



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Investigation of a music genre data classification method based on an improved ECAPA-TDNN algorithm

Dongyan Li, Shuhan Du

DOI: [10.1504/IJICT.2026.10076436](https://doi.org/10.1504/IJICT.2026.10076436)

Article History:

Received:	24 September 2025
Last revised:	10 November 2025
Accepted:	11 November 2025
Published online:	02 March 2026

Investigation of a music genre data classification method based on an improved ECAPA-TDNN algorithm

Dongyan Li and Shuhan Du*

School of Early Childhood Education,
Cangzhou Preschool Teachers College,
Cangzhou, Hebei, 062150, China
Email: 2628613121@qq.com
Email: 18630765969@163.com

*Corresponding author

Abstract: To address the issue of low accuracy in music genre data classification, this study proposes an improved model based on the integrated Res2Net module extended context-aware parallel aggregation-time delay neural network (ECAPA-TDNN). The study adopts the ‘pro-WGAN’ strategy to balance the fma_cedum dataset, and generates feature data through nested pro-WGAN loops for undersampling schools with more samples and schools with fewer samples. The results show that the improved ECAPA-TDNN model has a classification accuracy of 89%, which is 11.6% higher than the original ECAPA-TDNN model. The training time is only 32.22 seconds, and data balancing processing significantly improves classification performance. The results provide an efficient solution for music type classification.

Keywords: ECAPA-TDNN algorithm; music genre; data classification method; ResNet architecture; BiLSTM; pro-WGAN; multifeature fusion.

Reference to this paper should be made as follows: Li, D. and Du, S. (2026) ‘Investigation of a music genre data classification method based on an improved ECAPA-TDNN algorithm’, *Int. J. Information and Communication Technology*, Vol. 27, No. 13, pp.38–60.

Biographical notes: Dongyan Li graduated from Hebei Normal University and is currently an Associate Professor and the Director of Art Teaching and Research Section at Cangzhou Preschool Teachers College. Her research focuses on music education, vocal performance, and the performance of children’s songs. She has led or participated in seven provincial-level research projects, published 26 papers.

Shuhan Du earned her Bachelor’s degree from Southwest University and Master of Education from City University Malaysia. She currently teaches at Cangzhou Preschool Teachers College. Working in music education, her research interests include music pedagogy and vocal performance. She has led and participated in provincial and municipal research projects and has published multiple articles. She has also guided students to win awards in provincial and municipal skills competitions, demonstrating strong teaching outcomes.

1 Introduction

In the intersection of audio signal processing and machine learning, automatic music genre classification represents a core task. Its outcomes are widely applicable to music recommendation systems, audio retrieval, and copyright monitoring, directly influencing user experience and industry efficiency. Consequently, the development of related technologies has remained a key focus in the field (Yang et al., 2025). From a technological perspective, music genre classification has evolved through two main stages: traditional machine learning and deep learning. Traditional methods rely on manually designed features and a deep understanding of music theory and audio structures. Although effective for simple and structured data, these methods exhibit limited generalisation ability and classification accuracy when handling diverse and complex music styles.

With the rise of deep learning, recurrent neural network (RNN), convolutional neural network (CNN), attention-based models, and other architectures have been applied to this task. Models such as the bidirectional broadcast neural network (BBNN) have achieved performance breakthroughs on specific datasets. However, some models still suffer from parameter redundancy, long training times, and incomplete capture of long-term audio features. In-depth analysis of the performance of existing mainstream deep learning models in music classification tasks reveals significant limitations in capturing the complex temporal dynamic features of music signals.

Firstly, CNN models excel at extracting local spatial features from spectrograms, but their inherent translational invariance assumption and limited receptive field make it difficult to effectively model the melodic lines and rhythm patterns that evolve over time in music. Secondly, although RNNs and their variants (such as LSTM and GRU) theoretically have the ability to handle sequential data, they are often plagued by gradient vanishing or exploding problems in practical training. This results in insufficient modelling ability for long-range temporal dependencies in long sequences. In addition, although attention mechanisms can highlight the importance of keyframes through weighting, simple forward or self-attention mechanisms often ignore the inherent bidirectional contextual characteristics of music signals (i.e., the current note features are both influenced by past phrases and predict future melodic trends). The above limitations collectively lead to the difficulty of further improving the classification accuracy and generalisation ability of existing models when facing music works with complex structures and strong temporal dependencies.

Meanwhile, the extended context-aware parallel aggregation-time delay neural network (ECAPA-TDNN) algorithm has gradually emerged in the audio processing domain. By integrating channel attention mechanisms and residual network with multi-scale feature representation (Res2Net) multi-scale structures, ECAPA-TDNN demonstrates certain advantages in audio type recognition. Nevertheless, existing studies indicate that its feature extraction capability remains insufficient when processing music, which contains complex dynamic characteristics such as melody and rhythm, and it does not fully address the bidirectional past-future' information requirements of musical signals (Wang et al., 2025). Among them, there are literature supporting it. Zhang (2021) clearly point that ECAPA-TDNN only relied on one-way time delay layer modelling and could not effectively utilise the backward temporal information of music signals. This resulted in insufficient classification accuracy for music genres with complex rhythm changes (such as jazz and fusion music). The classification accuracy of jazz music on the

fma_medium dataset was 8.3% lower than that of the BiLSTM model. In addition, mainstream music datasets, such as fma_medium, generally exhibit class imbalance, further affecting model training performance and classification stability.

Based on these considerations, this study investigates an improved ECAPA-TDNN algorithm for music genre classification, with three main innovations.

- 1 In terms of architecture, a bidirectional long short-term memory (BiLSTM) network is embedded into ECAPA-TDNN. By leveraging BiLSTM's bidirectional information processing capability, the model compensates for the reliance on unidirectional information in traditional models and enhances the capture of temporal features in music signals.
- 2 Regarding data processing, a combined strategy of undersampling and pro-wasserstein generative adversarial network (pro-WGAN) feature generation is adopted. It aims to address the class imbalance in the fma_medium dataset, with iterative nested synthesis improving the quality of data for underrepresented genres.
- 3 In terms of mechanism optimisation, the synergy between channel attention and multi-head attention in ECAPA-TDNN is optimised. Additionally, multi-feature fusion is combined with principal component analysis (PCA) for dimensionality reduction, improving the efficiency of extracting key dynamic features from music.

2 Literature review

Regarding research on the ECAPA-TDNN algorithm, Chen et al. (2023) summarised, compared, and analysed the main research methods in current domestic and international speaker recognition. They proposed an improved ECAPA-TDNN algorithm, and experimental results demonstrated that the improved algorithm outperformed classical approaches in terms of accuracy and loss.

Desplanques et al. (2020) inspired by recent research findings, developed a network model based on an ECAPA-TDNN architecture. The improvement strategy involved integrating a convolutional block attention module with a multi-head attention mechanism.

Wang et al. (2023a) analysed the characteristics of Mel spectrograms and explored the use of deep learning-based ECAPA-TDNN methods for audio classification, with experiments confirming their effectiveness. It was found that deep learning-based ECAPA-TDNN methods using Mel spectrograms as features achieved high recognition accuracy (Trejos et al., 2025). Additionally, Wang et al. (2023a) proposed two CRET models and compared them with baseline models, including the Res2Net multi-scale backbone and ECAPA-TDNN, across different channels and datasets. The experimental results demonstrated that the proposed models exhibited strong performance in various experiments conducted on both training and testing sets. Meanwhile, Trejos et al. (2025) proposed a method framework for integrating and analysing different feature types, and used it as input for the well-known end-to-end encoder decoder attractor model, with a focus on time delayed neural network (ECAPA-TDNN) embedding and Geneva minimum acoustic parameter set.

