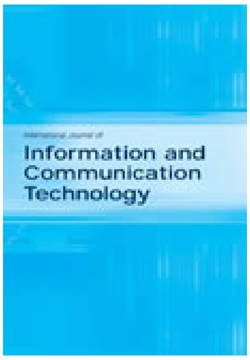
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Classification of classical music genres based on Mel-spectrogram and multi-channel learning

Lei Zhang

Article History:

Received:	30 December 2024
Last revised:	15 January 2025
Accepted:	15 January 2025
Published online:	21 March 2025

Classification of classical music genres based on Mel-spectrogram and multi-channel learning

Lei Zhang

Henan Academy of Drama Arts,
Henan University,
Zhengzhou 451464, China
Email: bassbaritone515@163.com

Abstract: Music genre classification has become a major focus of study as audio processing. Mel-spectrogram and multi-channel learning, MC-MelNet, is proposed in this work for the categorisation job of classical music genres. Combining the Mel-spectrogram and other audio characteristics, with a multi-channel learning framework, the model performs thorough modelling of audio signals. Complete use of the multidimensional information in the audio data enhances the categorisation accuracy. By means of end-to-end training, MC-MelNet simplifies the conventional feature engineering processes and simultaneously performs well in the tests, so attaining higher accuracy, precision, recall, and F1 socre than in the conventional approaches, which show the robustness and efficiency of multi-channel learning in the classification of classical music. The experimental results reveal that the MC-MelNet model can give significant support for the domains of audio classification and music information retrieval in the categorisation of classical music genres.

Keywords: classical music genre classification; Mel-spectrogram; audio feature fusion; multi-channel learning; MCL.

Reference to this paper should be made as follows: Zhang, L. (2025) 'Classification of classical music genres based on Mel-spectrogram and multi-channel learning', *Int. J. Information and Communication Technology*, Vol. 26, No. 5, pp.39–53.

Biographical notes: Lei Zhang received his Master's degree at Xi'an Conservatory of Music in 2008. He is currently an Associate Professor at Henan Academy of Drama Arts in Henan University. His research interests include machine learning, vocal music, and music education.

1 Introduction

Aiming at detecting and separating various kinds of music by automated computer techniques, music genre classification is a crucial activity in the field of music information retrieval (Casey et al., 2008; Sturm, 2014). More and more deep learning-based music genre categorisation techniques have been suggested and shown amazing results in view of the exploding expansion of audio data and the ongoing enhancement of computing capacity. Conventional audio classification techniques depend on hand-designed feature extraction, such Mel-spectrogram (Küçükbay et al., 2022), Mel-frequency cepstral coefficients (MFCC), Chroma features, etc., which are able to better express the frequency domain information of audio and somewhat improve the classification ability of the model. More complicated and several features are therefore required to improve the classification effect as the complexity and variety of music genres grow since a single audio feature sometimes cannot adequately describe the audio signal.

The Mel-spectrogram models the auditory perception mechanism of the human ear by separating the frequency spectrum into several Mel scales, so better capturing the low-frequency characteristics of audio signals and so stressing the frequency range to which the human ear is sensitive. Nevertheless, the Mel-spectrogram lacks the modelling of the time domain information and only offers the frequency domain information of the signal (Zhang et al., 2021; Tang et al., 2023), so it may have some restrictions when handling tasks with strong temporal aspects.

Recent studies have started to investigate the multi-channel learning (MCL) method in order to get above these restrictions (Fang et al., 2024; Al Islam et al., 2015). This method models the audio signal from several dimensions and enhances the classification performance generally by concurrently using several audio feature channels. To improve the diversity and robustness of the model, researchers sometimes mix the Meier spectrogram with other characteristics (e.g., MFCC, Chroma features, etc.), and apply the MCL framework. Particularly in the highly complicated choreography of music genre classification, it might help one become more adept in differentiating several genres.

End-to-end audio categorisation models are progressively going popular as deep learning technology – especially the application of convolutional neural network (CNN) and recurrent neural network (RNN) – develops rapidly (Banerjee et al., 2019). Through the concept of local perception and weight sharing, CNN can efficiently extract high-level features from time-frequency data including Meier spectrograms and capture spatial patterns in audio recordings. RNN, particularly long short-term memory (LSTM) network, are therefore ideally suited for processing audio tasks with temporal dependencies since they can replicate the temporal features in audio signals (Mirza et al., 2024).

Deep learning techniques still have significant difficulties even if they have shown amazing success in the categorisation of musical genres. First, complex audio signals and genre similarities call for more robust feature representation; second, the noise and diversity of audio data could compromise the model's robustness; still, how best to lower the computational complexity while yet guaranteeing the accuracy of the model remains a question of importance.

