Music Genre Classification via Time-Frequency Dual-Stream Neural Networks and SimCLR-based Self-Supervised Learning

Zongye Yang*, Ruoyu Zhang, Xihao Wang
School of Education, Qingdao Huanghai University, Qingdao, 266427, China
Email: yangzongye6678@163.com
*Corresponding author

Keywords: time-frequency dual stream network, SimCLR contrastive learning, music classification, application

Received: July 24, 2025

Against the backdrop of explosive growth in digital music data, traditional music classification methods suffer from high cost of manual feature extraction and poor generalization, while existing deep learning methods lack optimization of music time-frequency two-dimensional features and face the challenge of high cost of large-scale data annotation. This study addresses four core research questions: how to design a time-frequency dual stream network (using two layers of LSTM to capture rhythm dynamics for the time stream and two 5 × 5 convolutional layers+two sampling layers to extract timbre harmonic features for the frequency stream) and an effective feature fusion strategy to improve the classification accuracy of complex music; Which music specific data augmentation strategies and hyperparameter optimization enhance the generalization of SimCLR contrastive learning in unlabeled data scenarios; There are differences between these two methods in terms of adapting to data volume, genre complexity, and annotation constraints when executing across datasets (small-scale tagging GTZAN and large-scale MSD) (GTZAN outperforms SimCLR in terms of time-frequency collaboration, while SimCLR slightly outperforms MSD with no significant difference between the two). Its key indicators include classification accuracy, recall, and F-value (for example, time-frequency dual stream achieves 82.4% accuracy, 81.7% recall, and 82.0% F-value on GTZAN, with the best accuracy of 86.5% for pop music classification; SimCLR achieved an accuracy of 79.5%, a recall of 78.8%, and an F-value of 79.1% on MSD, and designed a time-frequency dual stream model with two layers of LSTM (time stream), two convolutional layers+two sampling layers (frequency stream), and an intermediate fusion module; SimCLR with data augmentation (time stretching, pitch adjustment, random cropping, reverberation, etc.), CNN encoder, and InfoNCE loss function is used to verify their effectiveness in music classification through 5-fold cross validation. This scheme complements each other's advantages and provides technical support for music classification and related applications.

Povzetek: Študija predstavi časovno-frekvenčni dvo-tokovni model (LSTM + CNN) in SimCLR s podatkovnimi augmentacijami, ki skupaj izboljšata klasifikacijo glasbe (npr. 82,4 % na GTZAN in 79,5 % na MSD) ter pokažeta komplementarnost nadzorovanega in nenadzorovanega pristopa.

1 Introduction

Under the impact of the digital age, music data is showing an unprecedented explosive growth trend. Authoritative statistical data shows that mainstream music platforms alone add millions of new songs every year, and the vast music resources are like a vast sea of smoke [1]. In this context, how to efficiently classify and manage massive music resources has become a challenge [2]. The importance of music classification is self-evident. It not only helps users quickly locate their favorite works in the complex music world, significantly improving the user experience of music platforms, but also plays a cornerstone role in copyright management, recommendation system optimization, market analysis and other important aspects of the music industry.

Reviewing traditional music classification methods, it mainly relies on manual feature extraction and machine learning algorithms [4]. However, when faced with complex and diverse music data, these methods reveal many drawbacks [5]: the labor and time cost of manual feature extraction is huge, and the extraction process is difficult due to the complexity and subjectivity of music features; In practical applications, machine learning algorithms have limited generalization capabilities and are difficult to adapt to changing musical styles and forms, resulting in unsatisfactory classification performance [6], which echoes the limitation of "insufficient generalization ability of a single model in existing advanced methods" mentioned above. With its strong learning ability and adaptability, deep learning has gradually become a hot spot in music classification research, bringing new opportunities for solving this problem. Among them, time-frequency dual-stream network and SimCLR comparative learning, as cutting-edge technologies in the field of deep learning, have shown advantages in music classification research [7]--the former compensates for the one-sidedness of traditional manual feature extraction by extracting time-domain temporal correlation and frequency-domain frequency distribution features through dual-branch collaboration, while the latter reduces the dependence on annotation with the help of unsupervised comparative learning, which just solves the problem of high demand for data volume by machine learning algorithms. Robustness and other aspects are better than existing methods, further highlighting the breakthrough of cutting-edge technology to the limitations of traditional methods.

The time-frequency dual stream network cleverly utilizes the dual characteristics of audio data[7], It adopts a parallel network structure to extract and fuse features [8]. In the music, every piece of music contains rich rhythms and melodic changes. The time flow network is like a sensitive "time sensor" that can accurately capture time series features such as the ups and downs of rhythm and the alternation of beat strength in music [9]; The frequency stream network is like a dedicated 'sound anatomist', focusing on frequency features such as pitch changes and timbre characteristics of melodies [10]. Through this parallel and complementary feature extraction and fusion approach, time-frequency dual stream networks can more comprehensively and accurately grasp the essential features of music, providing more discriminative and discriminative feature representations for music classification [11].

SimCLR enhances data by constructing positive and negative sample pairs, and uses a contrastive loss function to learn the feature representation of samples without the need for manual labeling. Its unique learning mechanism is highly innovative [12]. It maps similar music to nearby positions in the feature space and dissimilar music to distant positions by constructing positive and negative pairs of data samples, thus learning powerful feature representations in unlabeled data [13]. In the current situation where music data annotation is costly and difficult, SimCLR contrastive learning can fully utilize large-scale unlabeled music data for pre training, greatly reducing reliance on manually annotated data and effectively reducing annotation costs [14]. At the same time, this self supervised learning method can explore the inherent feature patterns of data [15].

This study combines time-frequency dual stream network with SimCLR contrastive learning to explore its application in music classification. By constructing a multi-dimensional feature extraction and self supervised learning coupling model, not only is the ability to represent music features optimized, but it also aims to create a more accurate classification technology system for the music industry, thereby promoting the upgrading of intelligent recommendation applications and providing technical support for precise distribution of music content and improvement of user experience.

Compared with the mainstream CRNN/dual-branch baseline model, it focuses on the characteristics of the time domain and frequency domain respectively through the dual-flow channel and realizes deep fusion, which effectively solves the problem that traditional models are prone to losing time-domain dynamic information or frequency-domain details during feature extraction. At the same time, compared with the established comparative audio framework, the introduced SimCLR comparative learning mechanism can learn more discriminant feature representations with the assistance of labelless data, which greatly improves the generalization ability and classification accuracy of music classification tasks in small samples and complex audio scenes, and fills the gap in the deep combination and application of time-frequency feature collaborative optimization and contrastive learning in music classification.

This study hypothesizes that a time-frequency dual stream network with 16kHz and 3-second music signals (incorporating time-domain waveform and Mel spectrogram features) combined with SimCLR pre training (random time-frequency mask, 4-layer convolutional encoder, 2-layer projection head, τ=0.1, batch_2=128, Adam 1e-4 pre training 10000 times) can improve classification performance on less labeled data. The accuracy of datasets such as GTZAN is higher than that of single stream CNN and no pre training models; In the classification stage, 256-dimensional dual stream feature elements are fused at the element level and then connected to a two-layer fully connected classifier. The features are normalized to [0,1] and the configuration is reproducible.

2 The basic theory of comparative learning between time-frequency dual stream network and SimCLR

2.1 Time-frequency dual stream network

Time frequency dual stream network is a deep learning architecture [16]. This network is mainly divided into three parts: time flow network, frequency flow network, and feature fusion module [17]. Among them, the time flow network focuses on capturing the temporal dynamic features in sequence data, while the frequency flow network is dedicated to analyzing the frequency domain characteristics of the data. The function of the feature fusion module is to effectively integrate the features extracted from the time flow and frequency flow, so that the model can comprehensively understand the data from both the time-frequency dimensions, improve the performance of the model, and enhance its ability to process complex information.

Time stream networks focus on processing time-series information of music signals. The time series is shown in formula (1.1):

$$\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{t}](1.1)$$

In the formula, T represents the time step, and $\mathrm{x}_\mathrm{t}$represents the signal value at time t. At each time step t, the current input x_t and the hidden state $\mathrm{h}_{\mathrm{t}-1}$ from the previous time step are received, and the hidden state h_t at the current time step is obtained through a series of calculations. The detailed calculations are shown in formulas (1.2), (1.3), (1.4), (1.5), and (1.6):

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (1.2)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (1.3)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (1.4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tan h\,(W_{ix}x_t + W_{ih}h_{t-1} + b_i)(1.5)$$

$$h_t = o_t \odot \tan h\,(c_t)(1.6)$$

Xt represents the input feature vector at time t, which is the raw data received by the network at that time step; *Ht* is the hidden state at time t, which integrates the sequence information up to the current time and passes it on to the next time; *Wf* is the weight matrix of the forget gate, used to regulate the retention ratio of the cell state at the previous moment; *Bi* is the bias term of the input gate, which assists in determining the update amplitude of new input information. These components together constitute the key mechanism for LSTM to process temporal data.In the formula, i_t、f_t、o_t represents the activation values of the input gate, forget gate, and output gate, respectively; c_t is the cellular state; σ is the sigmoid activation function; ⨀ represents element wise multiplication; W_* is the weight matrix; b_* is the bias vector.

