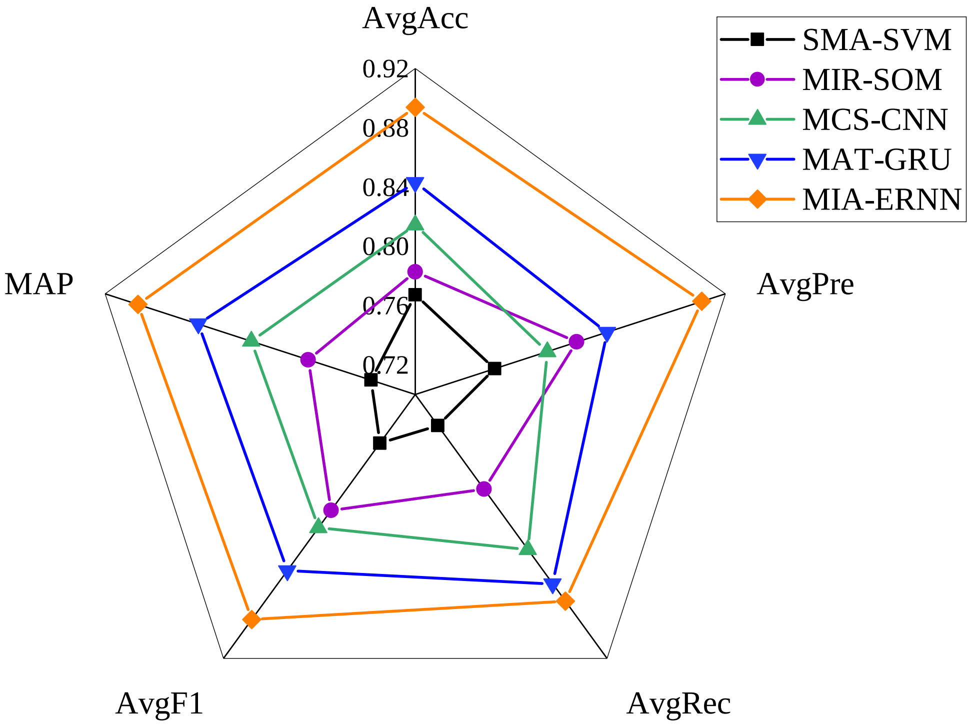

Figure 4. Comparison of prediction accuracy under different indicators

SMA-SVM only considers the emotional features of the lyrics text and inputs them into a traditional SVM classifier for music annotation, and thus performs the worst in all the metrics. MIR-SOM uses timbre and acoustic features as inputs to a self-organising neural network (SNN) for music annotation prediction, and does not incorporate the textual features. MCS-CNN fully incorporates both temporal and audio features to further improve the prediction. MAT-GRU and MIA-ERNN are both RNN-based music annotation methods, with the difference that MAT-GRU only uses a single audio feature as the input to the GRU and does not optimise the traditional GRU, resulting in a lower prediction accuracy than that of MIA-ERNN. In addition, the results of AvgF1 are more reflective of the accuracy strengths and weaknesses of the methods, SMA-SVM, MIR-SOM, MCS-CNN, MAT-GRU and MIA-ERNN have AvgF1 of 0.7407, 0.7965, 0.8111, 0.8469 and 0.8876, respectively, which suggests that MIA-ERNN achieves the best performance.

For the music intelligent annotation problem, the prediction accuracy cannot be used only as an evaluation index, but the imbalance of the dataset should also be taken into account, and the AUC can reflect the classifier's ability to classify the heavily skewed dataset more realistically and comprehensively. In this paper, 10 experiments were conducted on MIA-ERNN and the other four methods, and the box plots of the 50 AUC values obtained by the five methods were plotted, and the images obtained are implied in Figure 5. The mean and median of the AUC of MIA-ERNN is 0.903, which is larger than that of the comparison methods, and the box plot of MIA-ERNN is higher than that of the comparison methods, so that MIA-ERNN's AUC is higher than that of the comparison methods. The boxplots of MIA-ERNN are generally higher than the boxplots of AUC of the comparison methods, so the AUC of MIA-ERNN is generally larger than that of the comparison methods. In addition, the distribution of AUC of MIA-ERNN is more concentrated than that of the comparison method, so the automatic music annotation effect of MIA-ERNN is more stable.