In addition, Li et al. (2025) argued that speaker recognition played an increasingly important role in today's society. Currently, neural networks are widely used to extract

speaker features. Although the ECAPA-TDNN model that emphasises channel attention, propagation, and aggregation can obtain temporal context information to some extent through dilated convolution, this model has shortcomings in obtaining comprehensive speech features. To further improve the accuracy of the model, better capture temporal context information, and prevent ECAPA-TDNN from being affected by small frequency domain offsets, based on the ECAPA-TDNN model, a new CRET model is developed by combining two-dimensional convolutional networks (Conv2D), residual networks (ResNet), and ECAPA-TDNN.

Afterwards, Liu et al. (2025) compared the effectiveness of several mainstream attention mechanisms in the field of computer vision on the ECAPA-TDNN model through experiments. Next, their research focused on the substantial improvements brought by collaborative attention (CA) to the ECAPA-TDNN model. The introduction of CA could help the model embed time-frequency information into the channel representation. Even without using the AS-Norm model, the proposed model achieved a relative reduction of approximately 5.3% equal error rate and 5.5% minimum detection cost function on the Voxceleb-O and Voxceleb-H test sets compared to the ECAPA-TDNN baseline model. Moreover, Spandan and Goutam (2023) used the ECAPA-TDNN architecture and 40-dimensional MFCC to train a primary dialect classifier from training data of three dialects. They proposed a confidence evaluation algorithm and combined the TDNN performance of end-to-end and embedding extractor methods. Afterwards, they defined the semi open set scene box as a constrained optimisation problem. Through solving it, they ultimately proved that the performance degradation caused by unknown statements was minimised when the corresponding Softmax predictions in the target output were equally confused.

Then, Medetov et al. (2025) adopted the ECAPA-TDNN neural network architecture to construct a broadcaster speech model. Experimental studies showed that compared to a universal model based on all speech sounds, a speech model based solely on one model sound provided higher accuracy in identifying broadcasters. They also found that different phonemes provided different recognition accuracies for broadcasters. For example, in the case where the duration of the speech signal was 2–3 seconds, the accuracy of the generalised model for identifying the identity of the broadcaster was 75%. And using a model based solely on one phoneme ‘E’, with the same input data, the accuracy of broadcaster recognition was 85%, which was 10% higher than that of the generalised model.

Related to this, Kang et al. (2024) designed an online discussion activity level evaluation system based on voiceprint recognition technology. The application system developed for the research project was divided into two parts. The first part was to use audio segmentation technology to divide online discussion audio into multiple independent audio segments, and train a voiceprint recognition model to predict the identity of the speaker in each individual audio component. In the second part, they proposed an ECAPA-TDNN model for calculating the activity level of online conferences based on modified primary indicators. It used the frequency and total speaking time of each participant as the main indicators for activity level evaluation through traversal and calculation.

Regarding research on music genre data classification, Zaman et al. (2023) indicated that traditional classification methods typically rely on a deep understanding of music theory and audio structures, often involving the analysis of elements such as melody, rhythm, and harmony. These methods proved effective for relatively simple and

structured audio data. In addition, Chaudhury et al. (2022) combined Apache Spark with machine learning algorithms to study music genres and feature signals in the GTZAN Music Genre Dataset. They applied multiple scalable machine learning algorithms for genre classification and compared the performance of these classifiers, identifying random forest as the most effective classifier.

Nanni et al. (2021) evaluated and compared various acoustic and visual features, performing classification using a support vector machine (SVM) based on a fusion of these features, which improved accuracy. They also integrated multiple classifiers into an ensemble, retraining five pre-trained CNN models using six data augmentation techniques and four signal representations. Copiaco et al. (2021) proposed the use of spectral-temporal features and studied the performance of various architectural and parameter modifications applied to the original network. A new explainable deep learning model was developed by Zinemanas et al. (2021) for automatic sound recognition. It based prediction interpretation on the similarity of inputs to learned latent prototypes, which significantly boosted classification performance.

Furthermore, Maccagno et al. (2021) proposed a Mel spectrogram-based architecture for input audio frames, demonstrating high-precision classification in environmental sound classification tasks. Zhang (2021) transformed audio signals into spectrograms as a unified representation and proposed a convolutional RNN for music genre classification. Their strategy outperformed traditional handcrafted models and machine learning-based methods in terms of parameter efficiency and training time. Similarly, Yu et al. (2020) focused on attention-based music genre classification models, implementing both serial and parallel attention mechanisms, with parallel attention models showing superior performance. Gong et al. (2021) introduced the audio spectrogram transformer, a fully attention-based audio classification model without convolutional layers, achieving 98.1% accuracy in sound classification. In the same year, Liu et al. (2021) proposed a BBNN model, which primarily utilised an improved ‘tri-structure’ module to fully exploit the fundamental information of Mel spectrograms for classification decisions. With an accuracy of 93.7%, the BBNN set a new benchmark on the GTZAN Music Genre Dataset.

Wang et al. (2023a) studied and analysed the characteristics of Mel spectrograms, explored the application of deep learning ECAPA-TDNN method in audio type recognition, and verified its effectiveness through experiments. The experimental results showed that the deep learning ECAPA-TDNN audio type recognition method based on Mel spectrogram features had high recognition accuracy. In terms of classification research, Zhao and Liang (2024) proposed the hierarchical cascade fusion (HCF)-TDNN model by adding an HCF module to the ECAPA-TDNN model for sound event classification. In the HCF module, the audio signal was first converted into two-dimensional time-frequency features for segmentation. Then, the segmented features were convolved one by one to improve the small receptive fields that perceive details. Finally, after completing the convolution, the two adjacent parts were combined together and the next convolution was performed to expand the receptive field when capturing large targets. The improved model further enhanced scalability by emphasising channel attention and effective propagation and aggregation of feature information.

Afterwards, Shankar et al. (2025) argued that significant progress had been made in the field of audio analysis, revolutionising the processing of raw audio signals and integrating machine learning for sound classification. They utilised the Open AI Whisper model for transcription and ECAPA-TDNN architecture for speaker embedding, and

grouped the extracted embeddings using condensed hierarchical clustering. This process revealed complex speaker relationships and promoted meticulous fragment tagging. The research suggested new avenues for exploration in the dynamic field of audio analysis. Wang et al. (2023b) designed a PWPE-ECA-Res2Net TDNN model for speaker recognition to more effectively extract speaker feature embeddings in cross-scenario applications. Research showed that compared to ECAPA-TDNN models using MFCC features, ECAPA-TDNN models based on PWPE performed better in cross scene recognition accuracy, demonstrating stronger robustness and better noise resistance. In addition, even under the highest recognition rate conditions, the model could maintain a relatively short recognition time.

Overall, it can be seen that in the field of exploring ECAPA-TDNN algorithm, the current situation presents a multidimensional exploration trend. On one hand, numerous studies have been devoted to algorithm improvement, such as improvement strategies covering fusion convolutional block attention modules, multi head attention mechanisms, combined with two-dimensional convolutional networks, and residual networks. These improvements have achieved results in improving model accuracy, capturing temporal context information, enhancing noise resistance and robustness, and have also been applied to different scenarios such as speaker recognition, audio type recognition, sound event classification, etc., and their effectiveness has been verified. On the other hand, in research related to music genre data classification, traditional methods rely on music theory and audio structure analysis. However, in recent years, machine learning algorithms have been widely applied, and various models and architectures have emerged, such as CNN, recursive neural networks, and models based on attention mechanisms, which have achieved significant results in improving classification performance.