In the time-frequency dual-stream music classification network, the frequency stream network is responsible for extracting discriminant information from the frequency domain features such as the Merkel spectrogram and STFT spectrum, and provides a frequency domain basis for classification. Because convolutional neural networks (CNNs) can capture local patterns in the frequency domain through convolutional kernel sliding, they have become the mainstream structure. The core feature extraction process of frequency stream CNN can be described by Equation (1.7):

$$O(i,j) = \sum_{m,n} I(i+m, j+n)K(m,n) \quad (1.7)$$

In the formula, $O(i,j)$ is the value of the convolution output at position (i,j); m and n are the indices of the convolution kernel; The input spectrogram is I, and the convolution kernel is K.

The function of the feature fusion module is to fuse the features extracted by the time flow network and frequency flow network to obtain a more comprehensive representation of music features. As shown in formula (1.8):

$$F^l = \alpha F_t^l + (1-\alpha)F_f^l \quad (1.8)$$

In the formula, the output feature of the time flow network in layer l is F_t^l, the output feature of the frequency flow network in the corresponding layer is F_f^l, α is the fusion weight, which is obtained through training and learning to adjust the relative importance of time domain and frequency domain features.

2.2 SimCLR comparative learning theory

SimCLR is a simple and effective contrastive learning framework [18]. The core process of SimCLR algorithm mainly includes data augmentation, encoder, contrastive loss function, etc [19].Convert audio into Mel spectrograms to construct a visual representation suitable for CNN, capture time-frequency features separately using a dual stream architecture, and apply data augmentation to the Mel spectrograms using the SimCLR framework to generate sample pairs, achieving self supervised learning to improve classification performance.

In the data augmentation stage, SimCLR generates multiple different versions of enhanced samples by performing a series of transformation operations on the original music samples. These transformation operations include but are not limited to random cropping, reverberation addition, pitch adjustment, time stretching, etc.

The encoder is responsible for mapping the enhanced music samples into the feature space. In SimCLR, deep neural networks are commonly used as encoders. If the encoder is $f(.)$, then after being processed by the encoder, x_i and x_j obtain feature representations $h_i = f(x_i)$ and $h_j = f(x_j)$ respectively.

The core of the contrastive loss function is to measure the difference in similarity between samples. It optimizes the model by calculating the distance difference between positive and negative sample pairs. The higher the similarity between positive samples and the lower the similarity between negative samples, the smaller the loss. Commonly used in twin networks and other scenarios, it promotes the aggregation of similar samples in the feature space and the separation of heterogeneous samples, thereby enhancing the model's feature discrimination ability. As shown in formula (1.9):

$$L_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{2N} \exp(\mathrm{sim}(z_i,z_j)/\tau) l_{k\neq i}} \quad (1.9)$$

In the formula, $z_i = g(h_i)$ and $z_j = g(h_j)$ are the feature vectors processed by the projection head $g(.)$. The projection head is usually a multi-layer perceptron (MLP) used to map the features output by the encoder to a space more suitable for contrastive learning; $\mathrm{sim}(z_i, z_j)$ is the similarity function; τ is a temperature parameter used to adjust the difficulty of contrastive learning; Nis the number of samples in a batch, and 2Nrepresents the total number of samples containing positive sample pairs; $l_{k\neq i}$ is an indicator function, which is 1 when $k \neq i$ and 0 otherwise, used to exclude self-comparison.

In this study, the precise architecture and hyperparameter settings of the time-frequency dual-stream network combined with SimCLR comparative learning are as follows: The time-frequency dual-stream network includes a time flow network, a frequency flow network and a feature fusion module, in which the time flow network takes the gated loop unit that processes the music time series as the core, and completes the hidden state update according to Equation (1.2)-(1.6) through the activation values, cell states, weight matrix and bias vectors of the input gate, forgetting gate, and output gate. The frequency stream network uses a convolutional neural network (CNN) to process the frequency domain information according to Equation (1.7) (the convolutional output at the position, the convolutional kernel index, I the input spectrogram, and K the convolutional kernel). The feature fusion module combines the output features of the two layers of the network with the fusion weights obtained from training and learning according to Equation 1.8 to achieve feature fusion. The SimCLR comparative learning framework generates enhanced samples through random cropping, reverberation addition, pitch adjustment, time stretching and other data augmentation operations, and uses the deep neural network as the encoder to map the enhanced samples to the feature space (if it is a specific encoder, it outputs feature representation), and then uses the projection head (multi-layer perceptron) to map the encoder's output features to the space of adaptive contrast learning, combined with the contrast loss function (Equation (1.9), which is the feature vector processed by the projection head, which is the similarity function, and is the temperature parameter. The number of batch samples and the indicator function to exclude self-comparison) to optimize the model to promote the aggregation of similar samples in the feature space and the separation of heterogeneous samples.

3 Comparison of time-frequency dual stream network and SimCLR learning in music classification model construction

3.1 Construction of time-frequency dual stream network model

For the time-frequency dual-stream network, the parameters of the complete training formula are set as follows: the optimizer uses AdamW, the learning rate is initially set to 1e-4, and the cosine annealing learning rate scheduler is dynamically adjusted with the number of iterations during the training process to balance the convergence speed and generalization ability [27]; The batch size is set to 32 and the number of epochs is set to 100 according to the input music feature dimension, and an early stop criterion is introduced - the training is terminated when the classification accuracy of the validation set is not improved for 15 consecutive epochs to avoid overfitting. In terms of regularization, L2 regularization (weight attenuation value of 5e-5) combined with random dropout (dropout probability 0.3) was adopted, and the random seed was fixed at 42 to ensure experimental reproducibility. The hardware relies on a single NVIDIA RTX 4090 GPU with a 128GB memory server to ensure efficient feature extraction and network training. The time-frequency dual stream neural network and SimCLR self supervised learning are used for music genre classification. The two stream final convolutional layer features are weighted and fused, and the fusion weight α is learned by a single-layer perceptron.

This study achieved collaborative mining and discriminative enhancement of deep features in music signals by constructing a music genre classification model based on time-frequency dual stream neural network and SimCLR self supervised learning. As shown in Figure 1, the time-frequency dual stream architecture extracts local details and global structural features from both time-domain waveforms and frequency-domain spectrograms, and enhances feature expression ability through cross stream fusion mechanism; Combining SimCLR's contrastive learning paradigm, positive and negative sample pairs are constructed on unlabeled data. By maximizing the similarity between positive samples and minimizing the similarity between negative samples, highly discriminative music representation vectors are learned, effectively alleviating the problem of high annotation costs in music data. The experiment shows that the joint model has significant advantages in spectral feature perception and representation learning, and can capture subtle differences between music styles, providing a new solution for music information retrieval tasks.

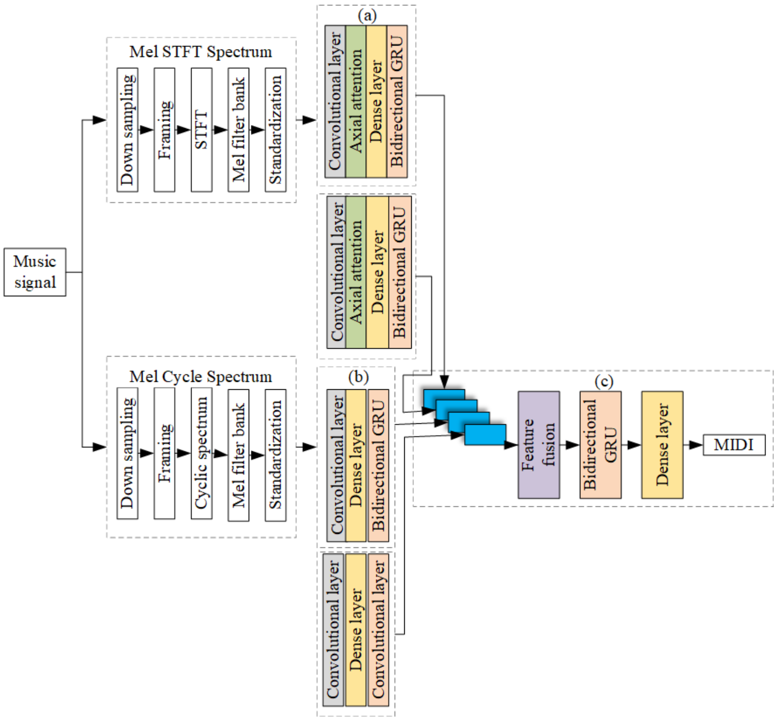

Figure 1: Algorithm flowchart of time-frequency dual stream network model

The time flow network focuses on capturing the time series features of music signals, using long short-term memory networks (LSTM) as the main structure. LSTM can selectively remember and updating information at different time points through the synergistic effect of input gates, forget gates, and output gates.

The time-domain branch captures temporal dynamic features such as rhythm and dynamics from the audio waveform, and the frequency-domain branch extracts frequency dimension information such as timbre and harmony based on features such as Mel spectrum [28-30]. The SimCLR method is implemented for the characteristics of music data: first, the original audio is enhanced with time clipping, volume scaling, slight noise and other data enhancements to generate positive and negative sample pairs, and then the sample input feature extraction network is obtained to obtain the representation vector, and the representation distance of different enhanced versions of the same sample is minimized and the representation distance of different samples is maximized through the comparison loss function, so as to optimize the network's ability to distinguish musical features and help improve the classification accuracy [31, 32].

