

# Performance Optimization of Digital Music Multi-Instrument Recognition Based on an Improved DCNN Model and Bi-LSTM

Junhui Zhao<sup>1</sup>, Xiaohang Jia<sup>2</sup>, and Yuxuan Zhou<sup>3</sup>

<sup>1</sup> Associate Professor, College of Art, Hebei University of Economics and Business, Shijiazhuang, 050062, China,  
E-mail: junhuizhao1@outlook.com (corresponding author).

<sup>2</sup> Postgraduate Student, Music Performance, Hebei University of Economics and Business, Shijiazhuang, 050062, China,  
E-mail: xiaohangjia@heu.uu.me

<sup>3</sup> Postgraduate Student, Vocal Performance, Hebei University of Economics and Business, Shijiazhuang, 050062, China,  
E-mail: yuxuanzh31@heu.uu.me

Project Management

Received January 21, 2026; revised March 29, 2026; accepted September 6  
Available online September 20, 2026

**Abstract:** To solve the problems of low accuracy and insufficient robustness in instrument recognition in multi-instrument scenarios of digital music, a parallel structure combining a standard Convolution Neural Network and a multi-scale Dilated Convolution Neural Network (CNN-DCNN) is designed to extract multi-scale spatial acoustic features from Mel spectrograms. The feature sequence extracted by CNN-DCNN is fed into a Bidirectional Long Short-Term Memory (Bi-LSTM) network for bidirectional context modeling, and a self-attention mechanism is introduced to dynamically weight temporal features, capturing the inherent long-range temporal dependencies of music signals. The findings demonstrate that the proposed fusion-improved model achieves optimal performance across multiple key metrics, with mean accuracy, macro-average F1 score, and mean precision reaching 88.92%, 87.11% and 88.34%, respectively, and significantly outperforming other comparative models. Ablation experiments further confirm the effectiveness of the parallel convolutional structure, temporal modeling, and attention mechanism. The proposed model, by deeply fusing multi-scale spatial features with bidirectional temporal contextual information, provides a high-performance solution for multi-instrument recognition in complex acoustic scenarios and offers valuable references for intelligent music analysis and retrieval.

**Keywords:** Multi-instrument recognition; deep learning; dilated convolutional neural network; bidirectional long short-term memory network; attention mechanism; mel spectrogram.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).  
DOI 10.32738/JEPPM-2026-91

## 1. Introduction

probability latent component analysis. These methods avoid complex feature engineering and demonstrate stronger feature learning capabilities. However, music signals themselves have strong temporal correlations, and some models in existing research (such as standard CNNs) have insufficient capture of the long-term dependencies of musical instrument note signals, failing to fully utilize the unique properties of multi-instrument audio spectrograms, which, to some extent, ignore the long-range temporal dependencies of musical instrument note signals. Simply relying on neural network architectures cannot fully capture the characteristics of time-frequency spectrograms in multi-instrument audio in digital music. Therefore, this research proposes fusing the dilated convolutions of a Dilated Convolutional Neural Network (DCNN) with the ordinary convolutions of a CNN, thereby forming a parallel structure that ensures discriminative learning of the original audio or its time-frequency representations (such as Mel spectrograms) and guarantees comprehensive feature selection. Secondly, a Bidirectional Long Short-term Memory (Bi-LSTM) network is introduced and fused to recognize both local spatial and global temporal features, thereby improving multi-instrument performance. The innovations of this research are reflected in two aspects: firstly, the parallel design of ordinary convolutions and dilated convolutions effectively captures feature information. Secondly, a series of high-level feature sequences extracted by CNN-DCNN is used as input, and a

Bi-LSTM is used to model temporal context and capture richer musical context information. Based on the theory of representation learning, it is hypothesized that by synergistically fusing multi-scale spatial feature extraction, bidirectional temporal context modeling, and attention-based feature weighting, a more robust and accurate representation of instrument timbre in complex polyphonic music can be constructed than a model that focuses solely on spatial or temporal dimensions. This theory suggests that a good feature representation should capture the invariance and variability of data across multiple levels and dimensions of abstraction. Meanwhile, it is also in line with sequence modeling theory, which requires models to capture long-range dependencies and contextual information to effectively process time-series data. This research aims to provide a higher-performance, more robust solution for multi-instrument recognition in digital music, thereby promoting its application in practical scenarios such as intelligent music retrieval, interactive music applications, and automated music production.