Through a systematic review of the above literature, it can be found that although BBNN, AST, and various hybrid models have achieved remarkable classification accuracy on specific datasets such as GTZAN, these advanced music classification methods still have some common and unresolved challenges. Firstly, in terms of temporal modelling, most models do not fully capture the bidirectional contextual dependencies in music signals over long periods of time, either lacking explicit temporal modelling like CNN or struggling to effectively learn long sequence dependencies like standard RNN/LSTM. Secondly, regarding model structure and efficiency, complex network architectures designed for high performance, like deep CNNs and RNNs in series, often come with huge parameter quantities and computational overhead. This results in long model training time and difficulty in rapid iteration and application deployment. Thirdly, at the data level, the vast majority of studies assume that the training data is class balanced, lacking systematic processing strategies and robustness evaluations for the uneven distribution of music data in the real world. This often results in poor classification performance of models on a few genres.

This study aims to address these three core limitations through collaborative innovation and improvement from three dimensions: data balancing (pro-WGAN), model structure (BiLSTM embedding), and feature processing mechanism (attention collaborative optimisation), thus constructing a more efficient and robust music classification model. This study is based on the research of music type data classification method using improved ECAPA-TDNN algorithm. In response to the problem of insufficient capture of complex music features by ECAPA-TDNN, the collaborative

mechanism of channel attention and multi head attention is optimised to enhance the ability to extract key dynamic features such as music melody and rhythm.

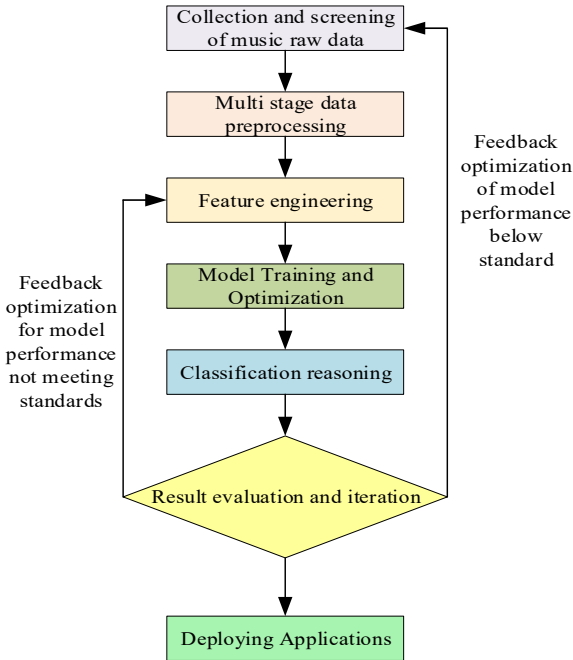
3 Design of a music genre data classification method based on the improved ECAPA-TDNN algorithm

3.1 Basic process and common methods of music genre data classification

The classification of music genre data constitutes a key problem in both audio signal processing and machine learning. It essentially involves analysing multidimensional features of music signals to establish a mapping between audio features and music genres, enabling automatic identification of unknown musical styles. This technology is widely applied in music recommendation systems, such as personalised playlist generation on Spotify, audio retrieval (e.g., filtering music libraries by genre), and copyright monitoring (e.g., identifying the genre of potentially infringing music). Its performance directly affects both user experience and industry efficiency.

From a technical perspective, the basic process of music genre data classification can be divided into six key stages: collection and selection of raw music data → multi-stage data preprocessing → feature engineering → model training and optimisation → classification inference → result evaluation and iteration. Each stage is closely interconnected, with feedback loops between them, and any omission can significantly reduce classification accuracy. The overall structural framework is illustrated in Figure 1.

Figure 1 Complete process and module interaction structure of music genre data classification



During the data collection and selection stage, both ‘data quality’ and ‘class balance’ must be considered. Priority is given to publicly available authoritative datasets or data obtained through legally compliant music platform application programming interfaces to ensure that audio files are intact and genre labels are accurate. Simultaneously, sample size differences among categories are controlled (the largest category contains no more than twice the number of samples as the smallest) to prevent the model from being biased toward majority classes.

In the preprocessing stage, format standardisation eliminates feature distortion caused by differing encoding formats. Noise suppression is performed using wavelet denoising, which preserves signal details in both time and frequency domains and performs better in complex noise scenarios. Its mathematical expression is given in equation (1). Audio segmentation uses a sliding window method combined with silence removal and padding rules to ensure a high proportion of effective data. Data augmentation applies diverse time-domain and frequency-domain operations to simulate real-world distortions, improving model generalisation.

$$W_{ij}^{denoised} = \begin{cases} \text{sing}(W_{ij}) \max(|W_{ij}| - \lambda, 0) & |W_{ij}| \geq \lambda \\ 0 & |W_{ij}| \leq \lambda \end{cases} \quad (1)$$

In equation (1), W_{ij} represents the wavelet coefficients, λ is an adaptive threshold calculated using Stein’s unbiased risk estimate (SURE), and $W_{ij}^{denoised}$ denotes the denoised wavelet coefficients.

Feature engineering serves as the bridge between raw audio and classification models, with its core objective being the extraction of discriminative features that differentiate music genres. Features across different dimensions exhibit distinct characteristics: time-domain features have low computational complexity ($O(N)$, where N is the number of sampling points, but limited discriminative power, making them suitable only for simple filtering tasks. Frequency-domain features have moderate computational complexity ($O(N \log N)$, including Fourier transforms, and can characterise timbre and harmonic distribution, aligning with human auditory perception. They are the mainstream choice; for instance, Mel-frequency cepstral coefficients (MFCC) simulate human auditory perception through Mel filter banks. The spectral centroid, calculated as in equation (2), represents the ‘centre of gravity’ of the spectrum, with rock music generally exhibiting higher spectral centroids than classical music, making it a key distinguishing metric:

$$\text{Spectral Centroid} = \frac{\sum_{k=1}^K k |X(k)|^2}{\sum_{k=1}^K |X(k)|^2} \quad (2)$$

In equation (2), k is the frequency index, $X(k)$ is the Fourier transform of the audio signal, and K is the total number of frequency points. Perceptual features, which integrate psychoacoustic models, have higher computational complexity but strong discriminative power, suitable for fine-grained classification.

In practical applications, a multi-feature fusion strategy is often adopted. For example, a 16-dimensional MFCC vector, combined with first- and second-order derivatives, forms a 48-dimensional feature set. When augmented with spectral centroid, bandwidth, and beats per minute (BPM), a 51-dimensional feature vector is obtained.

PCA is then applied to reduce dimensionality to 32, preserving over 95% of the information while reducing computational cost and avoiding the curse of dimensionality. The selection of the dimensionality reduction threshold is based on multiple sets of control variable experiments. The experiment tests the performance of the model when dimensionality is reduced to 16, 24, 32, 40, and 64 dimensions using the *fma_medium* dataset as a sample. The results show that when the dimensionality is reduced to 32 dimensions, the model accuracy reaches 82.1%, an increase of 5.8% compared to 76.3% in 16 dimensions, an increase of 2.6% compared to 79.5% in 24 dimensions, a decrease of only 0.7% compared to 82.8% in 40 dimensions, and a decrease of only 1.4% compared to 83.5% in 64 dimensions.

In terms of training efficiency, a single round of training for 32 dimensional features takes 1.2 seconds, which is 20% less than the 1.5 seconds for 40 dimensional features and 42.9% less than the 2.1 seconds for 64 dimensional features. In terms of storage efficiency, the 32 dimensional features only occupy 128 bytes per sample, which is 50% less than the 64 dimensional 256 bytes and significantly reduces memory consumption. Meanwhile, the study also compares PCA with other mainstream dimensionality reduction methods. Although linear discriminant analysis can optimise feature projection using category information, it is sensitive to data distribution assumptions and performs poorly in the non-Gaussian distribution features of the *fma_medium* dataset. Its accuracy drops to 78.5%, which is 3.6% lower than PCA. The t-SNE method performs well in feature visualisation and can clearly distinguish feature clusters from different schools of thought. However, its computational complexity is high. When reducing the dimensionality of 51 dimensional features, the single sample processing time reaches 0.8 milliseconds, which is 7 times longer than PCA's 0.1 milliseconds. Moreover, its generalisation ability is weak, and its accuracy on the test set decreases by 12.3% compared to the training set, while PCA only decreases by 3.5%.