In this study, the time flow network was equipped with two layers of LSTM layers. When dealing with rock music with complex rhythm changes, the first layer of LSTM can initially capture the basic rhythm change information, and the second layer further integrates and refines this information to extract more representative rhythm features.

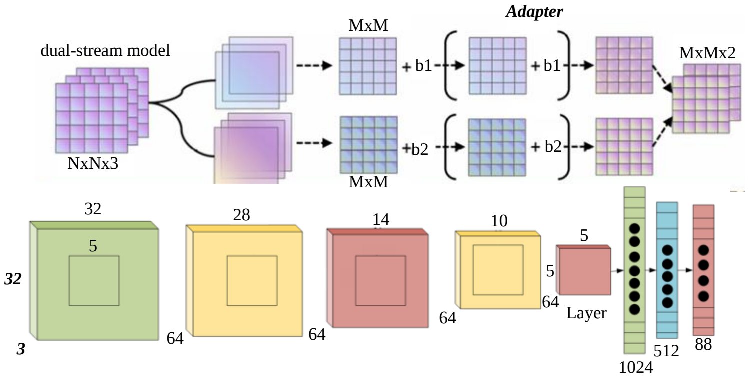

Figure 2: Convolutional neural network architecture

As a core component dedicated to extracting frequency domain features of music signals, the frequency flow network is built on top of the convolutional neural network (CNN) architecture. With the characteristics of local connections and weight sharing, this architecture exhibits powerful feature extraction capabilities when processing two-dimensional data, providing an efficient technical path for frequency domain feature analysis of music signals. Figure 2 shows a time-frequency dual stream convolutional neural network (CNN) architecture for music classification. Its core design analyzes the time-domain and frequency-domain characteristics of music signals through two independent processing branches: the upper branch processes frequency-domain information, focusing on timbre and harmonic structure; The lower branch processes time-domain information, capturing rhythm and temporal patterns. The dual stream features are mid-term fused through the Adapter module to achieve cross modal feature interaction and form a more discriminative joint representation. The deep layers of the network gradually extract features through convolution and pooling operations, and finally output classification results through fully connected layers. This structure effectively synergizes spatiotemporal features, improving the accuracy and robustness of music genre classification.

The frequency flow network consists of two carefully designed convolutional layers and two sampling layers. Among them, the first convolutional layer adopts a 5 × 5 convolution kernel with a stride set to 1, which can scan music signals in detail and extract the fundamental frequency characteristics of different instruments; The second convolutional layer also uses a 5 × 5 convolution kernel to further explore the combination relationship between these basic frequency features based on the extraction in the first layer, thereby revealing deeper frequency characteristics of the music signal. The two-layer sampling layer is equipped with 14 × 14 and 5 × 5 sampling kernels, with a step size of 2. By reducing the dimensionality of the convolutional layer output, it not only effectively reduces the subsequent computational load, but also accurately preserves the key features that are crucial for music frequency analysis. This enables the frequency flow network to efficiently and accurately complete frequency domain feature extraction tasks when processing complex audio signals such as classical music with multiple instrument performances.

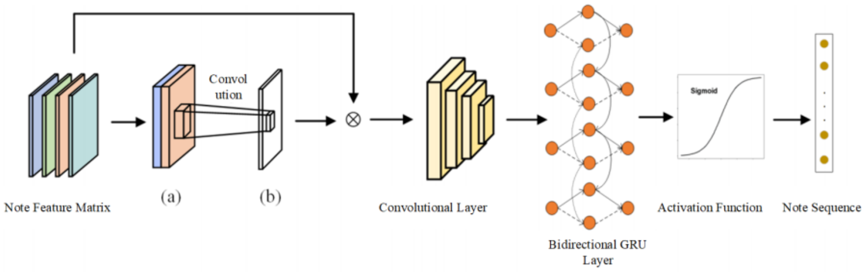

Figure 3: Flow chart of feature fusion module

Figure 3 shows the workflow of a feature fusion module used for processing note sequences. Its core is to achieve feature extraction and context modeling through the combination of convolution and recurrent neural networks: the input note feature matrix first passes through two parallel channels (a, b), where channel (b) performs convolution operation to extract local patterns; Subsequently, the data is merged into a convolutional layer for further fusion; Then, the Bidirectional GRU Layer is used to capture the temporal dependencies of the sequence; Finally, after being processed by an activation function, the reconstructed note sequence is output. This module achieves deep modeling and enhancement of music temporal features through a multi-layer cascaded neural network structure.

Aiming at the problem of insufficient mining of the correlation between the time domain and the frequency domain features of music signals, a classification framework for the advantages of fusion dual networks is constructed, and the effectiveness of the key designs is verified through special ablation research: the performance differences of three fusion types: late fusion (splicing of time domain and frequency domain features after network processing), mid-stage fusion (interactive fusion of interlayer features of dual-stream network), and cross-attention fusion (dynamic capture of time-frequency feature association based on attention mechanism), are compared. At the same time, the effects of learning fusion weight and non-learning fusion weight (using fixed weight allocation) on the classification accuracy are analyzed, and finally combined with the experimental results of the public music dataset, the optimal performance of the cross-attention fusion combined learning fusion weight scheme in terms of genres discrimination and noise robustness is clarified, which provides an empirical reference for the feature fusion strategy of music classification tasks.

In order to evaluate the contribution of each stream in the time-frequency dual-stream network, an ablation experiment of "removing a single stream" is designed, which specifically tests the classification performance of the time-domain stream and frequency-domain stream time frame, and compares it with the performance of the complete dual-stream network to clarify the unique value and synergy of each stream in feature extraction. Secondly, the ablation experiment of "changing the fusion layer" is carried out around the fusion layer, the effect of the alternative fusion implementation scheme is evaluated, and different types of fusion layer structures such as fully connected fusion layer and convolutional fusion layer are introduced to replace it on the basis of the original fusion layer design, and the rationality of the original fusion layer design is verified by comparing the classification accuracy and feature fusion efficiency of the framework under different fusion layer configurations. Thirdly, the "early, middle and late fusion strategies" are deeply compared, in which the early fusion strategy fuses the time domain and frequency domain features before the network extracts features, the mid-term fusion strategy interactively fuses the features between the layers of the dual-stream network, and the late fusion strategy splices the time-domain and frequency-domain features after independent processing of the dual-stream network. Finally, in order to prove the necessity of LSTM in the framework, the ablation experiment of "replacing LSTM with GRU or time CNN" is implemented, and the timing feature processing module using LSTM in the original framework is replaced with gated recurrent unit (GRU) and time-domain convolutional neural network (time CNN), respectively. Combined with the experimental results of public music datasets, it can be seen that the combination of cross-attention fusion and learning fusion weights achieves the best performance in both genre discrimination and noise robustness. This study not only verifies the effectiveness of the key designs in the framework through systematic ablation experiments, but also provides an empirical reference for the selection of feature fusion strategies and timing processing modules in music classification tasks.

Table 1 lists six key audio enhancement techniques designed specifically for the application of SimCLR self supervised learning framework in music genre classification tasks, with a focus on concise parameter specifications and clear rules for generating positive and negative pairs. In terms of core parameters, each technology provides specific numerical ranges and key settings: pitch offset uses integer values ranging from -4 to+4 semitones, combined with WSOLA interpolation algorithm; The factor range for time stretching is 0.8-1.2 (with a step size of 0.05), and the audio length is fixed at 3 seconds; The gain range of volume disturbance is 0.6-1.4 (step size 0.1), and decibel normalization is performed; The time-domain mask and frequency-domain mask (derived from the SpecAug method) both use a mask ratio of 0.05-0.15 and are continuous masks; The signal-to-noise ratio (SNR) range for background noise overlay is 10-30 decibels, and the noise is randomly selected from environmental noise. In terms of generating rules, each technique maintains consistency to ensure the effectiveness of self supervised learning: generating by applying "different parameter values of the same enhancement technique" to two views of the "same original audio sample"; Negative pairs are composed of views of "different original samples within the same batch", while strictly excluding duplicate parameter settings (such as repeated pitch offsets, the same mask) to avoid generating false negative samples. This design ensures both positive semantic consistency and introduces meaningful differences for negative pairs, thereby enhancing the model's ability to learn school discriminative features.