## 2. Design of a Multi-Instrument Recognition Algorithm for Digital Music

### 2.1. Music Feature Extraction based on Improved DCNN Model

The original digital music audio is a high-dimensional time-domain waveform signal. Direct processing is not only computationally intensive but also makes it difficult to reveal its inherent timbre attributes. Among many audio features, Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in audio analysis tasks due to their sensitivity to both high- and low-frequency regions and their nonlinear representation of sound frequencies. Therefore, this study extracts the Mel spectrum by first dividing the continuous audio signal into short time frames, then applying a Hamming window to each frame to ensure continuity between frames and reduce spectral leakage. Then, each windowed time-domain frame is processed with a fast Fourier transform to obtain its linear spectrum, which is then passed through a set of triangular Mel filters. Eq. (1) is the conversion between Mel frequency  $m$  and linear frequency  $f$ .

$$m = 2595 \cdot \log_{10}(1 + 700/f) \quad (1)$$

Calculating the energy output of each filter and taking its logarithm makes the result more closely match the human ear's perception of loudness. Arranging the logarithmic Mel-Energy spectra of all frames in chronological order generates a two-dimensional Mel-Energy spectrogram. The size of the analysis window directly affects the amount of information contained in the extracted features. Considering the duration of notes and the stable presentation of timbre features in instrumental performance, this study sets the analysis window size to 3 seconds to preserve high-quality timbre information.

CNNs are widely used in musical instrument recognition. They use filters composed of multiple neurons to form convolutional arrays, use receptive fields to analyze features in each region, and use activation functions to produce the output of the data processing. Eq. (2) is the expression of the convolution operation of CNN.

$$s(t) = x(t) * c(t) = \sum_{\kappa=-\infty}^{\infty} x(\kappa) c(t - \kappa) \quad (2)$$

In Eq. (2),  $x(t)$  is the input feature,  $c(t)$  is the Convolutional nuclei,  $s(t)$  is the feature map, and  $\kappa$  is the activation function. However, when CNN processes Mel spectrogram features, relying heavily on pooling operations to expand the receptive field may affect local fine textures and lead to loss of harmonic structural information (Richard et al., 2023). In view of this, the study proposes using a DCNN to perform signal processing. A DCNN can expand the receptive field without reducing image resolution, and the mathematical expression of its convolution signal is given by Eq. (3).

$$y[i] = \sum_{w=1}^W (x[i + r * w] * f[w]) \quad (3)$$

In Eq. (3),  $x[]$  is the input signal,  $y[]$  is the output signal after filtering,  $w$  is the filter channel width,  $W$  is the maximum channel width, and  $r$  is the dilation rate. When the dilation rate hyperparameter is 1, the dilated convolution degenerates into a standard convolution. When it is greater than 1, the convolution kernel will skip pixels when sampling the input feature map. By applying different dilations in different layers, dilated convolution can capture multi-scale contextual information without downsampling or changing the feature map resolution (Ji et al., 2023; Li and Wang, 2024). Therefore, this study proposes combining the advantages of standard and dilated convolutions to construct a parallel dual-branch network structure. In CNN-DCNN, the input Mel spectrogram is split into two parallel branches: a standard convolution branch and a dilated convolution branch. Residual connections are used in both branches to alleviate the gradient vanishing problem in deep networks. The feature maps extracted by the two branches are concatenated and fused in the channel dimension to obtain a composite feature representation that integrates multi-scale features and multi-level information. Eq. (4) is the mathematical expression of the fusion process.

$$F_{fused} = \text{Concat}(F_{conv}, F_{dconv}) \quad (4)$$

In Eq. (4),  $F_{conv}$  and  $F_{dconv}$  These are the output feature maps from the standard and dilated convolution branches, respectively. This parallel fusion design retains the standard convolution's ability to finely capture local details while leveraging dilated convolutions to capture global context, thereby achieving a more comprehensive and robust representation of musical instrument timbre features. To further clarify the structural details of the proposed CNN-DCNN parallel module, Table 1 provides a detailed hierarchical configuration.