Thus, this work presents a classical music genre classification model based on Mel-spectrogram and MCL, MC-MelNet.

This work offers the following original innovations:

1 Combining Mel-spectrogram with MCL. We creatively suggest in this work a model structure combining several audio feature channels with Mel-spectrogram. We model audio signals from several dimensions by combining several kinds of features using a MCL framework, so improving the feature representation capacity of the model and hence the accuracy of the categorisation of classical music genres.

2 End-to-end deep learning model. This work presents an MC-MelNet model that uses MCL along with Mel-spectrogram to extract audio features and fusion, therefore producing a more complete and accurate audio classification model. The model

improves the classification performance by using end-to-end training based on deep learning, hence simplifying the conventional feature engineering processes.

3 Improved feature fusion strategy. In order to avoid the restriction of a single feature, a new feature fusion strategy is suggested in this work using a weighted summation method to merge the information of several feature channels into a new feature vector. After feature fusion, the method significantly increases the representation capacity and produces more outstanding results in the classical music genre classification.

2 Relevant technologies

2.1 Mel-spectrogram

Widely applied in speech recognition, music classification, and other domains, Mel-spectrogram is an audio feature representation approach based on Mel scale that can efficiently extract the essential characteristics in the audio signal by converting the audio signal into time-frequency representation (Ustubioglu et al., 2023). First in the process of audio processing, the audio signal must be transformed into a frequency domain representation using short-time Fourier transform (STFT) (Zhu et al., 2007). The STFT of a particular audio signal $x(t)$ may be stated using the following equation:

$$X(t, f) = \int_{-\infty}^{\infty} x(\tau) w(t-\tau) e^{-j2\pi f \tau} d\tau \tag{1}$$

where the window function is $w(t-\tau)$, t represents the time step; f is the frequency. With the magnitude component reflecting the intensity of the signal at a given frequency and the real and imaginary parts representing the phase information, respectively, the STFT produces a complex number that reflects the information of the signal in the time-frequency plane.

The power spectrum $P(t, f)$ is then obtained by first considering the square of the magnitude of $X(t, f)$, therefore characterising the energy distribution of the signal in the time and frequency domains:

$$P(t, f) = |X(t, f)|^2 \tag{2}$$

Frequency analysis is built on the power spectrum $P(t, f)$, which offers details on the energy distribution of the signal at every frequency. The frequency axis must be mapped to the Meier scale if one is to more closely fit the perceptual characteristics of the human ear.

Based on the nonlinear perception of frequencies by the human ear, the Meier scale is built to produce a lesser resolution in the high frequency section and a higher resolution in the low frequency part (Wolfe et al., 2011). The map equation for the Meier scale follows:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

where f is the common frequency; f_{mel} is the related Mel frequency.

Through compressing the details of the high-frequency signal to a lower resolution, so enhancing the effectiveness of the representation of audio features, the Mel-spectrogram is able to better replicate the perception of frequencies by the human ear. Convert the frequency axis to the Mel scale.

The power spectrum must then be handled next using a Mel filter bank after the frequency mapping (Nadeu et al., 2001). The Meier filter bank is made of a sequence of triangle filters whose frequency response spans the Meier scale consistently (Kathania et al., 2019). One may define the frequency response $H_m(f)$ of the Mel filter as follows segmented function:

Regarding $f < f_{m-1}$ 和 $f \geq f_{m+1}$:

$$H_m(f) = 0 \tag{4}$$

That is to say, the filter responds with zero both before and after its upper border.

Regarding $f_{m-1} \leq f < f_m$:

$$H_m(f) = \frac{f - f_{m-1}}{f_m - f_{m-1}} \tag{5}$$

The filter responds linearly in this interval from f_{m-1} to frequencies between fm.

Regarding $f_m \leq f < f_{m+1}$:

$$H_m(f) = \frac{f_{m+1} - f}{f_{m+1} - f_m} \tag{6}$$

The filter responses in this period are linearly declining in frequency from f_m to f_{m+1}.

The Mel spectrum $M(t, m)$ is obtained by weighted and summed with the power spectrum as the signal travels through these filters:

$$M(t, m) = \sum_f H_m(f) P(t, f) \tag{7}$$

Transforming the power spectrum using a Mel filter bank yields the Mel spectrum $M(t, m)$, which on the Mel frequency scale denotes the energy distribution of an audio stream. Often a logarithmic adjustment of the Mel spectrum is done to further lower the dynamic range and better fit the perceptual qualities of human hearing. The logarithmic transformation reduces the huge dynamic range and accentuates the low-energy aspects. One computes the logarithmic Mel spectrum $L(t, m)$ as follows:

$$L(t, m) = \log\left(M(t, m) + \epsilon\right) \tag{8}$$

where in logarithmic operations a modest constant ϵ helps to avoid zeros. The Mel-spectrogram offers a time-frequency characteristic representation of the audio signal on the Mel scale, therefore representing the last logarithmic Mel spectrum $L(t, m)$. Particularly the ability to identify the spectral variations between different genres, the Mel-spectrogram can efficiently capture the time-frequency characteristics of audio in audio classification tasks including classical music genre classification, so providing strong feature support for machine learning models.