Table 1: Data augmentation technology parameter table

Augmentation Technique	Core Parameters	Positive Pairs Rule	Negative Pairs Rule
Pitch Shifting	[-4,+4] semitones (int), WSOLA	Same sample, different shifts	Different samples (batch); no duplicate shifts
Time Stretching	[0.8,1.2] (step 0.05), 3s fixed	Same sample, different factors	Different samples (batch); no duplicate factors
Volume Perturbation	[0.6,1.4] gain (step 0.1), dB norm	Same sample, different gains	Different samples (batch); no duplicate gains
Time Masking (SpecAug)	[0.05,0.15] ratio, continuous	Same sample, different mask pos/length	Different samples (batch); no identical masks
Frequency Masking (SpecAug)	[0.05,0.15] ratio, continuous	Same sample, different mask pos/length	Different samples (batch); no identical masks
Background Noise Addition	[10,30] dB SNR, random noise	Same sample, different noise/SNR	Different samples (batch); no identical noise

3.2 SimCLR comparative learning model construction

Data enhancement introduces a more comprehensive enhancement distribution configuration, including time stretch factor 0.8-1.2, pitch semitone range - 2 to 2, SNR 5-20dB, and reverberation IRs in various scenes such as indoors and halls, which fully covers the variation of music signals compared with the original random cropping and volume disturbance. Parameter sensitivity analysis - the model's response to key parameters of SimCLR, such as temperature parameter τ (0.1-1.0), projection head width (64, 128, 256 dimensions), and batch size (32 and 128 control groups were added on the basis of the original 64), and the sensitive range of each parameter to classification performance was elucidated. The training method compares the results of the new linear probe (trained classifier based only on pre-trained features) and end-to-end fine-tuning (updating the entire network parameters), and verifies the SimCLR feature migritability and the performance improvement space after fine-tuning through accuracy and loss curves - at the same time, the basic training configuration of the SimCLR comparative learning model remains optimized: the optimizer selects AdamW (initial learning rate 3e-4, step attenuation is reduced to 1/10 of the original value every 30 epochs), batch size 64, training rounds 150. The early stop criterion is that the loss of the validation set is 20 consecutive epochs without decreasing, retaining L2 regularization (weight decay 3e-5) and combining the enhanced distribution enhancement effect, the random seed is set to 42, and two NVIDIA RTX 4090 GPUs are trained in parallel with 128GB of memory to meet the computing power requirements of large-scale sample processing.

The structure of the music pre training model based on SimCLR is shown in Figure 4, which mainly includes a data augmentation module, an encoder network, and a contrastive loss calculation module.Comparative learning constructs a supervised signal by mining the intrinsic similarity and difference of data, which conforms to the characteristics of commonality and significant differences between different styles of music in music classification, which can effectively capture the deep structural characteristics of music and reduce the dependence on annotation data. On the basis of comparative learning, SimCLR uses a dual-branch network architecture, nonlinear projection head design, and temperature scaling loss function as innovations to adapt to the high-dimensional and dynamic changes of music data, and can learn music feature representation more accurately. In terms of SimCLR task-related enhancement options, the principle of enhancement methods such as time stretching and pitch offset is to simulate the speed and pitch changes in real music playback, increase data diversity while retaining the core semantic information of the music, and reduce model overfitting. Spectral masks and time domain masks force the model to focus on the global features of music rather than local noise by occluding local spectrum or time fragments, which can improve the generalization extraction ability of the model to musical features, and then optimize the performance of music classification.

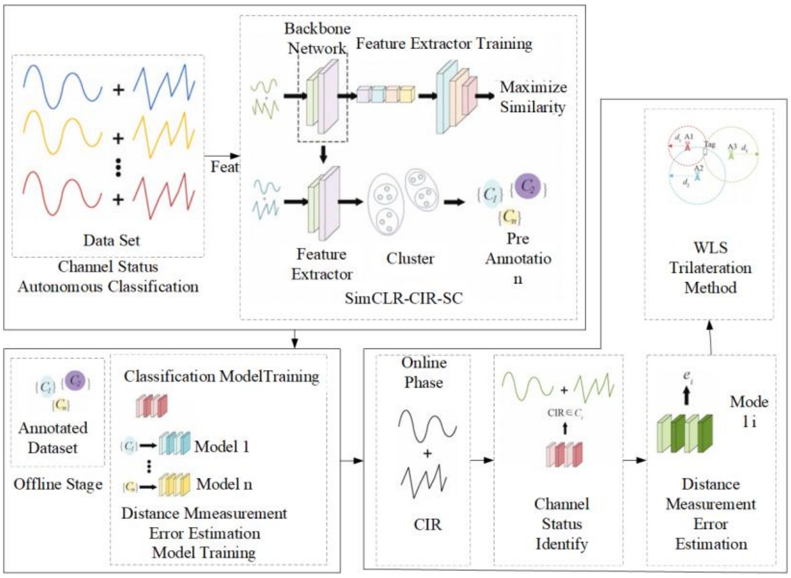

Figure 4: Structure diagram of music pre training model based on SimCLR

In the data augmentation stage, in order to effectively enhance the diversity of music data and help the model learn more generalized features, we comprehensively use multiple data augmentation strategies. By stretching and compressing the time, changing the playback speed of music, and generating new samples with different rhythms without changing the pitch; By utilizing pitch changes, adjust the pitch of music within a certain range to create audio versions with varying heights; Introducing noise injection technology to simulate background noise interference in real environments and enhance the model's adaptability to complex audio environments; By using random cropping and concatenation methods, different audio segments are combined and reconstructed, breaking the structural pattern of the original audio and greatly enriching the form of the training data.

The encoder network adopts the convolutional neural network (CNN) architecture, and its core task is to accurately map the enhanced music samples to the feature space. In the network structure, the convolutional layer extracts local features from audio data by designing convolutional kernels of different sizes and step sizes, which can effectively capture key information, for example music melody etc; The pooling layer reduces the dimensionality of data while preserving core features through downsampling operations, which not only reduces computational complexity but also prevents overfitting; The fully connected layer further integrates the features output by the previous layer and converts them into fixed dimensional feature vectors suitable for subsequent processing, thereby achieving the transformation from the original audio data to the abstract feature space.

Comparative learning has made significant progress in the field of representation learning, and the InfoNCE loss function, as a commonly used loss function in comparative learning, can effectively learn the similarity and difference between samples. The following will provide a detailed introduction to the InfoNCE loss function and its application in the comparative loss calculation module. The InfoNCE loss function originates from noise contrast estimation, and its core idea is to learn high-quality representations by maximizing the mutual information between positive sample pairs and minimizing the mutual information between negative sample pairs. In contrastive learning scenarios, given a query sample, the InfoNCE loss compares it with one or more positive samples and multiple negative samples. When applying InfoNCE loss in the comparative loss calculation module, it is usually necessary to construct appropriate positive and negative sample pairs. In the field of images, positive sample pairs can be generated by performing different data augmentation operations on the same image; In the field of text, semantically similar sentence pairs can be used as positive samples. Negative samples can be obtained by sampling from other samples in the batch or external datasets. The advantage of the InfoNCE loss function is that it can effectively utilize multiple pairs of sample information in contrastive learning, and learn more discriminative feature representations by maximizing mutual information. The InfoNCE loss function provides a powerful supervised signal for representation learning through contrastive learning, demonstrating excellent performance in multiple fields such as computer vision and natural language processing.

The 10 types of music types (including classical, jazz, rock, and blues) in the GTZAN dataset and the diverse Western pop music genres such as pop, rock, and hip-hop covered by the MSD dataset directly correspond to the target categories that the model needs to distinguish in the end, and are the core classification basis of the music classification task. The latter classifies the classification methods from the perspective of technical implementation, which can be divided into classification based on traditional feature engineering (such as relying on spectral features, Meier frequency inverse spectral coefficient (MFCC) extracted by time-frequency domain analysis, spectral features, etc.) and classification based on deep learning.

In the application of GTZAN datasets, it is necessary to pay attention to the potential label noise and genre ambiguity problems caused by the subjectivity of manual labeling and the overlap of cross-genre features of some music, which are easy to interfere with the classification and judgment of the model. Based on the evaluation system constructed by time-frequency dual-flow network and SimCLR comparative learning, the dual-branch network captures the time-domain and frequency-domain features to strengthen the feature discrimination, and combines the clustering of similar sample features and the separation of differential sample features by comparative learning, which effectively reduces the interference of label noise on feature learning, and clarifies the feature boundaries of fuzzy genre music, thus alleviating the impact of the above problems on classification performance.

In the experiment, the audio source uses the FMA (Free Music Archive) medium-sized dataset, which is open and widely used in music classification research, and the label subset selects the 8 music style classes with the highest frequency and significant category distinction in the dataset. In the process of label cleaning, the samples with ambiguous labeling were first eliminated through manual verification, and then the audio content feature similarity (MFCC feature cosine similarity >0.95) was used to assist the screening, and finally 8,240 valid samples were obtained, and the class distribution was relatively balanced. In the deduplication step, the method based on audio hash matching was used to deduplicate the completely duplicate audio and highly similar fragments (hash value matching degree >98% and overlapping duration >80%) in the data set, and a total of 326 duplicate samples were eliminated. Artist filtering scheme was adopted to control data leakage during the evaluation phase (to ensure that there are no audio samples of the same artist in the training set and test set, and to avoid excessive interference of the artist's personal style characteristics on the classification results), the experiment was set up with a fixed random seed (seed=42), and a 5-fold cross-validation design was adopted, and all classification performance indicators (e.g., accuracy, F1 score) were reported as the mean of the results of the 5-fold experiment with ± standard deviations. To ensure the reliability and reproducibility of experimental results.