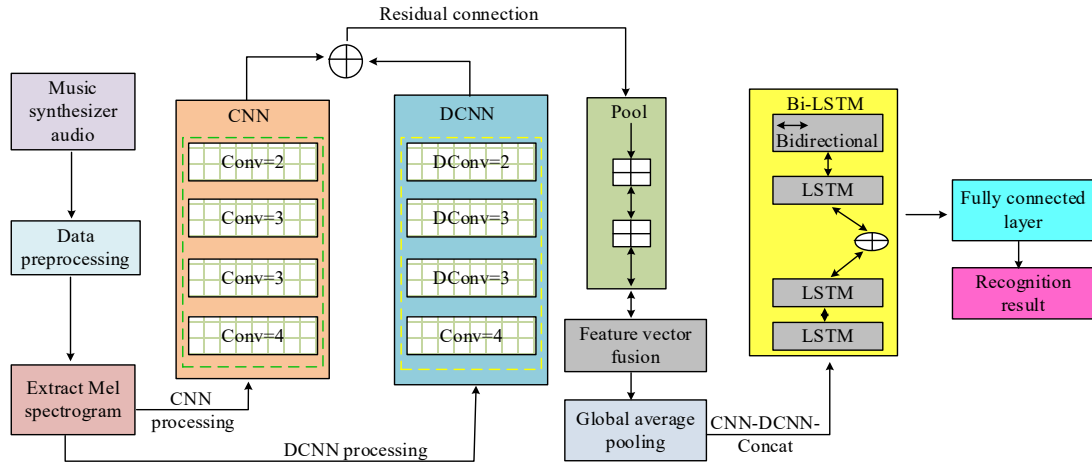
### 2.2. Design of an Instrument Recognition Algorithm Integrating Bi-LSTM and CNN-DCNN Models

The CNN-DCNN can effectively obtain important musical information by focusing on the "spatial" features of the spectrogram, but its ability to capture the "temporal" sequence attributes is limited. Therefore, the study introduced a Bi-LSTM based on CNN-DCNN to better extract temporal features from musical instrument audio data. The unidirectional

sequence characteristic of LSTM means that its output at a given time point depends only on the input at that time point and on previous time points (Tomar et al., 2025). Bi-LSTM captures contextual information from the past (forward) and future (backward) by combining a forward and a backward LSTM (Gelbukh et al., 2024). By concatenating the forward HS  $\overrightarrow{ht}$  and the backward HS  $\overleftarrow{ht}$  at time  $t$ , the final temporal feature representation  $ht = \overrightarrow{ht} \oplus \overleftarrow{ht}$  at that time can be obtained (Xiao et al., 2025). The recognition algorithm structure that integrates Bi-LSTM and CNN-DCNN models is shown in Fig. 1.