2.2 Multi-channel learning

MCL is a method of knowledge extraction and fusion utilising several input channels (Guo et al., 2022; Ye et al., 2021). Especially in audio classification jobs, it has been extensively applied in audio signal processing chores. Learning features from several sources helps the model to get acoustic information from several points of view, hence strengthening classification accuracy and resilience. In the job of classifying classical music, the audio signal can be feature extracted from several angles including MFCC, time domain waveform, and the Meier spectrogram. Every feature captures a separate component of the audio signal; so, integrating features from several channels helps the model to have better identification ability.

Suppose in MCL we extract multiple types of features for the audio signal $x(t)$, each matching an input channel. These properties are $f_1(t)$, $f_2(t)$, …, $f_n(t)$ accordingly, where every $f_i(t)$ shows the signal's representation on a certain channel. By now the audio signal's multichannel characteristics could be seen as a set of vectors:

$$F(t) = [f_1(t), f_2(t), ..., f_n(t)] \tag{9}$$

Each feature has a matching representation at some time point t; these feature vectors can contain various kinds of features including Meier spectrograms, time-domain features, zero-crossing rates (ZCRs), etc.

Deep learning models typically extract more abstract feature maps by means of CNN or another network design, therefore processing the features of every channel. Under that channel, these feature maps can mirror the local qualities of the audio signal. A time-domain waveform, for instance, tells about the instantaneous changes in the audio signal while a Meier spectrogram shows the time-frequency distribution of the audio signal (Wei et al., 2018). Following the feature extraction, an independent convolution procedure generates fresh set of feature representations for every channel.

Combining the features of each channel by concatenation or weighted summation is typical practice to efficiently integrate the information from many channels. Let a CNN handle the features $f_1(t)$, $f_2(t)$, …, $f_n(t)$ of every channel to generate a high-dimensional feature map. The concatenation approach combines sequentially the feature maps of every channel to generate a high-dimensional feature vector $z(t)$:

$$z(t) = [f_1(t), f_2(t), ..., f_n(t)] \tag{10}$$

This splicing helps to maintain the feature information of several channels by joining their elements into one huge vector. Feature fusion can be used for some jobs where the contributions between channels might not be the same (Dai et al., 2024). This uses weighted fusion. In weighted fusion, the weighted fused feature vector can be stated as the features of each channel given varying weights w_i depending on their relevance:

$$z(t) = \sum_{i=1}^{n} w_i \cdot f_i(t) \tag{11}$$

where w_i is the weighting coefficient of channel i, often learnt during the training procedure. This weighted fusion dynamically changes the contribution of every channel in the final feature vector, therefore enabling the model to pay more attention to the channels that support the classification objective.

The classifier will get the fused feature vector $z(t)$ for ultimate classification. Practically, a fully connected layer usually handles these combined characteristics. With MCL, the output $\hat{y}(t)$ of the classifier can be stated as:

$$\hat{y}(t) = \sigma\left(W \cdot z(t) + b\right) \tag{12}$$

Usually using sigmoid activation function (for binary classification problems) or softmax activation function (for multiclassification problems), where W is the weight matrix, b is the bias term, and σ is the activation function produces the prediction results.

By use of MCL, the model can build a multi-level and rich feature representation of the audio signal by aggregating features from several channels, so augmenting the basic information of the signal from a single channel. Especially in the classification assignment for classical music, this method can efficiently improve the audio classification performance by means of integrated use of multi-channel information and greatly increase the classification accuracy and resilience.

3 Classical music genre classification model based on Mel-spectrogram and MCL

3.1 Model framework: MC-MelNet

This work proposes a classical music genre classification model based on Mel-spectrogram and MCL, named MC-MelNet, which combines the frequency-domain characteristics of Mel-spectrogram and multi-dimensional feature inputs of MCL for effective classification by deep neural network. Not only can MC-MelNet extract audio frequency features from Mel-spectrogram but also fusing time-domain features such ZCR and short-time energy (STE) to acquire a more complete audio signal representation using the multi-channel input and feature fusion approach. To provide a more complete picture of the audio signal, extracts audio frequency characteristics as well as merges time-domain aspects including ZCR and STE.