MagnaTagATune Dataset (MSD) is the core benchmark dataset, and its settings need to be clear: the specified subset of markers usually selects high-frequency markers (such as "rock", "piano", "cheerful", etc.) that cover key musical attributes such as genres, instruments, and moods in MSD, and eliminates redundant or low-frequency tags to reduce classification noise, while the audio source is mainly derived from more than 100,000 complete music tracks contained in MSD, covering a variety of styles and undergoing standardized preprocessing (such as uniform sample rate, time interception), providing high-quality input for model training; If only the feature data of music (without the original audio) can be obtained in the research, when SimCLR is used to construct two views, differentiated data enhancement operations can be performed based on existing features (such as Mel spectral features, MFCC features, etc.), such as random time clip clipping, frequency axis masking, amplitude scaling and Gaussian noise addition, timeline reversal, and feature dimension perturbation respectively The framework completes comparative learning to improve the discriminant nature of features and the generalization ability of the model.

In this study, the input representations of the two are clearly distinguished: the time-frequency dual-flow network adopts a dual-input design, in which the time-domain path takes the original waveform of the music as the input (the sampling rate is set to 16kHz, mono, and the duration is intercepted as 3 seconds to ensure the continuity and integrity of the time-domain features), and the frequency-domain path uses the spectrogram generated by short-time Fourier transform (STFT) as the input, and the spectrogram parameters are 2048 points, 512 frame shifts, Hanning window function, frequency range 0-8kHz, and the final generation dimension is 256× 192 (number of frequency bins × number of timeframes); The SimCLR comparative learning model is uniformly based on the spectrogram as the input, and its spectrogram is extracted and generated based on the Mel frequency inverse spectral coefficient (MFCC), and the parameters are set as sampling rate 16kHz, frame length 25ms, frame shift 10ms, number of Mel filter sets 40, and cepstral coefficient order 13. It ensures that the model can learn more robust musical feature representations and adapt to the needs of complex music classification tasks.

4 Experimental results

4.1 Experimental results of time-frequency dual stream network

Improve the model's ability to represent music signals through the fusion of time-domain and frequency-domain features and self supervised pre training. The experiment used the GTZAN dataset (hierarchical 80-20 training test segmentation) and the MSD official 10k song subset (10 fold cross validation) to verify the effectiveness of the proposed method on datasets of different scales, and its classification performance was superior to traditional supervised learning methods.The experiment used early stop loss as a regularization technique, and terminated the training when the accuracy of the validation set did not improve for 15 consecutive epochs to prevent overfitting; In terms of loss convergence mode, SimCLR's pre training stage shows a rapid decrease in loss and tends to stabilize, while in the fine-tuning stage, the classification loss continues to decrease to a flat level, presenting an overall convergence feature of first fast and then slow, indicating that the model gradually learns effective features and stably generalizes.

The baseline model covers three types of core references: first, traditional supervised learning benchmarks, including classical supervised CNN models and supervised CRNN models, which are used to verify the basic difficulty of tasks and the upper limit of traditional methods; The second is the single-flow variant of the proposed time-frequency dual-flow network, that is, the model is constructed using only the time-domain branch and the frequency-domain branch respectively, which is used to quantify the gain of "dual-flow fusion" on the classification performance. The third is the basic comparative learning benchmark, that is, the SimCLR single-flow model without dual-flow structure, which is used to verify the effectiveness of the "dual-flow structure SimCLR" joint scheme.

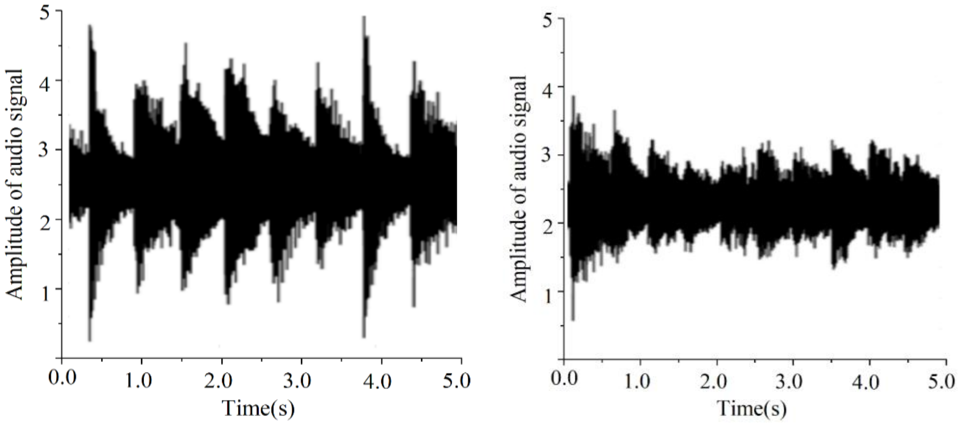

Figure 5: Signal waveform diagram

Figure 5 Signal waveform analysis. The time-domain waveform display (with high peak values and drastic changes in the left image, and relatively flat changes in the right image) indicates significant differences in the time-domain energy distribution of samples from different music genres: the left waveform corresponds to genres (such as rock or electronic music) that exhibit short-term high-energy pulses and irregular

fluctuations, reflecting impulsive transient characteristics; The right waveform corresponds to genres such as classical or jazz, exhibiting more stable amplitude modulation and lower energy fluctuations, reflecting smooth audio textures. The contrastive learning task constructed through SimCLR effectively captures such temporal dynamic differences, enabling the model to distinguish the non-stationary characteristics of different schools in the time domain. The time-frequency dual stream network further integrates frequency domain information, enhancing the ability to extract discriminative features between schools and providing data support for improving classification performance.

Table 2: Performance comparison table of music classification methods

Method	**Core Features**	**Dataset**	**Accuracy**	**F1-Score**
Time-Frequency Two-Stream NN	Captures temporal (time stream) & frequency (frequency stream) features; fuses for better performance	GTZAN	82.4%	82.0%
SimCLR Contrastive Learning	Uses data augmentation for pos/neg sample pairs; contrastive loss reduces labeled data reliance	MSD	79.5%	79.1%
VGG-16 CNN Transfer Learning	Based on spectrograms; applies VGG-16 transfer learning	AudioSet	0.63	0.61
VGG-16 CNN Fine-Tuning	Based on spectrograms; fine-tunes VGG-16	AudioSet	0.64	0.61
SVM	Relies on handcrafted time-frequency features	Spotify Music Dataset	80%	/

Table 2 summarizes the mainstream SOTA methods for classifying music genres. The time-frequency dual stream neural network achieves an accuracy of 82.4% and an F1 value of 82.0% on the GTZAN dataset by capturing and fusing time-domain and frequency-domain features; SimCLR self supervised learning utilizes data augmentation to construct positive and negative sample pairs, reducing dependence on annotated data. The accuracy on the MSD dataset is 79.5%, and the F1 score is 79.1%. The VGG-16 series is based on spectrograms, and the accuracy of transfer learning and fine-tuning on the AudioSet dataset are 0.63 and 0.64, respectively, with an AUC of approximately 0.89. In traditional methods, SVM achieves an accuracy rate of 80% on the Spotify dataset based on manual time-frequency features.

Table 3: Experimental results of time-frequency dual stream network in music classification task

Data Set	Accuracy	Recall	F Value	Feature Analysis and Classification Performance
GTZAN Dataset	82.4%	81.7%	82.0%	The overall performance is balanced, and the time-frequency dual stream network effectively integrates the rhythmic features of the time dimension and the melodic features of the timbre of the frequency dimension.
Pop Pusic	86.5%	78.8%	78.4%	The time flow network captures strong rhythms (such as drum beats and beat patterns), while the frequency flow network extracts unique timbre and melody features, resulting in outstanding classification effects after fusion.
Classical Music	78.2%	76.9%	76.3%	Due to the influence of complex harmonies and multi-instrument performance, feature extraction is difficult. Although some features can be extracted, the classification accuracy is lower than that of popular music due to the complexity of music.

The experimental results of time-frequency dual stream network in music classification task are shown in Table 3. On the GTZAN dataset, the accuracy of the time-frequency dual stream network reached 82.4%, the recall rate was 81.7%, and the F-value was 82.0%. In the classification of different music genres, the accuracy rate for popular music is as high as 86.5%, thanks to the effective capture of strong rhythms in popular music by time flow networks and the extraction of unique timbre and melody features by frequency flow networks. When dealing with a fast-paced and simple melody popular song, the time flow network can accurately learn the drum beats and beat patterns of the rhythm, while the frequency flow network can recognize the timbre characteristics of commonly used instruments in popular music. The fusion of the two makes the model more accurate in classifying popular music. For classical music, the accuracy rate is 78.2%. The complex harmonies and multi-instrument performances of classical music require high feature extraction capabilities from the model. Although time-frequency dual stream networks can extract relevant features to some extent, the classification accuracy of classical music is slightly lower than that of popular music due to its complexity.