Although autoencoder methods can adaptively extract key features through deep learning, they require additional neural network training processes, have high training costs, and are prone to overfitting in small sample scenarios, with an accuracy rate of less than 75%. Taking into account the balance of accuracy, efficiency, and generalisation, PCA dimensionality reduction to 32 dimensions is ultimately chosen as the optimal solution for this study. This choice significantly improves the training and storage efficiency of the model while ensuring classification accuracy, and maintains good generalisation performance. PCA performs dimensionality reduction through the covariance matrix and eigenvalue decomposition as expressed in equation (3):

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \quad (3)$$

In equation (3), Σ is the covariance matrix, x_i represents sample feature vectors, $\bar{x}$ is the sample mean vector, n is the number of samples, and T denotes matrix transpose.

Feature optimisation includes normalisation and standardisation. Min-Max normalisation, given in equation (4), maps feature values to the $[0, 1]$ interval and is suitable for uniformly distributed features:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

In equation (4), $X_{\max}$ and $X_{\min}$ represent the maximum and minimum values of the feature, respectively.

Z-score standardisation, given in equation (5), transforms features to have zero mean and unit variance, suitable for scenarios with outliers:

$$x_{std} = \frac{x - \mu}{\sigma} \quad (5)$$

In equation (5), μ represents the mean of the feature, and σ denotes the standard deviation of the feature.

Model selection depends on data scale and accuracy requirements. Traditional machine learning methods focus on ‘manually designed features + simple models.’ SVM models identify the optimal hyperplane as expressed in equation (6), providing stability for small datasets. Random forests can automatically assess feature importance, while k-nearest neighbours (k-NN) is simple to implement but sensitive to distance metrics:

$$\min_{w,b} 1/2 \|w\|^2 + C \sum_{i=1}^n \xi_i, \text{ s.t. } y_i (w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (6)$$

In equation (6), w is the hyperplane normal vector, b is the bias, C is the penalty parameter, ξ_i denotes the slack variable, and y_i represents the sample label.

In the model training and optimisation stage, the dataset is divided in a 7:2:1 ratio. Hyperparameters are tuned using grid search, random search, or Bayesian optimisation, while cross-validation, early stopping, and regularisation are employed to prevent overfitting. During classification inference, the features of unknown music are input to the model, which outputs the class probability distribution via the Softmax function [equation (7)]. The label corresponding to the highest probability is selected as the classification result.

$$P_i = \frac{\exp(z_i)}{\sum_{j=1}^C \exp(z_j)} \quad (7)$$

In equation (7), z_i is the output of the model’s final layer, C is the number of music genres, p_i is the predicted probability for class i .

Accuracy, precision, recall, and F1-score are used as evaluation metrics, with their definitions provided in equations (8)–(11). The confusion matrix further assists in detecting categories prone to confusion. Error analysis informs iterative improvements, such as feature adjustment, model structure enhancement, or dataset expansion, forming a ‘training-evaluation-optimisation’ closed loop.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

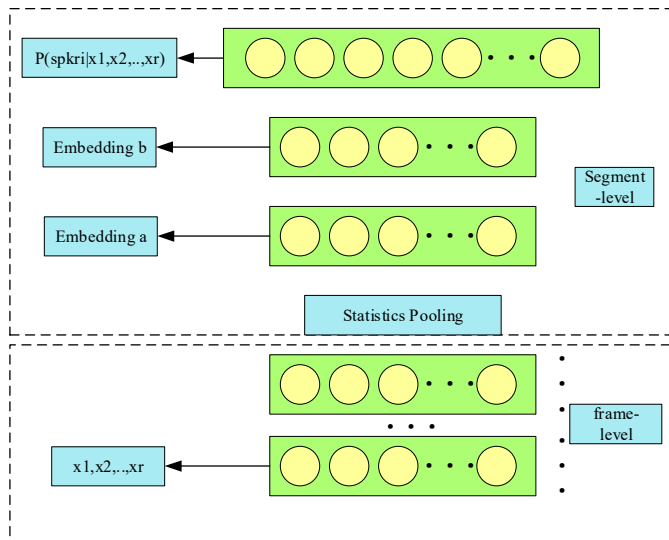
Here, TP denotes true positives, TN refers to true negatives, FP represents false positives, and FN signifies false negatives. Overall, traditional machine learning methods remain valuable in small datasets or low-computation scenarios, whereas deep learning methods, leveraging automatic feature learning, have become mainstream in large-scale, high-precision applications. The improved ECAPA-TDNN, through multi-module optimisation, demonstrates superior performance and generalisation capability in music genre classification tasks.

3.2 Construction of a music genre data classification model based on the improved ECAPA-TDNN algorithm

3.2.1 Data-driven ECAPA-TDNN model

To validate the effectiveness of data balancing, the ECAPA-TDNN model is trained on both balanced and unbalanced music datasets for comparison experiments. In 2020, Desplanques et al. (2020) developed the ECAPA-TDNN model, applying it to speaker recognition with audio features. This model was improved based on the x-vectors algorithm, incorporating the squeeze-and-excitation (SE) module and the channel attention mechanism from SE-Net. The x-vectors framework employs a statistics pooling layer to convert audio signals of arbitrary length into fixed-length feature vectors. The model architecture features multiple frame-level TDNN layers, followed by a statistics pooling layer, two fully connected layers, and a concluding Softmax layer. Figure 2 illustrates the model architecture.

Figure 2 Architecture of the x-vectors model (see online version for colours)



Traditional x-vectors consider only 15 frames during learning. In contrast, ECAPA-TDNN incorporates the SE module to capture global information of data features. Additionally, the model employs a Res2Net structure to achieve multi-scale feature representation. In conventional ResidualNetwork (ResNet) architectures, input information first passes through a 1×1 convolution to reduce feature channels, which limits attention to individual pixels without considering neighboring pixels. The output then undergoes a 3×3 convolution with downsampling, followed by a 1×1 convolution to restore the feature channels for residual connections, maintaining input-output feature dimensionality. Figures 3 and 4 illustrate the ResNet and Res2Net structures, respectively.

Figure 3 ResNet structure (see online version for colours)

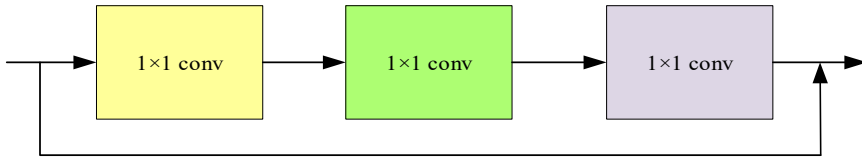
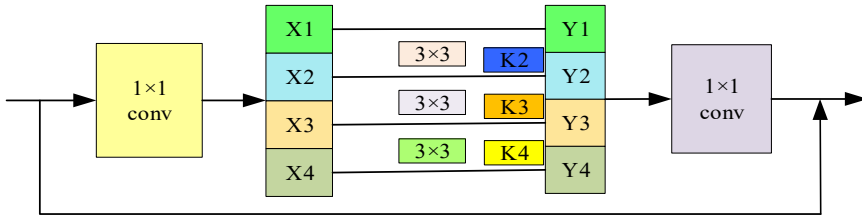


Figure 4 Res2Net structure (see online version for colours)



The Res2Net structure contains a residual connection and a split-transform-merge (STM) module, which is its core component. The STM module partitions the input feature map into submaps, processes each via a designated transformation, and then merges them to form the final output. Res2Net enhances overall classification performance and, together with techniques such as stochastic transformations and DropBlock regularisation, improves model robustness and generalisation.