First, STFT transforms the audio signal $x(t)$ into a Mel-spectrogram $M(t)$, which is then created by Mel filter bank processing and frequency domain feature of the model. Calculated as the distribution of the audio signal in frequency and time, the Mel-spectrogram may fairly depict these aspects.

$$M(t) = F_{Mel}\left(F_{STFT}\left(x(t)\right)\right) \tag{13}$$

MC-MelNet additionally generates other features from the time domain of the audio signal, such as $ZCR(t)$ and $STE(t)$, so enhancing the feature representation of the model. These equations respectively allow one to determine the ZCR and STE:

$$ZCR(t) = \sum_{n=1}^{N} \Pi\left(x(n) \cdot x(n-1) < 0\right) \tag{14}$$

$$STE(t) = \sum_{n=1}^{N} \left|x(n)\right|^2 \tag{15}$$

These time-domain characteristics enable dynamic audio signal recording and reveal information on the timing of the audio stream. MC-MelNet synthesises several facets of the audio stream by combining Mel-spectrograms with time-domain characteristics, therefore improving the classification of classical music genres.

MC-MelNet uses CNN to do multi-channel feature fusion following feature extracting. To derive more representative features, the output $f_i(t)$ of every feature channel will be convolved with the learnt convolution kernel by convolution operation. One may depict this procedure as follows:

$$f_i'(t) = C_i\left(f_i(t)\right) \tag{16}$$

where C_i is the convolution process; $f_i'(t)$ is the i^{th} channel's output feature map following convolution. A cross-channel fusion procedure will link the outputs of every channel once each one has been convolved. More especially, all the convolved features will be spliced along the feature dimensions into a new feature vector $z(t)$:

$$z(t) = concat\left(f_1'(t), f_2'(t), \ldots, f_n'(t)\right) \tag{17}$$

where $concat(\cdot)$ represents the sewing of features output from several channels to generate a consistent feature representation $z(t)$. By means of convolutional procedures, this fusion approach not only retains the information of every feature channel but also automatically extracts more complex and discriminative features. MC-MelNet makes full use of the multi-dimensional information of several channels by means of cross-channel feature splicing, hence improving the model's classification capacity.

The deep CNN receives the fused feature vector $z(t)$ to further extract high-level audio signal features and eventually applied for genre categorisation. Through the spliced multidimensional features, the CNN learns the local patterns of the audio signal, therefore collecting important audio characteristics as notes and rhythms. Following multi-layer convolution and pooling procedures, the final output classification result $\hat{y}(t)$ will be utilised to produce the probability distribution of every genre via the softmax function, so defined by the formula:

$$\hat{y}(t) = \text{softmax}\left(W \cdot C(t) + b\right) \tag{18}$$

where W is the weight matrix of the fully connected layer; $C(t)$ is the feature map produced by the convolution and pooling layers; b is the bias term; and the softmax function generates the genre to which the audio signal falls.

MC-MelNet is able to effectively fuse the multi-channel information of Mel-spectrograms and time-domain features using this modelling framework, and concurrently extract deep-level audio features using CNNs, so enabling accurate classification of classical music genres. This framework improves the expressiveness and accuracy of the classification model by recording local acoustic patterns as well as by combining several feature information.

3.2 Model training and inference process

First, MCL and feature fusion trains and infers the Meier spectrogram and other pertinent characteristics by loading and processing the input audio data.

In this case, x_i is the audio data and y_i is the label – that is, music genre – the dataset $D = \{(x_i, y_i)\}$ comprises audio samples x_i and their matching labels y_i. Features like Mel-spectrogram, ZCR, and STE are obtained for every audio file.

The model's training runs through numerous phases. First, hyperparameters like learning rate, batch size, and number of training rounds are defined together with starting weights for the convolutional and fully-connected layers. Every training round the dataset is split into several tiny batches, each of which is used for forward propagation, loss computation, gradient update and other activities of the model.

Every batch of audio samples x_i' will be extracted with features including Mel-spectrogram, ZCR and STE, and feature extraction will be done via convolution operation throughout the training process. Feature fusion will combine the outputs of several feature channels; thereafter, the fused features will be forward propagated via a neural network. Back propagation generates the loss value between the model prediction result and the actual label; it also modulates the model weights.

Using a test set, the model is assessed following completion of the training. The audio data is handled in the evaluation phase using the same feature extraction and convolution, then fed into the trained model for inference, therefore producing the projected labels for every sample.