**Table 1.** Hierarchical structure of the CNN-DCNN parallel feature extraction module

| Branch                           | Layer type       | Core size/step            | Dilation rate | Number of output channels | Activation function | Normalization |
|----------------------------------|------------------|---------------------------|---------------|---------------------------|---------------------|---------------|
| Input (Shape: $128 \times 129$ ) | Mel spectrogram  | -                         | -             | 1                         | -                   | -             |
| Standard convolution branch      | Conv2D + BN      | $3 \times 3 / 1 \times 1$ | 1             | 64                        | ReLU                | Batch Norm    |
|                                  | Residual Block 1 | $3 \times 3 / 1 \times 1$ | 1             | 64                        | ReLU                | Batch Norm    |
|                                  | Residual Block 2 | $3 \times 3 / 1 \times 1$ | 1             | 128                       | ReLU                | Batch Norm    |
| Hollow convolution branch        | Conv2D + BN      | $3 \times 3 / 1 \times 1$ | 2             | 64                        | ReLU                | Batch Norm    |
|                                  | Residual Block 1 | $3 \times 3 / 1 \times 1$ | 4             | 64                        | ReLU                | Batch Norm    |
|                                  | Residual Block 2 | $3 \times 3 / 1 \times 1$ | 8             | 128                       | ReLU                | Batch Norm    |
| Fusion                           | Concatenation    | -                         | -             | 256                       | -                   | -             |



**Fig. 1.** Identification algorithm structure integrating Bi-LSTM and CNN-DCNN models

In Fig. 1, the output of the CNN-DCNN dual-branch structure is processed by residual, pooling and fusion, and then imported into Bi-LSTM. Bi-LSTM can extract temporal dependencies in instrument audio data, capture forward and backward relationships in music signals, and thus improve the performance of music-signal feature recognition. However, melodic instruments account for a large proportion of the performance market and energy in the overall music signal. Therefore, to better calculate the weight of the output at different times, the study uses the Self-Attention Mechanism (SAM)

to recognize and compute features, thereby reducing interference from irrelevant information. The attention network will perform a weighted sum over the feature vectors and obtain the final recognition result using the sigmoid function (Lekshmi and Rajan, 2025). Eq. (5) is the mathematical expression of the attention weight  $\partial n$  after normalization calculation.

$$\partial n = \frac{\exp(an)}{\sum_{n=1}^N \exp(an)} \quad (5)$$

In Eq. (6),  $an$  represents the attention score. Before sending the feature maps extracted by CNN-DCNN to Bi-LSTM, dimension transformation is required. Specifically, after being processed by the CNN-DCNN module and the max pooling layer, a four-dimensional feature tensor with shape (Batch, Channels,  $\text{Freq}_{\text{dim}}$ ,  $\text{Time}_{\text{dim}}$ ) is obtained. Average pooling is performed on the frequency dimension to obtain a three-dimensional tensor with the shape of (Batch, Channels,  $\text{Time}_{\text{dim}}$ ). Subsequently, the dimension is transposed to form a feature sequence with time steps as the sequence unit, which serves as the input to the Bi-LSTM. The self-attention mechanism used in the study is the scaled dot-product attention. For the hidden state sequence output by the Bi-LSTM at each time step, the query  $Q$ , key  $K$ , and value  $V$  Matrices are first transformed using three different linear transformations, and the attention score can be computed using Eq. (6).

$$\text{Attention}(Q, K, V) = \text{Soft max}(QK^T / \text{Sqrt}(dk))V \quad (6)$$

In Eq. (6),  $dk$  is the dimension of the key vector is used for scaling to prevent gradients from being too small,  $T$  is the transpose of the matrix, and the resulting context vector is the weighted sum of all time-step value vectors. This vector is then fed into the fully connected layer for the final classification decision. This enables the model to dynamically focus on the time frames that contribute the most to the recognition task. Fig. 2 shows a schematic diagram of the final improved fused CNN-DCNN-Bi-LSTMA M network structure.