In the ECAPA-TDNN model, Desplanques et al. (2020) adapted the Res2Net structure by replacing the original 2D convolutions with 1D convolutions. Convolutional kernels of size 3 are used as intermediate dilated kernels, with dilation rates of 2, 3, and 4. This design emphasises channel attention, information propagation, and aggregation. The SE-Res2Block structure and the overall ECAPA-TDNN model architecture are shown in Figures 5 and 6.

Figure 5 SE-Res2Block structure in ECAPA-TDNN (see online version for colours)

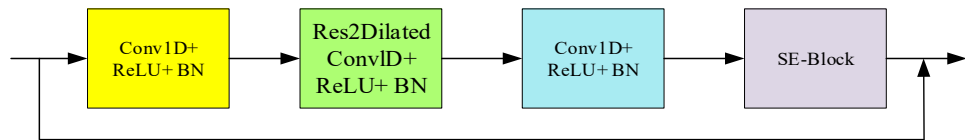
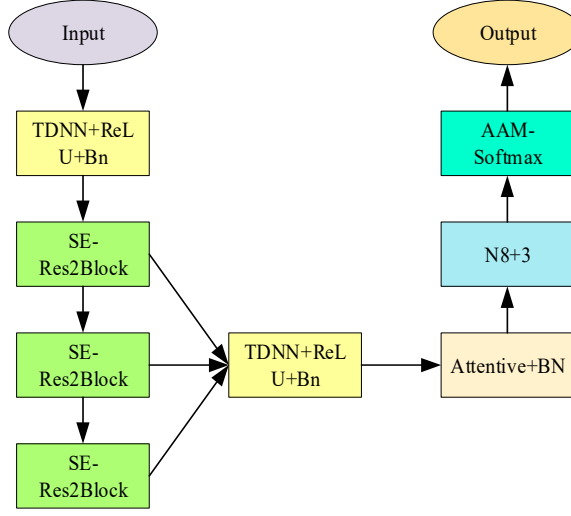


Figure 6 Overall ECAPA-TDNN model architecture (see online version for colours)

3.2.2 Model optimisation and improvement

The long short-term memory (LSTM) network is composed of three gates – input, forget, and output – along with a memory cell that maintains the memory state. The input gate governs the information entering the cell, the forget gate specifies what should be erased, and the output gate controls the information released. The memory cell state is controlled by these gates, determining when it should be updated or forgotten. The gating mechanisms are automatically learned by the network.

Compared to traditional RNN models, LSTM selectively propagates past and current information through gate control, mitigating gradient vanishing and exploding issues. This enables LSTM to handle long-sequence data more effectively. The LSTM structure can be formally expressed as follows:

- Forget gate:

$$f_t = \sigma(U_f h_{t-1} + W_f x_t) \quad (12)$$

$$k_t = c_{t-1} \odot f_t \quad (13)$$

- Input gate:

$$i_t = \sigma(U_i h_{t-1} + W_i x_t) \quad (14)$$

$$g_t = \tanh(U_g h_{t-1} + W_g x_t) \quad (15)$$

$$j_t = g_t \odot i_t \quad (16)$$

$$c_t = j_t + k_t \quad (17)$$

- Output gate:

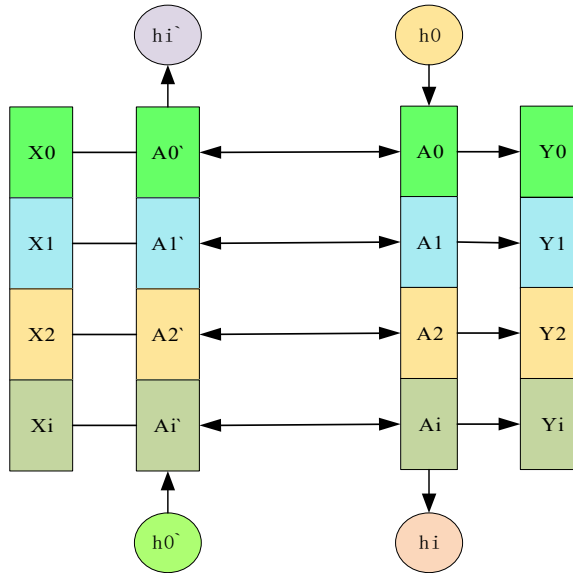
$$\sigma_t = \sigma(U_o h_{t-1} + W_o x_t) \quad (18)$$

$$h_t = \tanh(c_t) \odot o_t \quad (19)$$

In the above equations, the sigmoid function $\sigma(\cdot)$ maps values into the range $[0, 1]$, controlling gate openness (0 = fully closed, 1 = fully open). $\odot$ denotes element-wise multiplication. The forget gate f_t regulates the fraction of the prior cell state c_{t-1} to forget. U_f and W_f are the forget gate weight matrices for the previous hidden state h_{t-1} and current input x_t , respectively. k_t represents the residual memory of c_{t-1} after forgetting. i_t (input gate) controls how much current input enters the cell state. U_i and W_i are the weight matrices used to generate g_t for h_{t-1} and x_t , respectively. j_t (gated update) is the element-wise product of i_t and g_t , representing the portion of current input added to the cell state. c_t (updated cell state) is obtained by adding the retained memory k_t and the current update j_t . The output gate o_t regulates the extent to which the cell state is revealed to the hidden state. U_o and W_o are the output gate weight matrices for h_{t-1} and x_t . h_t (hidden state) is computed by applying $\tanh(\cdot)$ to c_t and multiplying element-wise with o_t .

BiLSTM is a variant of the standard LSTM architecture designed to capture both past and future contextual information in sequential data. Unlike conventional LSTM models that propagate information in a single temporal direction and thus can only utilise past context at a given time step, BiLSTM combines a forward-pass LSTM and a backward-pass LSTM to enable bidirectional information flow. This property is particularly advantageous for music audio signals, where both preceding and subsequent temporal information contribute to accurate feature representation. Figure 7 presents the architecture of the BiLSTM model.

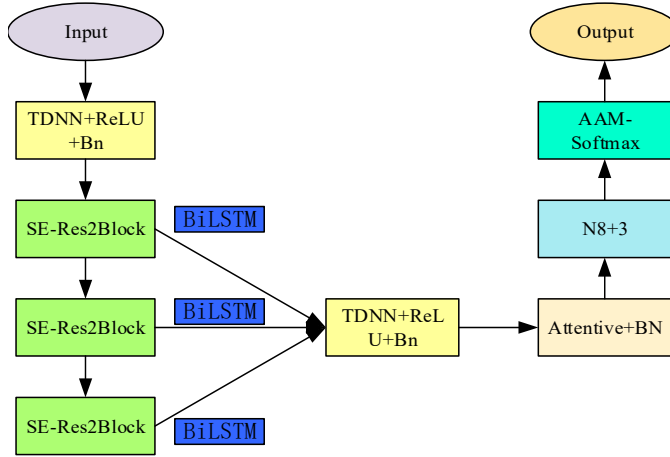
Figure 7 BiLSTM model structure (see online version for colours)



In the BiLSTM framework, the input sequence $(x_0, x_1, x_2, \dots, x_i)$ is first processed by the forward LSTM to generate forward hidden states. Simultaneously, the same sequence is processed in reverse by the backward LSTM, which produces backward hidden states. The forward and backward hidden states are then concatenated at each time step to produce the final output, integrating information from both directions.