Algorithm 1 exhibits the pseudo-code for inference and training:

Algorithm 1 Pseudo-code for inference and training MC-MelNet

```
begin
    for each audio file in dataset $D$ do
        Load audio file $x_i$;
        Preprocess $x_i$ to extract features: Mel-spectrogram, ZCR, STE;
    Store the extracted features as $x_i'$ for later processing;
    end for
    Initialise convolutional layers with random weights;
    Initialise fully connected layers with random weights;
    Set the model hyperparameters (e.g., learning rate, batch size, etc.);
    for epoch = 1 to E do
        Shuffle dataset D to randomise the training order;
        for batch = 1 to B do
            for each audio sample $x_i'$ in the batch do
                Extract Mel-spectrogram, ZCR, STE from $x_i'$;
                Perform convolutional operations for each feature channel;
                Perform feature fusion by concatenating the outputs of each channel;
            end for
            Pass the fused features through the neural network;
            Compute the predicted genre \( \hat{y_i} \) for each sample in the batch;
            Compute the loss using cross-entropy or another suitable loss function;
            Compute the gradient of the loss with respect to each parameter;
            Update the weights of convolutional and fully connected layers using gradient
            descent or other optimiser;
        **end for**
    **end for**
        **for** each test sample $x'$_test in test dataset $D$_test **do**
        Extract Mel-spectrogram, ZCR, STE from $x'$_test;
        Perform convolution and feature fusion as in training;
        Pass the fused features through the trained neural network;
        Output the predicted genre \( \hat{$y_i$} {test} \);
    **end for**
    **Return** the trained model weights $W$_final;
    **Return** the classification results for test dataset $D$_test;
**end**
```

We measured the model's performance using the four evaluation criteria listed below throughout the training and assessment process:

1 Accuracy

One of the most often used classification evaluation measures, accuracy shows the fraction of properly categorised samples among all the samples. Its calculating formula is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{19}$$

2 Precision

The proportion of all the samples expected to fit a certain category that really fall into that category is indicated by the accuracy rate. Its computation follows the formula:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{20}$$

where *TP* counts the actual cases and *FP* counts the false positives.

3 Recall

Recall is the percentage of all the samples that the model can effectively identify into a category. Its computation follows the formula:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{21}$$

where FN stands for the false negative count.

4 F1-score

The F1-score strikes a compromise between recall and accuracy by averaging both. The computation approach is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

These evaluation criteria provide a thorough evaluation of the MC-MelNet model in audio genre classification tasks, therefore facilitating analysis of the model's classification accuracy, stability, and efficiency.

4 Experimental results and analyses

4.1 Datasets

The GTZAN music dataset, a classical dataset for music classification activities due to its standardisation, genre diversity and wide range of applications, was selected for this experiment in order to evaluate the performance of the MC-MelNet model in the classical music genre classification task.

Designed for audio classification tasks, the GTZAN dataset was made public by machine learning field researchers. There are ten different music genres in it, and each has one hundred audio samples, for a thousand audio files overall. Using a 16-bit mono recording technique, the audio samples run 30 seconds, and a 22,050 Hz sampling rate. Uniform audio file structure of the GTZAN dataset helps model training and experimental evaluation.

Table 1 lists the salient characteristics and specifics of the GTZAN dataset.

Table 1 GTZAN music dataset statistical information

Attribute	*Description*
Number of genres	10 genres
Number of samples per genre	100 audio samples per genre
Total number of samples	1,000 audio samples
Audio duration	30 seconds per audio sample
Audio sampling rate	22,050 Hz
Audio format	16-bit mono audio format
Main genres	Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock

Training, validation, and testing sets separate the dataset to enable efficient model training and evaluation of generalisation capability. Eighty percent of the data is utilised for training, 10% for validation, and 10% for testing – the particular division ratio is thus. To guarantee a balanced data split, the audio samples of every genre are randomly distributed to several subgroups.

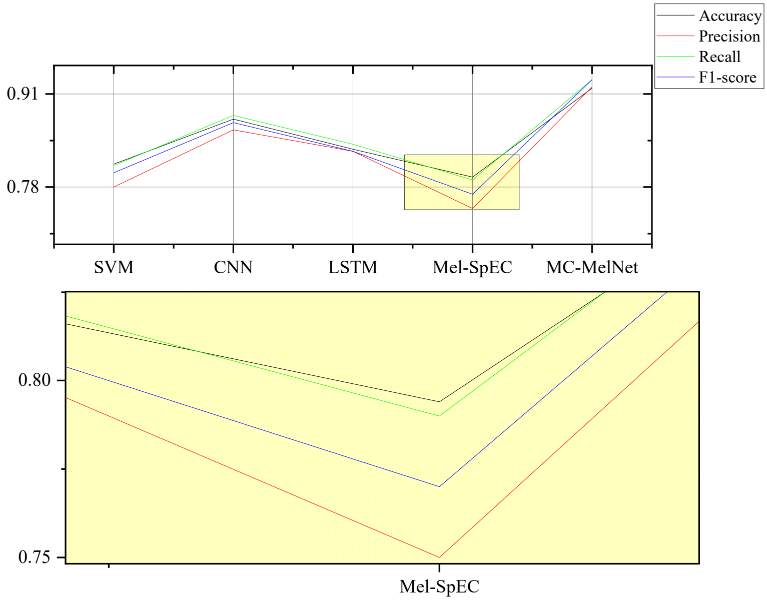
Prior to model training, the audio data in the GTZAN dataset is preprocessed as follows: first, each audio sample is sliced into multiple 2-second-long segments in order to make it easier for the model to capture detailed features in the audio. Each audio sample is split into 15 segments. Next, the time-domain signal is converted into a frequency-domain signal using STFT to compute the Mel spectrogram. All Mel spectrograms are normalised for zero mean and unit variance. In addition, to increase the diversity and robustness of the data, the audio clips were subjected to data enhancement operations such as time shifting, volume scaling and spectral transformation.