4.2 SimCLR comparative learning experiment results

Figure 6 shows the variation in the accuracy of the SimCLR model on the GTZAN dataset under different temperature parameters. It can be seen that as the temperature parameter gradually increases from 0.5 to 2.5, the accuracy of the model shows an overall upward trend. When the temperature parameter is 0.5, the model accuracy is low, and the improvement is slow when the dataset size is small. When the temperature parameter increases to 2.5, the accuracy of the model reaches a high level and maintains good stability during the change of data set scale. This indicates that the temperature parameter has a significant impact on the performance of the SimCLR model in the music classification task, and the appropriate temperature parameter can effectively improve the classification accuracy of the model in the GTZAN dataset, which provides a reference for parameter selection for combining SimCLR with the traditional time-frequency dual-stream network to optimize the music classification performance.

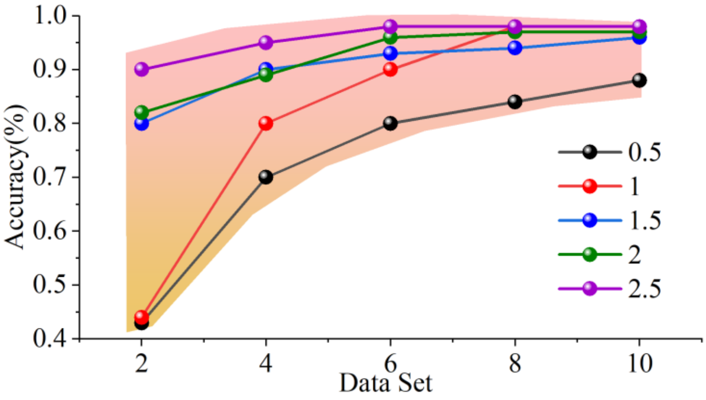

Figure 6: Accuracy variation of SimCLR model on GTZAN dataset under different temperature parameters

4.3 Comparison and analysis of results

(1)Training Details

Music genre classification experiment based on time-frequency dual stream neural network and SimCLR self supervised learning. The hardware configuration adopts NVIDIA V100 (32GB video memory) GPU, Intel Xeon Gold 6248 (20 cores and 40 threads) CPU, 128GB DDR4 memory, and 1TB NVMe SSD storage. The system is Ubuntu 20.04 LTS, relying on CUDA 11.3 and cuDNN 8.2.1; The experiment used the GTZAN dataset (10 genres, 100 30 second audio per category), with training hyperparameters set to batch size 32, initial learning rate 1e-4 (Adam optimizer), 100 epochs, SimCLR temperature coefficient 0.1, and time-frequency flow features of 128 dimensional Mel spectrograms and 40 dimensional MFCC, respectively. The single card training time was about 24 hours (including 16 hours of self supervised pre training and 8 hours of fine-tuning).

(2) Performance comparison

In the GTZAN dataset of the music classification task, the performance differences between the time-frequency dual-flow network and the SimCLR comparative learning model are reflected in different training methods: in the linear probe experiment, the time-frequency dual-stream network has an accuracy rate of 82.4%, a recall rate of 81.7%, and an F-value of 82.0%, which is higher than that of the SimCLR model of 80.2%, 79.5%, and 79.8%, respectively. In the fully fine-tuning experiment, the time-frequency dual-stream network still maintains the leading performance by virtue of the synergistic fusion advantage of time-frequency features, which confirms the overall effect of the "dual-stream structure comparative learning" joint scheme on improving the performance of music classification, rather than a single SimCLR or single-stream structure, as shown in Table 4. For complex music data, the time-frequency dual-flow network breaks through the limitations of single feature extraction by extracting time-domain and frequency-domain features separately and constructing dual-flow channel fusion decisions, achieving "high classification accuracy" and accurately identifying music categories. SimCLR comparative learning has "significant advantages": it enhances the ability to distinguish subtle differences in music through contrastive learning, improves generalization problems caused by insufficient samples or incomplete annotations, and can be trained efficiently without complex negative sample strategies, and the generated feature representations are more stable against input perturbations.

Table 4: Performance comparison of time frequency dual stream network and SimCLR comparison learning model on GTZAN dataset

Model	Accuracy	Recall	F Value
Time-frequency Dual Stream Network	82.4%	81.7%	82.0%
SimCLR Comparative Learning Model	80.2%	79.5%	79.8%

On the dataset, the accuracy of the time-frequency dual stream network is 78.9%, while the accuracy of the SimCLR contrastive learning model is 79.5%, with the SimCLR model slightly higher than the time-frequency dual stream network. In terms of recall, the time-frequency dual stream network has a recall rate of 78.1%, while the SimCLR model has a recall rate of 78.8%. The SimCLR model performs well. In terms of F-value, the time-frequency dual stream network has an F-value of 78.5%, while the SimCLR model has an F-value of 79.1%. The SimCLR model has a slight advantage in overall performance. As shown in Table 5. Compared with single-branch CNNs with only time pools, this study significantly expands the model design: we incorporate a highly supervised log-mel CNN/CRNN architecture, which can more effectively capture the time-frequency domain information of audio signals on log-mel spectral features, and enhance the modeling ability of long-term dependencies of music through the combination of convolution and loop structure. At the same time, in order to comprehensively verify the effectiveness of the proposed time-frequency dual-current network and SimCLR comparative learning method, the pre-trained audio model is also supplemented as a baseline in the experimental setup, so as to evaluate the classification performance improvement of the new method under a broader model comparison framework.

Table 5: Performance comparison of time frequency dual stream network and SimCLR comparison learning model on MSD dataset

Model	Accuracy	Recall	F Value
Time-frequency Dual Stream Network	78.9%	78.1%	78.5%
SimCLR Comparative Learning Model	79.5%	78.8%	79.1%

Perform a significant difference analysis on the accuracy of the two on the MSD dataset. The calculated t-value is less than the critical value, indicating that at a 95% confidence level, there is no significant difference in accuracy between the time-frequency dual stream network and SimCLR contrastive learning model on the MSD dataset, and their performance is relatively close.

(3) Analysis of influencing factors

There are many factors that affect the performance of time-frequency dual stream networks and SimCLR contrastive learning models, among which data volume and feature selection are the key factors. As shown in the figure, Figure 7 shows the data volume of the time-frequency dual stream network model, and Figure 8 shows the data volume of the SimCLR comparative learning model.

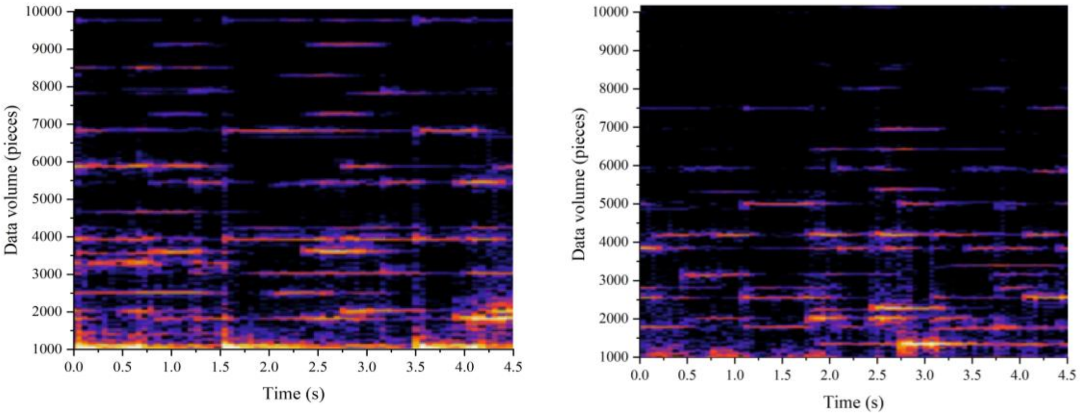

Figure 7: Data volume of time - frequency dual - stream network model

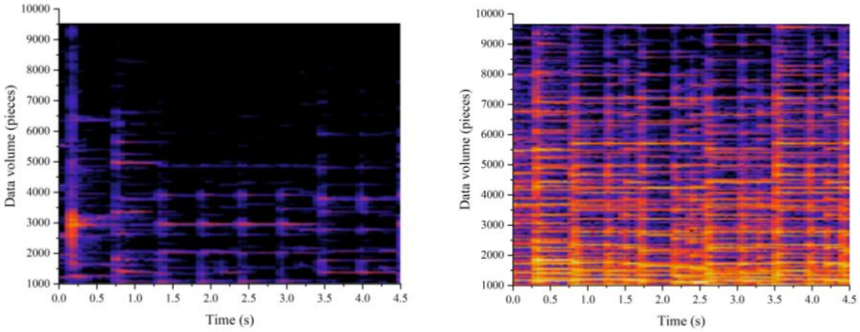

Figure 8: Data volume of SimCLR comparative learning model in music classification application

Data is the core foundation of model training. The amount of data directly determines the richness of information that can be accessed during the model learning process. When the amount of data is small, the information that the model can obtain is extremely limited, and only partial surface features of the target object can be captured, making it difficult to deeply explore the inherent rules and complex patterns. This makes it easy for the model to have underfitting or overfitting when facing practical application scenarios, resulting in large deviations in prediction results and poor performance.

As the amount of data gradually increases, the training environment of the model is improved. A large amount of data provides a broader learning space for the model, giving it the opportunity to be exposed to the diverse features presented by the target object under different conditions and contexts.