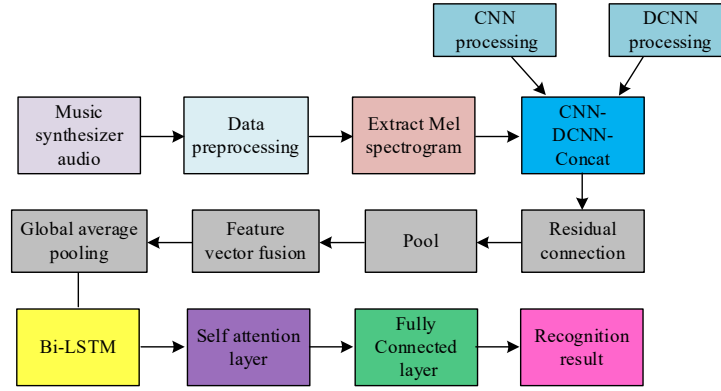


Fig. 2. Schematic diagram of the fusion CNN-DCNN-Bi-LSTM-AM network structure

In Fig. 2, CNN-DCNN extracts multi-instrument features from music, Bi-LSTM captures temporal dependencies in the music sequence, the self-attention mechanism weights key musical information for scoring, and the model outputs recognition results through fully connected layers to further improve instrument recognition accuracy. In the fusion model, the modified linear unit is selected as the primary activation function, and batch normalization and regularization are applied to each convolutional layer.

### 3. Performance Test of Digital Music Multi-instrument Recognition Fusion of CNN-DCNN-Bi-LSTM-AM

#### 3.1. Experimental Environment and Parameter Settings

To better analyze the capability of the introduced algorithm, the researchers selected the high-quality multi-instrument music dataset MedleyDB for testing. This dataset contains a large number of music clips played by multiple instruments and provides instrument annotations accurate at each time point. The study selected audio clips containing six common instruments (Piano, Guitar, Violin, Flute, Drums, and Saxophone). Audio clips containing at least one of these six instruments were selected. All audio clips were uniformly resampled to 44.1kHz. Continuous audio streams were divided into 3-second segments with a 1.5-second overlap to increase data volume and ensure note integrity. If multiple target instruments are present within a 3-second segment, the segment is labeled as a multi-label sample. All processed audio clips were randomly divided into training and test sets in a 6:4 ratio to ensure segmentation was performed at the song level, preventing clips from the same song from appearing in both sets simultaneously and preventing data leakage. The final dataset contained approximately 25,000 training samples and 5,500 test samples. To analyze the distribution of musical instruments in the dataset, the study counted the number of 3-second segments containing each instrument in the training set. The sample size of pianos was 7,850, accounting for 31.4%; The sample size of guitars was 6,525, accounting for 26.1%; The sample size of violin was 4,550, accounting for 18.2%; The sample size of the flute was 3,125, accounting for 12.5%; The sample size of drums was 8200, accounting for 32.8%; The sample size of Sachs was 2,875, accounting for 11.5%. There is a certain degree of category imbalance in the dataset, and, based on this, the research mainly used indicators such as macro-averaging to evaluate model performance. All experiments were conducted in a unified hardware and software environment. The experimental environment configuration for the research was as follows: In terms of hardware, Intel® Core™ i9-12900K CPU and NVIDIA GeForce RTX 3090 GPU with 24GB of graphics memory and 64GB of DDR5

memory were used. In terms of software, on Ubuntu 20.04, Python 3.8 and PyTorch 1.12.1 were used, supplemented by libraries such as Librosa 0.9.2, NumPy 1.21.5, and SciPy 1.7.3. In terms of parameter settings, the audio was uniformly processed at a sampling rate of 44100 Hz, with an analysis window length of 3 seconds and 128 Mel filters. The dilated convolution kernels in the model were 3x3, with dilation rates of 1, 2, 4, and 8. The number of hidden units in Bi-LSTM was 256, and the model was trained with an initial learning rate of 0.001 and a batch size of 32.

Simultaneously, indicators such as accuracy, macro-precision, Mean Average Precision (mAP), error value, and computational cost were selected for model validation. The macro averaging indicator independently computes category-specific indicators and averages them, so that categories with smaller sample sizes receive the same weight as those with larger sample sizes, enabling a fairer evaluation of the model's overall performance on imbalanced datasets.