4.2 Comparison experiments

We performed comparison studies with numerous classical audio classification models to comprehensively assess the performance of the MC-MelNet model in the genre categorisation in classical music. Support vector machine (SVM), CNN, LSTM, and the conventional Mel-spectrogram (Mel-Spec) feature extracting technique are among the comparison models.

Figure 1 shows the experimental findings, therefore illustrating the performance of many models on the four evaluation criteria. Every experiment was carried out ten independent runs and the findings were averaged to guarantee their dependability.

Figure 1 Results of the comparison experiment (see online version for colours)



From the experimental results, it is evident that the MC-MelNet model exhibits outstanding performance in all the assessment criteria, particularly in the accuracy rate, precision rate, recall rate, F1-score, which has a major advantage over other models. With a precision rate of 0.92, a recall rate of 0.93, and an F1-score of 0.93, MC-MelNet specifically achieves 91.8% in accuracy – higher than the other compared models – and shows the strong ability of the model in the audio classification task.

Though placed second in terms of accuracy (87.5%), the CNN model performs somewhat worse in terms of precision rate, recall rate, and F1-score, with 0.86, 0.88 and 0.87, respectively, compared with MC-MelNet .With a better balance particularly in terms of recall rate (0.84) and F1-score (0.83), the LSTM model also performs better; yet, the general accuracy rate is still lower than MC-MelNet at 83.3%.

In this experiment, the conventional SVM model and the Mel-Spec approach based on Mel-spectrograms performed somewhat poorly. Although its performance is still reasonable in some situations, the SVM, with an accuracy of 81.2%, a precision of 0.78, a recall of 0.81, and an F1-score of 0.80, obviously cannot compete with the deep learning approaches in challenging audio classification tasks. Mel-SPEC Although it may be efficient in some simple tasks, as a conventional feature extraction technique its performance in this trial is somewhat low, with an accuracy of just 79.4%, a precision of 0.75, a recall of 0.79, and an F1-score of 0.77.

By means of its integration of Mel-spectrograms and MCL, the MC-MelNet model dramatically enhances the accuracy and robustness of the model in the audio classification task overall. Based on its great classification performance, MC-MelNet is still the best option in practical applications even if its training time is somewhat longer than that of other models.

4.3 Ablation experiments

By progressively deleting various modules from the MC-MelNet model, we hope to evaluate the contribution of specific components to the general performance of the model in the ablation studies. To investigate the effect of these modules on the model performance, we specifically created three sets of ablation experiments each deleting one important module of the model. The ablation studies aim to identify which modules are most important in the audio classification task and which module removal causes appreciable performance reduction.

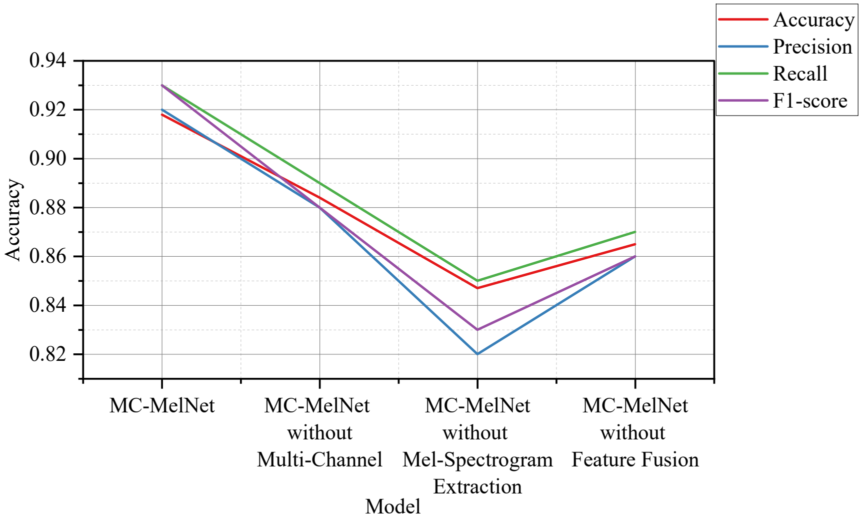
First series of tests removes the MCL module (MC-MelNet without MCL). The model still takes the Mel-spectrogram as input in this experiment, but it no longer uses MCL; all feature channels will be straightly merged into a single feature representation. The accuracy falls from 91.8% to 88.4% and the precision, recall, and F1-score also somewhat drop following the removal of the MCL module, correspondingly. This suggests that feature fusion and model performance enhancement depend much on the MCL module. Different kinds of features can be handled by MCL, thereby allowing the model to learn from several dimensions and so get better classification accuracy. The model loses full use of the various feature information after eliminating this module, which results in performance deterioration.