To further enhance classification performance in music genre recognition, this study incorporates the BiLSTM module into the ECAPA-TDNN architecture, resulting in a novel improved model structure. The specific fusion strategy is to use a ‘serial connection’ approach: after the frame level TDNN layer of the ECAPA-TDNN network completes feature extraction, the output feature sequence (rather than the speaker embedding vector after statistical pooling) is used as the input of the BiLSTM module. The BiLSTM network then performs bidirectional forward and backward scanning on the sequence, fully capturing its long-term temporal contextual information. The hidden state of the last time step of BiLSTM (containing both forward and backward final contextual information) is extracted, and then concatenated with the global statistics (such as mean and standard deviation) of the feature sequence processed by SE-Res2Block in the ECAPA-TDNN framework and weighted by channel attention in the feature dimension. This enhanced feature vector combines local multi-scale spectral features (from ECAPA-TDNN) and global bidirectional temporal context information (from BiLSTM), and is finally fed into the statistical pooling layer and subsequent classification layer. This deep serial fusion method ensures that the model can simultaneously gain a deep understanding of music signals from two key dimensions: frequency domain and time domain.

Figure 8 Improved ECAPA-TDNN model structure (see online version for colours)



In terms of optimising the attention mechanism, this study addresses the problem of relatively single attention forms in the original structure of ECAPA-TDNN, and introduces the multi head self-attention (MHSA) mechanism, which is co-designed with the original channel attention (squeeze excitation, SE) module. This study does not simply parallelise or concatenate the two, but designs a lightweight ‘gate control collaborative fusion’ module. This module receives channel weighted features output by the SE module and time frame relationship weighted features output by the MHSA module. Then, a learnable scalar weight parameter α (generated by a fully connected layer and Sigmoid activation function, with values between 0 and 1) adaptively determines the contribution of channel attention and temporal attention in the final feature representation. The final feature output can be obtained from equations (20) and (21):

$$F_{fused} = \alpha \times F_{se} + (1 - \alpha) \times F_{mhsa} \quad (20)$$

$$\alpha = \sigma(W \times [F_{se} \parallel F_{mhsa}] + b) \quad (21)$$

F_{mhsa} is the time frame relationship weighted feature output by the MHS module, α is the learnable gating weight parameter with a range of $[0,1]$, F_{se} is the channel weighted feature output by the SE module, σ represents the Sigmoid activation function, W and b are the weight matrix and bias vector of the fully connected layer, and $\parallel$ represents the feature concatenation operation.

This design allows the model to dynamically adjust its attention ratio to spectral channel importance and temporal frame importance based on the specific characteristics of the input music segment, thereby achieving more refined and adaptive efficient capture of music dynamic features. Figure 8 presents the modified ECAPA-TDNN with BiLSTM integration.

4 Experimental results of the music type data classification method based on the improved ECAPA-TDNN algorithm

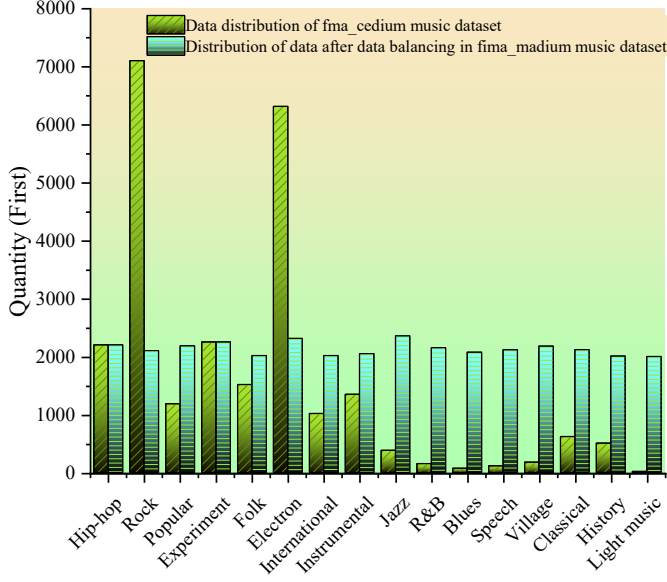
4.1 Selection of experimental test data samples

To address the imbalance in the distribution of genre samples within the `fma_medium` music dataset, this study employs a combined strategy of undersampling and pro-WGAN-based feature generation for data balancing. For genres with a large number of samples, such as rock and electronic music, undersampling is applied by randomly selecting 2,300 rock samples and 2,000 electronic samples. For genres with a moderate number of samples, such as experimental music, the original sample size is maintained. For genres with relatively few samples, new genre-constrained music data are generated using the pro-WGAN feature generation network to expand the dataset.

In particular, for extremely scarce genres, such as light music and blues, the initially generated samples are further used as new inputs for the generation network to enable recursive nested generation, thereby improving both the quality and diversity of the synthetic data. This strategy effectively alleviates sample distribution disparities across genres, ensuring that the balanced `fma_medium` dataset is more suitable for subsequent training and evaluation. Figure 9 illustrates the data distribution and balancing results. The pro-WGAN (progressive wasserstein generative adversarial network) used here includes the following key parameters and designs in implementation: the generator adopts a multi-layer perceptron structure with 3 fully connected layers, the hidden layer dimension is set to 512, ReLU is used as the activation function, and the output layer uses Tanh function to normalise the data to the $[-1,1]$ interval. The critic is also a 3-layer fully connected network with a hidden layer dimension of 512, using LeakyReLU as the activation function. The loss function adopts the standard Wasserstein loss and is supplemented by a gradient penalty (GP) term, where the gradient penalty coefficient λ is set to 10 to enforce the Lipschitz constraint. The optimiser uses Adam, with a learning rate set to 0.0001. The exponential decay rate for first-order moment estimation is $\beta_1 = 0.5$, and the exponential decay rate for second-order moment estimation is $\beta_2 = 0.9$. The model is trained for a total of 5,000 epochs, during which the discriminator is updated 5 times and then the generator is updated once, and the batch size is fixed at 64.

These parameters have been determined through preliminary grid search and can achieve a good balance between generating sample quality and training stability.

Figure 9 Data distribution of the fma_medium music dataset and data distribution after balancing (see online version for colours)



4.2 Testing and analysis of music genre classification models

4.2.1 Classification accuracy and training time

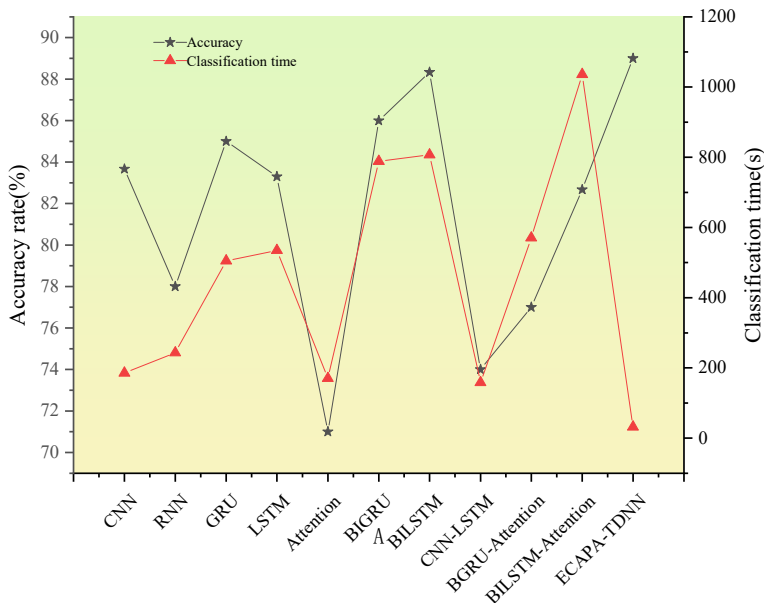
To ensure the fairness and reproducibility of all comparative experiments, this study establishes a unified experimental environment configuration and hyperparameter setting standard. In terms of hardware configuration, the experiment uses a single NVIDIA RTX 3090 GPU as the main computing unit, with a 24GB video memory capacity sufficient to support the video memory requirements during the training process of all comparison models and the improved model in this study. The central processing unit uses Intel Xeon Silver 4210, with a clock speed of 2.20GHz and 128GB DDR4 system memory, providing sufficient computing resource support for data preprocessing and model training. Regarding software environment, the experimental platform is built on Ubuntu 20.04 LTS operating system, and the deep learning framework adopts PyTorch 1.12.1 version, combined with CUDA 11.6 parallel computing architecture and cuDNN 8.4.0 acceleration library, to ensure efficient execution of the model training process. The programming environment is Python 3.8 and relies on scientific computing libraries such as NumPy 1.21.6 and SciPy 1.7.3 for data processing.