### 3.2. Performance Evaluation of CNN-DCNN-Bi-LSTM-AM Algorithms

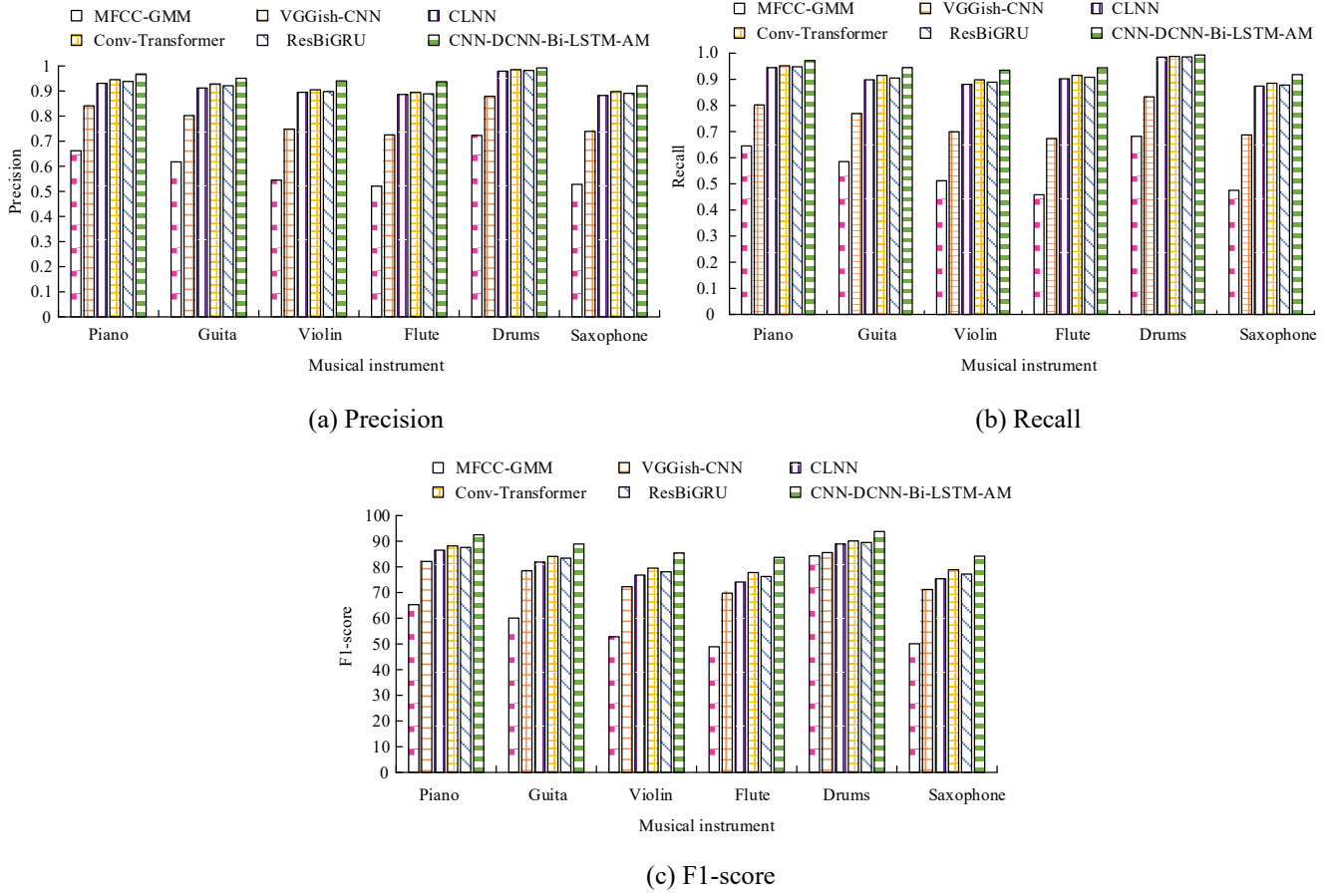
To verify the effectiveness of the proposed CNN-DCNN-Bi-LSTM-AM recognition algorithm, it was compared with various baseline models, including Mel Frequency Cepstrum Coefficient-Gaussian Mixture Model (MFCC-GMM) (Germen and Karadoğan, 2024), Visual Geometry Group Network (VGGish-CNN; He, 2024), Convolutional Neural Network Long Short-Term Memory (CLNN) (Wang and Dai, 2025), Convolutional Transformer (Conv-Transformer) (Yang et al., 2024), and ResNet-50-Bi-GRU (ResBiGRU) (Vaithianathan, 2024). To ensure fairness in comparison, all baseline models were retrained or evaluated under a unified framework using the same audio preprocessing, data augmentation, and segmentation strategies as the model proposed in the study. The performance of various algorithms on the multi-instrument recognition task was then compared, and the outcomes are presented in Table 2.