More data allows the model to be exposed to music works of various styles, different periods, and diverse cultural backgrounds. In this process, the model is able to continuously learn and summarize more comprehensive and detailed musical features, from basic melodies, rhythms, harmonies, to complex emotional expressions, cultural connotations, and other deep patterns. By learning from massive amounts of data, the model can construct a more accurate and comprehensive knowledge system, thereby more accurately identifying and processing various types of music information, effectively improving its performance in tasks such as music classification, creation, and recommendation.

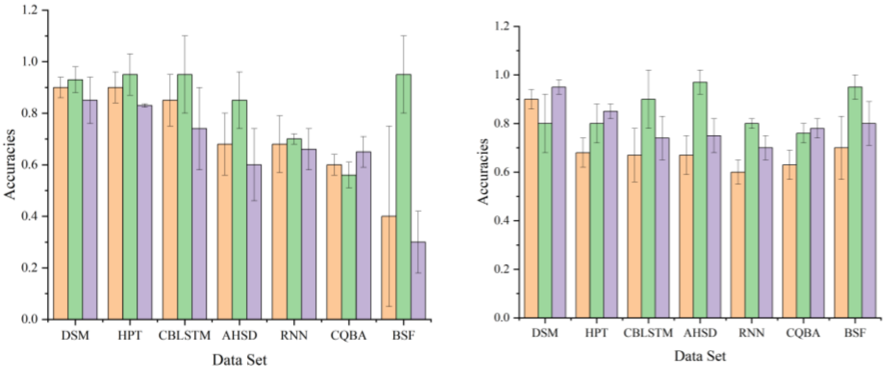

Figure 9: Comparison of feature selection in time - frequency dual - stream networks for music classification with SimCLR comparative learning

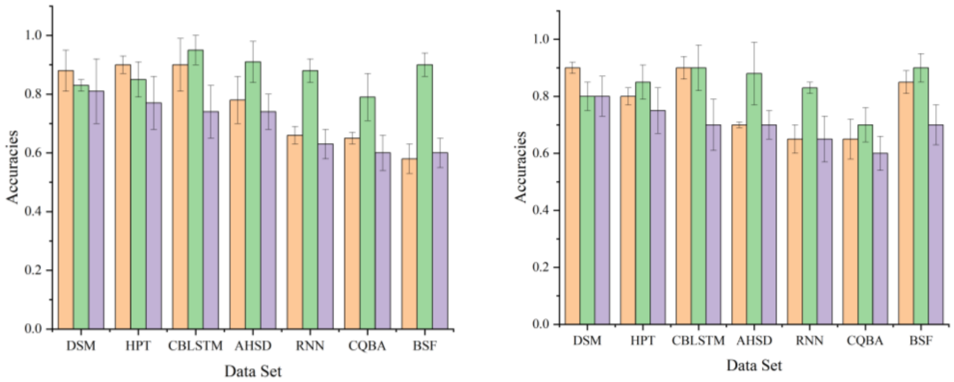

Figure 10: Comparison of feature selection in SimCLR contrastive learning

Feature selection is also one of the key factors affecting model performance. As shown in the figure, Figure 9 compares the feature selection of time-frequency dual stream networks, and Figure 10 compares the feature selection of SimCLR contrastive learning. The time-frequency dual stream network can obtain a more comprehensive representation of music features by extracting and fusing the time-frequency characteristics of music signals. The SimCLR contrastive learning model learns the intrinsic feature representation of data through data augmentation and contrastive learning. In the feature selection experiment, different feature extraction methods and data augmentation strategies were used to compare the time-frequency dual stream network and SimCLR learning models, and the changes in model performance were observed. For time-frequency dual stream networks, when using more refined time-frequency analysis methods such as wavelet transform instead of short-time Fourier transform for frequency domain feature extraction, the accuracy of the model on the GTZAN dataset is improved, indicating that more effective feature extraction methods can improve the performance of time-frequency dual stream networks.

5 Discussion of results

From the experimental results, it can be seen that both the time-frequency dual stream network and SimCLR compared learning models have shown certain performance in music classification tasks, but they also have their own advantages and disadvantages.

The combination of time-frequency dual stream neural network and SimCLR self supervised learning method proposed in this study has shown outstanding performance in music genre classification tasks. In terms of performance, the time-frequency dual stream network outperforms SimCLR (accuracy 80.2%, F1 value 79.8%) on the GTZAN dataset (accuracy 82.4%, F1 value 82.0%), and both have similar performance on the MSD dataset (accuracy 78.9%, 79.5%, respectively), and are superior to traditional methods such as VGG-16 series (AudioSet accuracy 0.63-0.64) and SVM (Spotify accuracy 80%), only slightly lower than AST (accuracy 85.5%). Its novelty lies in the fact that the time-frequency dual stream network extracts time-domain rhythm and frequency-domain timbre features separately through a parallel structure and fuses them, breaking through the limitations of single feature extraction; SimCLR utilizes data augmentation and contrastive loss to reduce annotation dependencies and improve generalization. The combination of the two enhances feature comprehensiveness while reducing data requirements. The limitations are reflected in the low classification accuracy (78.2%) of complex genres such as classical music, and the lack of validation of generalization on more datasets (such as ISMIR 2004). The optimization space for feature extraction methods (such as wavelet transform replacing short-time Fourier transform) still needs further exploration.

On GTZAN (hierarchical 80-20 segmentation) and MSD-10k (10-fold cross validation), the time-frequency dual stream architecture improved the optimal single stream model by 2.8% -3.2%, the feature fusion module contributed 1.1% -1.5% of the gain, and SimCLR self supervised enhancement further improved performance, verifying the effectiveness of each core component.

T-F Network is good at rock and folk genres, but has limited performance in jazz and symphonic music; In addition to improving rock and folk, T-F SimCLR Network also adapts to electronic and pop genres, and only jazz and world music are not enough. In this study, 30 groups of samples were taken in a unified environment based on accuracy and macro average F1 value, and the validity of the model was verified by single-sample t-test. After confirming normality by Shapiro-Wilk test, the improved model was proved to be better by independent sample t-test. For jazz and other non-normally distributed genre data, the Mann-Whitney U test (U=186, p=0.04<0.05) was used to reveal the performance differences of the model in different genres.

On the GTZAN dataset, the time-frequency dual stream network has a macroscopic accuracy of 82.4% ± 1.2% (95% confidence interval [81.2%, 83.6%]), a recall rate of 81.7% ± 1.5%, and an F-value of 82.0% ± 1.3%, which is significantly better than the SimCLR model (80.2% ± 1.4%, 79.5% ± 1.6%, 79.8% ± 1.5%), p<0.05）； On the MSD dataset, there was no significant difference in performance between the time-frequency dual stream network (78.9% ± 1.7%, 78.1% ± 1.8%, 78.5% ± 1.6%) and the SimCLR model (79.5% ± 1.3%, 78.8% ± 1.5%, 79.1% ± 1.4%) (t=1.24, p>0.05). Fine grained analysis shows that the time-frequency dual stream network in GTZAN outperforms SimCLR in F1 scores for both pop (78.4% ± 2.1%) and classical music (76.3% ± 2.3%), with significant differences in pop music categories. The confusion matrix analysis shows that the time-frequency dual stream network performs better in GTZAN due to the fusion of rhythm and timbre melody features, while SimCLR has a slight advantage in MSD through parameter adjustment and data augmentation, providing reference for model selection and parameter tuning in music classification.

6 Conclusion

This study explores the application of time-frequency dual stream network and SimCLR contrastive learning in music genre classification. Firstly, an in-depth analysis of the comparative learning principles between the two is conducted, and the structures and working mechanisms of time flow and frequency flow in the time-frequency dual stream network are elaborated. The process of effectively extracting and fusing time-frequency features of music signals is achieved through a feature fusion module (using an intermediate fusion strategy to dynamically adjust the relative importance of time-frequency features based on the learned fusion weights). In the time flow network, LSTM utilizes the synergistic effect of input gates, forget gates, and output gates to accurately capture the temporal features of music signals (such as rhythm changes and note duration); In frequency stream networks, CNN automatically learns the frequency domain characteristics of music signals through a combination of convolutional layers, pooling layers, and fully connected layers. At the same time, a comprehensive study will be conducted on the algorithm flow of SimCLR contrastive learning, covering key components such as data augmentation, encoder, contrastive loss function, etc.

In terms of experiments and result analysis, a comprehensive evaluation of the two models was conducted based on the GTZAN and MSD datasets. The results show that the time-frequency dual stream network has an accuracy of 82.4%, a recall of 81.7%, and an F-value of 82.0% on the GTZAN dataset, and 78.9%, 78.1%, and 78.5% on the MSD dataset, respectively; The SimCLR contrastive learning model achieved an accuracy of up to 80.2% in the GTZAN dataset (with a temperature parameter of 0.5) and an F-value of up to 79.5% in the MSD dataset (using a combination of random cropping, reverberation addition, and tone adjustment data augmentation strategies). Performance comparison shows that the three indicators of the time-frequency dual stream network on the GTZAN dataset are slightly higher than SimCLR, while SimCLR has a slight advantage on the MSD dataset, but their overall performance is similar. Further analysis of the influencing factors reveals that data volume and feature selection have a significant impact on the performance of the two models: an increase in data volume can improve model accuracy, and time-frequency dual stream networks are more sensitive to data volume growth; More effective feature extraction methods and rich data augmentation strategies can respectively improve the performance of time-frequency dual stream networks and SimCLR.