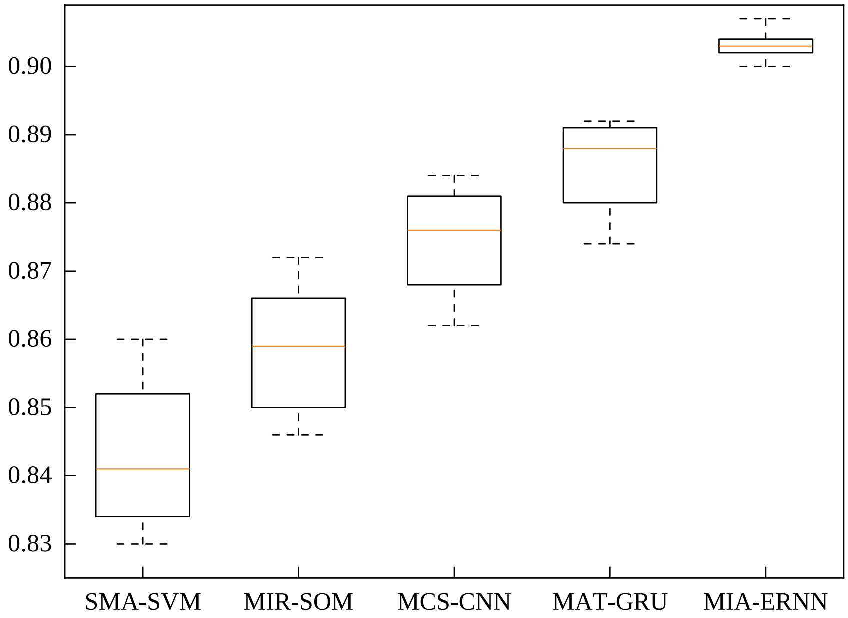

Figure 5. AUC box plots for different methods

5.2. **Results of ablation experiments with different components.** To better verify the impact of the feature extraction module, feature aggregation module and optimization RNN module in MIA-ERNN, this paper carries out an ablation study on MIA-ERNN, and the results are shown in Table 1. The ablation experiments of each component are as below.

(1) Remove the audio feature extraction module. Directly use music semantic features as input to ERNN for music annotation prediction, denoted as MIA-ERNN/AF.

(2) Remove the semantic feature extraction module. Directly use audio features as input to ERNN for music annotation prediction, denoted as MIA-ERNN/SE.

(3) Remove the attention aggregation feature module. Splicing music audio features and semantic features as input to ERNN for music annotation prediction, denoted as MIA-ERNN/AM.

(4) Remove the optimization module for RNN. The result of the aggregation of musical audio features and semantic features is used as the input of the traditional RNN for music labelling prediction, denoted as MIA-ERNN/ET.

Table 1. Ablation results of different components in MIA-ERNN

Method	AvgAcc	AvgPre	AvgRec	AvgF1	MAP
MIA-ERNN/AF	0.8492	0.8634	0.8398	0.8514	0.8461
MIA-ERNN/SE	0.8726	0.8776	0.8684	0.8730	0.8637
MIA-ERNN/AM	0.8655	0.8741	0.8491	0.8614	0.8745
MIA-ERNN/ET	0.8348	0.8512	0.8273	0.8391	0.8285
MIA-ERNN	0.8937	0.9033	0.8724	0.8876	0.8967

As can be seen from Table 1, MIA-ERNN/SE and MIA-ERNN/AM both outperform the other groups of models and the gap between them and the full model is small, which indicates that the semantic features and attention aggregation features have a certain impact on the prediction accuracy. The performance of MIA-ERNN/ET is worse than the other four methods, which indicates that optimizing the traditional RNN models greatly improves the music labelling prediction efficiency, demonstrating the importance of optimizing the RNN. The next worst performance of MIA-ERNN/AF indicates that the audio features of the music itself have a greater impact on the prediction results compared to the semantic features. In conclusion, MIA-ERNN, which incorporates all components, achieves the best performance.

6. **Conclusion.** Due to the presence of noise in real music annotation datasets, the performance of music intelligent annotation task is adversely affected. To address this issue, this article suggests an ERNN-based music intelligent annotation technique. Firstly, the interpolation method is used to optimize the RNN, and the data correlation within and between streams is effectively captured by an efficient hierarchical learning network, so as to achieve the accurate estimation of missing values. Then the music audio signal is preprocessed and feature extracted, the logarithmic energy of the ERB spectrum is taken as the audio feature, and the semantic features of the music tags are captured using the LDA model, and finally the two features are aggregated using the attention mechanism as the enhanced feature representation. Finally, the sequence of aggregated feature vectors is transformed into a 3D tensor, which is used as the input of ERNN, and the tensor interpolation results and the prediction results of ERNN are fused to output the final predicted music tag annotation sequence. Experiments indicate that the suggested technique has high AvgAcc, AvgF1 and MAP, which is a big improvement compared with the existing methods.