Removal of the Mel-spectrogram feature extraction module (MC-MelNet minus Mel-spectrogram extraction) forms the second set of studies. In this experiment the model directly employs raw audio data or other basic audio properties instead of Mel-spectrogram as input. With a precision of 0.82, a recall of 0.85, and an F1-score of 0.83, the results demonstrate that the accuracy of the model drastically drops to 84.7%, following the Mel-spectrogram feature extraction. This implies that, especially for the audio classification task, the Mel-spectrogram is an efficient audio feature able to capture the spectral properties in the audio signals. The model loses this crucial feature information after deleting the Mel-spectrogram, which causes a notable reduction of the classification performance.

The module with feature fusion deleted (MC-MelNet without feature fusion) forms the third set of tests. In this experiment, we eliminated the fusion mechanism among several feature channels in the model and instead categorised each feature channel independently and just combined their outputs. The accuracy of the model decreases to 86.5% with a precision of 0.86, a recall of 0.87, and an F1-score of 0.86.5% once the feature fusion module is removed. Though the performance declines, the change is minor when compared to the tests without the MCL module and the Meier spectrogram feature extraction module. This implies that, particularly in cases of effective integration between several features, which can enhance the resilience and classification capacity of the model, the feature fusion module aids to help to further improve the performance of the model. Still, feature fusion's importance is more subdued than that of Mel's spectrogram features and MCL.

Figure 2 displays the experimental findings:

Figure 2 Results of the ablation experiment (see online version for colours)



By means of these ablation studies, we can deduce that the core elements in the performance of the MC-MelNet model are the MCL module and the Mel-spectrogram feature extraction module; whereas, the feature fusion module has a rather minor influence even if it helps to increase the classification effect. The significant function of Mel-spectrogram and MCL in the classification job of classical music is validated by the experimental results.

5 Conclusions

In this work, we propose a classical music genre classification model (MC-MelNet) based on Mel-spectrogram and MCL, which essentially improves the accuracy and robustness of classical music genre classification by combining Mel-spectrogram features and a MCL framework. We carried comparison and ablation studies and matched the model with conventional audio classification techniques on numerous assessment criteria to confirm its efficacy. Verifying the excellence of the method in the classical music genre classification problem, the experimental findings reveal that the MC-MelNet model beats other benchmark approaches in terms of accuracy, precision, recall, and F1-score.

The MC-MelNet approach has certain restrictions even if it has improved outcomes in the classification of classical music genres. First of all, the model is now limited to the classification challenge of classical music genres; other kinds of music genres or audio data may need different feature extraction and model structure to fit. Secondly, the training process of the model is more complex, especially in MCL and feature fusion, which requires a lot of computational resources and time. As the size of the dataset increases, the training time and computational overhead also increase significantly, which is a challenge for resource-constrained environments. Furthermore, in some circumstances – especially in more complicated audio environments – the Meier spectrograms employed as audio cues in this work might not be able to fully represent all the important information in the audio stream. While MCL can extract features from several dimensions, the performance of the model may degrade in cases of poor quality of the input audio or additional noise. At last, the present model mostly depends on audio genre classification and has not yet been completely validated even if it performs well on various evaluation criteria.

Future studies can be enhanced and broadened in the following respects:

1 Cross-genre generalisation ability is improved. Although the MC-MelNet model has shown outstanding performance in the classification of classical music genres, its generalisation capacity has not been entirely confirmed. Larger multi-genre datasets will help the model's cross-genre learning capacity to be improved going forward. Combining a greater spectrum of music genres, including modern music and electronic music, for instance, will let one investigate how to raise the model's capacity for discriminating between several genres.

2 A more efficient feature extraction method. Though in some situations may not be able to completely capture the detailed information in the audio, Mel-spectrograms have been extensively validated as a portrayal of audio properties. Future exploration of more sophisticated audio feature extraction techniques is warranted. Furthermore, investigating automatic feature learning techniques grounded in deep learning could help to lessen the dependence on handcrafted features, hence enhancing the accuracy and efficiency of the model.