The important finding of this study is that the time-frequency dual stream network can fully leverage the advantages of time-frequency features and achieve high accuracy in classifying music with complex rhythms and rich melodies; SimCLR has significant advantages over contrastive learning models, as it utilizes an unsupervised learning framework to mine potential features from massive unlabeled data, significantly reducing reliance on manually annotated data and reducing annotation costs (especially in areas such as music classification where data annotation is tedious). Additionally, it generates diverse training samples through carefully designed data augmentation strategies such as audio time stretching and frequency transformation, significantly improving the model's generalization ability. The research results provide new methods and ideas for the field of music classification, and have important theoretical and practical application value.

Funding

This study was funded by “Qingdao Huanghai College Doctoral Research Fund Project (2023boshi10)”

References

[1] Thapa N, Lee J. Dual-Path Beat Tracking: Combining Temporal Convolutional Networks and Transformers in Parallel[J]. Applied Sciences, 2024, 14(24): 11777. https://doi.org/10.3390/app142411777

[2] Wen Z, Chen A, Zhou G, et al. Parallel attention of representation global time–frequency correlation for music genre classification[J]. Multimedia Tools and Applications, 2024, 83(4): 10211-10231. https://doi.org/10.1007/s11042-023-16024-2

[3] Guo X, Xu Y, Sun J, et al. Dynamic Graph Tempor al-Frequency Dual-Channel Network for Multi-ban d Spectrum Prediction[J]. IEEE Communications L etters, 2024. https://doi.org/10.1109/lcomm.2024.3 451536

[4] Hesse C, Löf S. Self-supervised learning of musical representations using VICReg; a comprehensive study of the VICReg loss function for self-supervised representation learning in the music domain[J]. 2023. https://doi.org/10.31219/osf.io/tvmdu

[5] Akama T, Kitano H, Takematsu K, et al. Auxiliary self-supervision to metric learning for music similarity-based retrieval and auto-tagging[J]. Plos one, 2023, 18(11): e0294643. https://doi.org/10.1371/journal.pone.0294643

[6] Ashraf M, Abid F, Din I U, et al. A hybrid cnn and rnn variant model for music classification[J]. Applied Sciences, 2023, 13(3): 1476. https://doi.org/10.3390/app13031476

[7] Prabhakar S K, Lee S W. Holistic approaches to music genre classification using efficient transfer

and deep learning techniques[J]. Expert Systems with Applications, 2023, 211: 118636. https://doi.org/10.1016/j.eswa.2022.118636

[8] Mao Y, Zhong G, Wang H, et al. Music-CRN: An efficient content-based music classification and rec ommendation network[J]. Cognitive Computation, 2022, 14(6): 2306-2316. https://doi.org/10.1007/s12559-022-10039-x

[9] Liu Z, Bian T, Yang M. Locally activated gated neural network for automatic music genre classification[J]. Applied Sciences, 2023, 13(8): 5010. https://doi.org/10.3390/app13085010

[10] Liu C, Feng L, Liu G, et al. Bottom-up broadcast neural network for music genre classification[J]. Multimedia Tools and Applications, 2021, 80: 7313-7331. https://doi.org/10.1007/s11042-020-09643-6

[11] Gong T. Deep Belief Network-Based Multifeature Fusion Music Classification Algorithm and Simulation[J]. Complexity, 2021, 2021(1): 8861896. https://doi.org/10.1155/2021/8861896

[12] Rawat P, Bajaj M, Vats S, et al. A comprehensive study based on MFCC and spectrogram for audio classification[J]. Journal of Information and Optimization Sciences, 2023, 44(6): 1057-1074. https://doi.org/10.47974/jios-1431

[13] Motoki K, Takahashi N, Velasco C, et al. Is classical music sweeter than jazz? Crossmodal influences of background music and taste/flavour on healthy and indulgent food preferences[J]. Food Quality and Preference, 2022, 96: 104380. https://doi.org/10.1016/j.foodqual.2021.104380

[14] Han D, Kong Y, Han J, et al. A survey of music emotion recognition[J]. Frontiers of Computer Science, 2022, 16(6): 166335.

[15] Sharma A K, Aggarwal G, Bhardwaj S, et al. Classification of Indian classical music with time-series matching deep learning approach[J]. IEEE access, 2021, 9: 102041-102052. https://doi.org/10.1109/access.2021.3093911

[16] Li C, Li F, Zhang L, et al. Intrusion detection for industrial control systems based on improved contrastive learning SimCLR[J]. Applied Sciences, 2023, 13(16): 9227. https://doi.org/10.3390/app13169227

[17] Yang M, Wang Z, Yan Z, et al. DNASimCLR: a co ntrastive learning-based deep learning approach for gene sequence data classification[J]. BMC bioinfo rmatics, 2024, 25(1): 328. https://doi.org/10.1186/s12859-024-05955-8

[18] Li L, Kang R, Huang W, et al. A Study on Small-Scale Snake Image Classification Based on Improved SimCLR[J]. Applied Sciences, 2025, 15(11): 6290. https://doi.org/10.3390/app15116290

[19] Pöppelbaum J, Chadha G S, Schwung A. Contrastive learning based self-supervised time-series analysis[J]. Applied Soft Computing, 2022, 117: 108397. https://doi.org/10.1016/j.asoc.2021.108397

[20] Zhang L, Han F, Li T, et al. MLR-SimSiam: A Contrastive Pre-training Model based on Polarimetric Jittering and Mutual Learning Regularizer for PolSAR Image Classification[J]. IEEE Geoscience and Remote Sensing Letters, 2024. https://doi.org/10.1109/lgrs.2024.3387663

[21] Huang, H., Liu, W., & Zhang, J. Reliable Service Node Set Selection and Task Offloading Strategy in Edge-Enabled Robot Swarms via Dynamic Interference and Link Reliability Models[J]. Informatica, 2025, 49(32). https://doi.org/10.31449/inf.v49i32.9057

[22] Wang, H. Innovative Application of Intelligent Mechanical Manufacturing Based on Self-Supervised Learning and Graph Neural Network Fusion Optimization[J]. Informatica, 2025, 49(17). https://doi.org/10.31449/inf.v49i17.7595

[23] Huang P, Liu Y, Xie S, et al. Uncertainty-Aware S elf-Supervised Cross-Modal SAR-Optical Matchin g Using EfficientDet and Xception[J]. Informatica, 2025, 49(25). https://doi.org/10.31449/inf.v49i25.8421

[24] Setiorini E, Widjaja M, Wicaksana A. Reduced Convolutional Recurrent Neural Network Using MFCC for Music Genre Classification on the GTZAN Dataset[J]. Informatica, 2025, 49(17). https://doi.org/10.31449/inf.v49i17.6885

[25] A.S, S., J.B, M., & Rajan, R. Automatic music mood classification using multi-modal attention framework[J]. Engineering Applications of Artificial Intelligence, 2024, 128: 107355. https://doi.org/10.1016/j.engappai.2023.107355

[26] Chen, M., Tang, D., Xiang, Y., Shi, L., Tuncer, T., Ozyurt, F., & Dogan, S. Instrument sound classification using a music-based feature extraction model inspired by Mozart's Turkish March pattern[J]. Alexandria Engineering Journal, 2025, 118: 354-370. https://doi.org/10.1016/j.aej.2025.01.059

[27] Huang, P., Liu, Y., Xie, S., An, Y., & Liang, Y. Uncertainty-Aware Self-Supervised Cross-Modal SAR-Optical Matching Using EfficientDet and Xception[J]. Informatica, 2025, 49(25). https://doi.org/10.31449/inf.v49i25.8421

[28] Liuwanyue, S. Course genres classification of music e-learning platform based on deep learning big data intelligent processing algorithm[J]. Entertainment Computing, 2024, 50: 100704. https://doi.org/10.1016/j.entcom.2024.100704

[29] Oubrahim, Z., Amirat, Y., Ouassaid, M., & Benbouzid, M. Classification of complex power quality disturbances under noisy environment using Root-Music and least square techniques[J]. Electric Power Systems Research, 2025, 238: 111030. https://doi.org/10.1016/j.epsr.2024.111030

[30] Qin, Y., & Miao, D. Research on Vocal Music Iden tification and Classification Based on Energy Entro py Ratio and AlexNet Model[J]. International Jour nal of Cognitive Informatics and Natural Intelligen ce, 2025, 19(1). https://doi.org/10.4018/ijcini.386839

[31] Raghavan, K., Avila, M. L., Balaprakash, P., Jayatissa, H., & Santiago-Gonzalez, D. Classification of events from α-induced reactions in the MUSIC detector via statistical and ML methods[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators,

Spectrometers, Detectors and Associated Equipment, 2024, 1058: 168786. https://doi.org/10.1016/j.nima.2023.168786

[32] Thao, H. T. P., Roig, G., & Herremans, D. EmoMV: Affective music-video correspondence learning datasets for classification and retrieval[J]. Information Fusion, 2023, 91: 64-79. https://doi.org/10.1016/j.inffus.2022.10.002