REFERENCES

[1] F. Rahman, and J. Siddiqi, "Semantic annotation of digital music," *Journal of Computer and System Sciences*, vol. 78, no. 4, pp. 1219-1231, 2012.

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303-319, 2010.

[3] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20-30, 2018.

[4] Y. S. Ghatas, M. B. Fayek, and M. M. Hadhoud, "Generic symbolic music labeling pipeline," *IEEE Access*, vol. 10, pp. 76233-76242, 2022.

[5] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035-1047, 2005.

[6] Y.-T. Kuo, and M.-C. Chuang, "A proposal of a color music notation system on a single melody for music beginners," *International Journal of Music Education*, vol. 31, no. 4, pp. 394-412, 2013.

[7] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088-1110, 2011.

[8] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640-221653, 2020.

[9] M. Rospocher, "Explicit song lyrics detection with subword-enriched word embeddings," *Expert Systems with Applications*, vol. 163, 113749, 2021.

[10] B. K. Bolla, S. R. Pattnaik, and S. Patra, "Detection of Objectionable Song Lyrics Using Weakly Supervised Learning and Natural Language Processing Techniques," *Procedia Computer Science*, vol. 235, pp. 1929-1942, 2024.

[11] R. Miotto, and G. Lanckriet, "A generative context model for semantic music annotation and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1096-1108, 2011.

[12] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41-51, 2018.

[13] Y. Yu, S. Tang, F. Raposo, and L. Chen, "Deep cross-modal correlation learning for audio and lyrics in music retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1, pp. 1-16, 2019.

[14] J. Singh, "An efficient deep neural network model for music classification," *International Journal of Web Science*, vol. 3, no. 3, pp. 236-248, 2022.

[15] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied Soft Computing*, vol. 52, pp. 28-38, 2017.

[16] M. Furner, M. Z. Islam, and C.-T. Li, "Knowledge discovery and visualisation framework using machine learning for music information retrieval from broadcast radio data," *Expert Systems with Applications*, vol. 182, 115236, 2021.

[17] H. Tang, Y. Zhang, and Q. Zhang, "The use of deep learning-based intelligent music signal identification and generation technology in national music teaching," *Frontiers in Psychology*, vol. 13, 762402, 2022.

[18] J. Grekow, "Music emotion recognition using recurrent neural networks and pretrained models," *Journal of Intelligent Information Systems*, vol. 57, no. 3, pp. 531-546, 2021.

[19] S. Rajesh, and N. Nalini, "Musical instrument emotion recognition using deep recurrent neural network," *Procedia Computer Science*, vol. 167, pp. 16-25, 2020.

[20] Y.-H. Lin, and H. H. Chen, "Tag propagation and cost-sensitive learning for music auto-tagging," *IEEE Transactions on Multimedia*, vol. 23, pp. 1605-1616, 2020.

[21] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19-46, 2024.

[22] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, p. 2339, 2023.

[23] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, p. 1977, 2023.

[24] J. Go, K. Sohn, and C. Lee, "Interpolation using neural networks for digital still cameras," *IEEE Transactions on Consumer Electronics*, vol. 46, no. 3, pp. 610-616, 2000.

[25] H. Zhang, Z. Wang, and D. Liu, "A comprehensive review of stability analysis of continuous-time recurrent neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 7, pp. 1229-1262, 2014.

[26] G. D. Saxena, N. A. Farooqui, and S. Ali, "Extricate Features Utilizing Mel Frequency Cepstral Coefficient in Automatic Speech Recognition System," *International Journal of Engineering and Manufacturing*, vol. 12, no. 6, 14, 2022.

[27] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48-62, 2021.

[28] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Learning efficient sparse and low rank models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1821-1833, 2015.

[29] S. Bengani, S. Vadivel, and J. A. A. Jothi, "Efficient music auto-tagging with convolutional neural networks," *Journal of Computer Science*, vol. 15, no. 8, pp. 1203-1208, 2019.