In terms of hyperparameter settings for model training, this study adopts a systematic configuration strategy. The optimiser uniformly uses the Adam algorithm, with parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, epsilon = $1e-8$, and an initial learning rate uniformly set to 0.001. The cosine annealing scheduling strategy is used to dynamically adjust the learning rate. The batch size is fixed at 32, which takes into account GPU

memory limitations while ensuring the stability of gradient estimation. The maximum number of training epochs is set to 100, and an early stopping mechanism is introduced to automatically terminate the training when the validation set loss does not improve for ten consecutive training epochs. This setting effectively prevents the occurrence of overfitting in the model. For model initialisation, all network parameters are initialised using the Xavier uniform initialisation method to ensure that the variance of inputs at each layer remains consistent, thereby accelerating model convergence. The loss function uses cross entropy loss and sets a Dropout ratio of 0.2 in the fully connected layer to enhance the model's generalisation ability. This study ensures that all comparative experiments are conducted under completely consistent conditions. Hence, the performance differences between different model architectures can truly reflect their inherent characteristic differences, rather than being caused by external factors, thereby ensuring the scientific and comparable results of the experiments.

To evaluate model performance, this study compares ten benchmark models – CNN, RNN, Gate Recurrent Unit (GRU), LSTM, Attention, Bidirectional Gated Recurrent Unit (BiGRU), BiLSTM, CNN-LSTM, BiGRU-Attention, and BiLSTM-Attention – against the proposed improved ECAPA-TDNN model. The classification accuracy and training time of each model are shown in Figure 10.

Figure 10 Comparison of model accuracy and model running time (see online version for colours)



As shown in Figure 10, the improved ECAPA-TDNN model achieves the highest classification accuracy of 89%, which is 7.55% higher than the average accuracy of 81.45% and 18% higher than the lowest-performing Attention model (71%). This demonstrates the superior classification performance of the proposed model. Moreover, the improved ECAPA-TDNN also exhibits the shortest training time of only 32.22 seconds, which is 14.19 times faster than the average training time of 457.32

seconds, and 32.15 times faster than the BiLSTM-Attention model, which has the longest training time of 1,035.88 seconds. The findings demonstrate that the proposed ECAPA-TDNN model achieves higher classification accuracy while also enhancing computational efficiency.

4.2.2 Verification of classification performance

To further validate the effectiveness of data balancing and denoising strategies, the original ECAPA-TDNN model was tested on three datasets: the unbalanced dataset, the dataset balanced with a conventional GAN, and the dataset balanced with pro-WGAN. Table 1 summarises the experimental results.

Table 1 Experimental comparison results

<i>Dataset</i>	<i>Original dataset</i>	<i>Balanced with GAN</i>	<i>Balanced with pro-WGAN</i>
MAP	0.51	0.71	0.77
Acc	65.9%	79.8%	82.4%
Macro-F1	0.63	0.75	0.80
AUC-ROC	0.89	0.92	0.94

As shown in Table 1, both balancing and denoising substantially improve classification performance. However, in datasets with imbalanced categories, accuracy and mean accuracy (MAP) may not fully reflect the true performance of the model across all categories, especially for a few categories. To this end, two evaluation metrics that are more sensitive to category imbalance are added: Macro-F1 score and area under the receiver operating characteristic curve (AUC-ROC). As shown in Table 1, when using the original imbalanced dataset, the Macro-F1 score of the model is only 0.63, and the AUC-ROC is 0.89. After using traditional GAN for data balancing, these two indicators improved to 0.75 and 0.92, respectively. After adopting the proposed pro-WGAN strategy for data balancing, the model achieves optimal overall performance, with Macro-F1 score further improved to 0.80 and AUC-ROC also reaching 0.94. The significant improvement of Macro-F1 indicates that the model has better classification consistency for all music genres, especially those with small sample sizes. The improvement of AUC-ROC proves that the model has stronger ability to distinguish true and false positive samples under different decision thresholds. This provides more comprehensive and solid multidimensional evidence for the effectiveness of the pro-WGAN data balancing strategy. On the dataset balanced with pro-WGAN, the model's mean average precision (MAP) increases by 0.252 compared with the original dataset and by 0.057 compared with the GAN-balanced dataset. Classification accuracy improves by 16.8% and 2.6%, respectively. These findings indicate that pro-WGAN-based data balancing achieves the best performance for music genre classification.

4.2.3 Analysis of ablation experiment

To quantitatively decompose and verify the respective contributions of the three core improvements proposed (BiLSTM embedding, attention mechanism optimisation, and pro-WGAN data balancing) to the final model performance, this designs and conducts a systematic ablation study. The experiment set up five different model configurations:

- 1 complete model (including all three improvements)
- 2 remove the BiLSTM module (retaining only the ECAPA-TDNN backbone and attention optimisation)
- 3 remove the attention collaborative optimisation mechanism (using only the SE attention of the original ECAPA-TDNN and concatenating it with BiLSTM)
- 4 remove data balancing strategy (train the complete model using the original imbalanced dataset)
- 5 the original ECAPA-TDNN model (as a performance baseline).

The experimental results are detailed in Table 2.

Table 2 Results of ablation experiment

<i>Model configuration</i>	<i>Accuracy (%)</i>	<i>Macro-F1</i>	<i>Training time (s)</i>
Complete model	89.0	0.80	32.22
Remove BiLSTM	83.5	0.72	28.91
Remove attention optimisation	85.2	0.75	30.45
Remove data balance (using the original dataset)	78.3	0.63	31.88
The original ECAPA-TDNN	78.0	0.62	35.10

The ablation experiment results clearly reveal the independent value of each improvement:

- 1 the embedding of BiLSTM brings about an accuracy improvement of about 5.5% and a Macro-F1 gain of 0.08, which directly proves its effectiveness in enhancing bidirectional temporal modelling of music signals
- 2 the optimisation of attention mechanism contributes about 3.8% accuracy improvement and 0.05% Macro-F1 gain, indicating its important role in improving feature discriminability
- 3 the impact of data balancing strategy is most significant, as its removal leads to a significant decrease of about 10.7% in accuracy and 0.17% in Macro-F1.

This highlights that solving the problem of class imbalance is a prerequisite for achieving high performance of the model. Meanwhile, the complete model achieves the shortest training time, indicating that while various improvements improving performance, structural optimisation actually increases computational efficiency. In summary, the ablation experiment strongly demonstrates the necessity and synergistic effect of the three core improvement measures proposed here.