3 Real-time performance optimisation of the model. Future research can concentrate on maximising the inference efficiency of the MC-MelNet model so that it can operate real-time audio classification in resource-limited contexts since the model takes high computational resources during training. Techniques like model pruning and quantisation, for instance, can help to lower the computational complexity of the model thereby enabling its adaptation to embedded devices or real-time audio stream processing systems.

Declarations

All authors declare that they have no conflicts of interest.

References

Al Islam, A.A., Islam, M.J., Nurain, N. et al. (2015) 'Channel assignment techniques for multi-radio wireless mesh networks: a survey', *IEEE Communications Surveys & Tutorials*, Vol. 18, No. 2, pp.988–1017.

Banerjee, I., Ling, Y., Chen, M.C. et al. (2019) 'Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification', *Artificial Intelligence in Medicine*, Vol. 97, pp.79–88.

Casey, M.A., Veltkamp, R., Goto, M. et al. (2008) 'Content-based music information retrieval: current directions and future challenges', *Proceedings of the IEEE*, Vol. 96, No. 4, pp.668–696.

Dai, H., Wang, J., Zhong, Q. et al. (2024) 'A GAN-based anomaly detector using multi-feature fusion and selection', *Scientific Reports*, Vol. 14, No. 1, p.5259.

Fang, H., Sha, L. and Liang, J. (2024) 'Multimodal recommender system based on multi-channel counterfactual learning networks', *Multimedia Systems*, Vol. 30, No. 5, p.242.

Guo, Y., Hu, T., Zhou, Y. et al. (2022) 'Multi-channel data fusion and intelligent fault diagnosis based on deep learning', *Measurement Science and Technology*, Vol. 34, No. 1, p.015115.

Kathania, H.K., Shahnawazuddin, S., Ahmad, W. et al. (2019) 'Role of linear, mel and inverse-mel filterbanks in automatic recognition of speech from high-pitched speakers', *Circuits, Systems, and Signal Processing*, Vol. 38, pp.4667–4682.

Küçükbay, S.E., Yazıcı, A. and Kalkan, S. (2022) 'Hand-crafted versus learned representations for audio event detection', *Multimedia Tools and Applications*, Vol. 81, No. 21, pp.30911–30930.

Mirza, F.K., Gürsoy, A.F., Baykaş, T. et al. (2024) 'Residual LSTM neural network for time dependent consecutive pitch string recognition from spectrograms: a study on Turkish classical music makams', *Multimedia Tools and Applications*, Vol. 83, No. 14, pp.41243–41271.

Nadeu, C., Macho, D. and Hernando, J. (2001) 'Time and frequency filtering of filter-bank energies for robust HMM speech recognition', *Speech Communication*, Vol. 34, Nos. 1–2, pp.93–114.

Sturm, B.L. (2014) 'The state of the art ten years after a state of the art: future research in music information retrieval', *Journal of New Music Research*, Vol. 43, No. 2, pp.147–172.

Tang, N., Zhou, F., Wang, Y. et al. (2023) 'Differential treatment for time and frequency dimensions in mel-spectrograms: an efficient 3D spectrogram network for underwater acoustic target classification', *Ocean Engineering*, Vol. 287, p.115863.

Ustubioglu, B., Tahaoglu, G. and Ulutas, G. (2023) 'Detection of audio copy-move-forgery with novel feature matching on Mel spectrogram', *Expert Systems with Applications*, Vol. 213, p.118963.

Wei, X., Zhou, L., Chen, Z. et al. (2018) 'Automatic seizure detection using three-dimensional CNN based on multi-channel EEG', *BMC Medical Informatics and Decision Making*, Vol. 18, pp.71–80.

Wolfe, J., John, A., Schafer, E. et al. (2011) 'Long-term effects of non-linear frequency compression for children with moderate hearing loss', *International Journal of Audiology*, Vol. 50, No. 6, pp.396–404.

Ye, F., Guo, Y., Xia, Z. et al. (2021) 'Feature extraction and process monitoring of multi-channel data in a forging process via sensor fusion', *International Journal of Computer Integrated Manufacturing*, Vol. 34, No. 1, pp.95–109.

Zhang, T., Feng, G., Liang, J. et al. (2021) 'Acoustic scene classification based on Mel spectrogram decomposition and model merging', *Applied Acoustics*, Vol. 182, p.108258.

Zhu, X., Beauregard, G.T. and Wyse, L.L. (2007) 'Real-time signal estimation from modified short-time Fourier transform magnitude spectra', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 5, pp.1645–1653.