5 Conclusions

This study addresses challenges in music genre classification by first handling the imbalance in the `fma_medium` dataset and then improving the ECAPA-TDNN model through the integration of BiLSTM and optimisation of the attention collaboration mechanism. At the level of data preprocessing, this study innovatively adopts a strategy

of combining undersampling with pro-WGAN loop nested generation to systematically address the issue of class imbalance in the *fma_medium* dataset. At the level of model architecture improvement, this study has made two core innovations to the ECAPA-TDNN algorithm. The improved ECAPA-TDNN model is validated through comparative experiments. The results demonstrate significant improvements in both classification accuracy and training efficiency. Data balancing further enhances performance. The main conclusions are as follows. First, the improved ECAPA-TDNN model achieves the highest classification accuracy. Among ten baseline models, including CNN and RNN, it reaches 89%, which is 7.55% higher than the average of 81.45% and 18% higher than the lowest-performing Attention model. This confirms that the architectural modifications effectively enhance feature extraction. Second, the model shows remarkable training efficiency. Its training time is only 32.22 seconds, which is 14.19 times faster than the average and 32.15 times faster than the slowest BiLSTM-Attention model. This demonstrates the advantages of the optimised model in terms of parameter efficiency and computational speed. Third, data balancing has a notable effect. Training on the pro-WGAN balanced *fma_medium* dataset increases the MAP by 0.252 compared to the original dataset and by 0.057 compared to the GAN-balanced dataset. Classification accuracy improves by 16.8% and 2.6%, respectively. These results highlight the critical role of data preprocessing in improving model performance.

Of particular note is the analysis of indicators such as Macro-F1 and AUC-ROC, which are more sensitive to class imbalance. It is found that the improved model also shows significant improvements in the classification performance of a few classes, with Macro-F1 increasing from 0.63 to 0.80 and AUC-ROC increasing from 0.89 to 0.94. This fully demonstrates the effectiveness of the proposed method in alleviating class imbalance problems. In order to further explore the independent contributions of each improvement module, this study conducts a systematic ablation experiment analysis. The experimental results show that the three major improvement measures all have a significant impact on the final performance improvement: BiLSTM embedding alone contributes 5.5% accuracy improvement, mainly reflected in better modelling of music temporal features. The optimisation of attention mechanism has brought a performance gain of 3.8%, effectively improving the discriminability of feature selection. The data balancing strategy has been proven to be a key factor in performance improvement, contributing over 10% of accuracy improvement alone, highlighting the fundamental role of high-quality data preprocessing in music classification tasks. More importantly, when the three improvement measures work together, a ‘ $1 + 1 + 1 > 3$ ’ effect is produced, indicating good complementarity and synergy among the improvement modules.

The main limitation of this study is that experiments are conducted only on the *fma_medium* dataset. Other datasets with different style distributions, varying sample lengths, or niche genres are not included. Future work should expand the dataset to cover additional sub-genres, such as classical, folk, and electronic music. It should also consider real-world data disturbances, including noise interference and audio quality degradation, to enhance the generalisability and practical applicability of the model.

Declarations

The data supporting the findings of this study are available within the article.

This research did not receive any specific funding.

All authors declare that they have no conflicts of interest.

References

- Chaudhury, M., Karami, A. and Ghazanfar, M.A. (2022) 'Large-scale music genre analysis and classification using machine learning with Apache Spark', *Electronics*, Vol. 11, No. 16, pp.2567–2573.
- Chen, H., Ling, X., Zhang, X. et al. (2023) 'Speaker recognition based on improved ECAPA-TDNN network', *Proceedings of SPIE*, No. 260, pp.1–15.
- Copiaco, A., Ritz, C., Abdulaziz, N. et al. (2021) 'A study of features and deep neural network architectures and hyper-parameters for domestic audio classification', *Applied Sciences*, Vol. 11, No. 11, p.4880.
- Desplanques, B., Thienpondt, J. and Demuynck, K. (2020) 'ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification', Vol. 5, No. 14, pp.3–5.
- Gong, Y., Chung, Y.A. and Glass, J. (2021) 'AST: audio spectrogram transformer', *arXiv preprint*, arXiv:2104.01778.
- Kang, H., He, B., Song, R. et al. (2024) 'ECAPA-TDNN based online discussion activity-level evaluation', *Scientific Reports*, Vol. 14, No. 1, p.14744.
- Li, P., Hoi, M.L., Wang, Y. et al. (2025) 'Enhancing speaker recognition with CRET model: a fusion of CONV2D, RESNET and ECAPA-TDNN', *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2025, No. 1, p.9.
- Liu, C., Feng, L., Liu, G. et al. (2021) 'Bottom-up broadcast neural network for music genre classification', *Multimedia Tools and Applications*, Vol. 80, pp.7313–7331.
- Liu, S., Song, Z. and He, L. (2025) 'Improving ECAPA-TDNN performance with coordinate attention', *Journal of Shanghai Jiaotong University (Science)*, <https://doi.org/10.1007/s12204-024-2726-z>.
- Maccagno, A., Mastropietro, A., Mazziotta, U. et al. (2021) 'A CNN approach for audio classification in construction sites', *Progresses in Artificial Intelligence and Neural Systems*, Vol. 28, No. 8, pp.371–381.
- Medetov, B., Nurlankyzy, A., Namazbayev, T. et al. (2025) 'Speaker recognition by ultrashort utterances', *Eastern-European Journal of Enterprise Technologies*, Vol. 2, No. 9, pp.62–69.
- Nanni, L., Maguolo, G., Brahnam, S. et al. (2021) 'An ensemble of convolutional neural networks for audio classification', *Applied Sciences*, Vol. 11, No. 13, pp.5796–5799.
- Shankar, R.D., Manjula, R.B. and Biradar, R.C. (2025) 'Revolutionizing speaker recognition and diarization: a novel methodology in speech analysis', *SN Computer Science*, Vol. 6, No. 1, pp.1–12.
- Spandan, D. and Goutam, S. (2023) 'Addressing the semi-open set dialect recognition problem under resource-efficient considerations', *Speech Communication*, Vol. 152, p.102957.
- Trejos, A.I.J., Diez, L.A. and Ramos, D. (2025) 'Feature integration strategies for neural speaker diarization in conversational telephone speech', *Applied Sciences*, Vol. 15, No. 9, p.4842.
- Wang, C., Xu, L., Zhu, H. et al. (2025) 'Robustness study of speaker recognition based on ECAPA-TDNN-CIFG', *Journal of Computational Methods in Sciences and Engineering*, Vol. 24, Nos. 4–5, pp.3287–3296.
- Wang, J., Wang, Z., Han, X. et al. (2023a) 'An investigation of ECAPA-TDNN audio type recognition method based on Mel acoustic spectrograms', *Electronics*, Vol. 12, No. 21, pp.23–25.

- Wang, S., Zhang, H., Zhang, X. et al. (2023b) ‘Voiceprint recognition under cross-scenario conditions using perceptual wavelet packet entropy-guided efficient-channel-attention-Res2Net-time-delay-neural-network model’, *Mathematics*, Vol. 11, No. 19, p.20.
- Yang, C., Sun, L., Zhang, J. et al. (2025) ‘Research on the quantification and automatic classification method of Chinese cabbage plant type based on point cloud data and PointNet++’, *Frontiers in Plant Science*, 17 January, Vol. 15, p.1458962, DOI: 10.3389/fpls.2024.1458962.
- Yu, Y., Luo, S., Liu, S. et al. (2020) ‘Deep attention based music genre classification’, *Neurocomputing*, Vol. 372, No. 19, pp.84–91.
- Zaman, K., Sah, M., Direkoglu, C. et al. (2023) ‘A survey of audio classification using deep learning’, *IEEE Access*, Vol. 53, No. 9, pp.521–526.
- Zhang, J. (2021) ‘Music feature extraction and classification algorithm based on deep learning’, *Scientific Programming*, Vol. 20, No. 3, pp.1–9.
- Zhao, B. and Liang, J. (2024) ‘Hierarchical-concatenate fusion TDNN for sound event classification’, *PLoS ONE*, Vol. 19, No. 10, p.e0312998, <https://doi.org/10.1371/journal.pone.0312998>.
- Zinemanas, P., Rocamora, M., Miron, M. et al. (2021) ‘An interpretable deep learning model for automatic sound classification’, *Electronics*, Vol. 10, No. 7, p.850.