**Table 2.** Comprehensive performance comparison of various models in multi-instrument recognition tasks

| Model               | Accuracy (%) | Macro-Precision (%) | Macro-Recall (%) | Macro-F1 (%) | mAP (%) | Hamming Loss |
|---------------------|--------------|---------------------|------------------|--------------|---------|--------------|
| MFCC-GMM            | 58.34        | 55.12               | 52.89            | 53.45        | 54.78   | 0.1987       |
| VGGish-CNN          | 76.89        | 74.56               | 73.21            | 73.67        | 75.12   | 0.1245       |
| CLNN                | 81.45        | 79.88               | 78.34            | 79.05        | 80.23   | 0.0989       |
| Conv-Transformer    | 84.12        | 82.45               | 81.67            | 82.05        | 83.56   | 0.0856       |
| ResBiGRU            | 83.56        | 81.89               | 81.12            | 81.50        | 82.78   | 0.0889       |
| CNN-DCNN-Bi-LSTM-AM | 88.92        | 87.45               | 86.78            | 87.11        | 88.34   | 0.0623       |

Table 2 shows that the introduced CNN-DCNN-Bi-LSTM-AM model significantly outperformed all comparison models across all metrics, achieving an accuracy of 88.92% and a macro-average precision of 87.45%, which were 30.58% (accuracy) and 33.66% (macro-average precision) higher than the MFCC-GMM model, respectively. Conv-Transformer (Accuracy 84.12%) and ResBiGRU (Accuracy 83.56%) showed similar performance, whereas the MFCC-GMM model, relying solely on statistical features, failed to capture the complex dynamics of instrument sounds, resulting in an average accuracy of only 54.78%. VGGish-CNN (Accuracy: 76.89%), which is based on pre-trained audio features, significantly outperformed MFCC-GMM. Due to the unique local harmonic structure and long-range dependency characteristics of music signals, the Conv-Transformer and ResBiGRU models still performed slightly worse than the specifically designed CDBLA model. Analysis of the music recognition performance of different algorithms is shown in Fig. 3.

In Fig. 3(a), the traditional MFCC-GMM model had the lowest accuracy, particularly for the flute and saxophone, indicating difficulty in capturing the subtle acoustic features of these instruments. The macro average accuracy of the VGGish CNN was 0.789. The research model performed particularly well on piano and drum, achieving accuracies of 0.968 and 0.992, respectively, with a macro average accuracy of 0.952, surpassing AST's 0.938. The macroscopic average recall rate of MFCC-GMM in Fig. 3(b) was only 0.560. CRNN, Conv Transformer, and ResBiGRU had lower recall rates on saxophones and violins, which may be related to their timbres being easily masked in complex mixing. The average recall rate of the CNN-DCNN Bi-LSTM-AM model was 0.951. In Fig. 3(c), the CNN-DCNN Bi-LSTM-AM model achieved the highest F1 score for identifying all six instruments, outperforming the other models. The results of music recognition errors and parameters of different algorithms were compared and analyzed, and the results are shown in Table 3.



**Fig. 3.** Recognition performance of different algorithms in different musical instruments

Table 3 shows that the MFCC-GMM model, relying solely on shallow cross-cue matching, achieved MSE and MAE values of 0.312 and 0.443, respectively, at the single-instrument level. The ResBiGRU model had a training MSE error of 0.292 and a testing MSE error of 0.309 for a single instrument. The CNN-DCNN-Bi-LSTM-AM fusion model achieved a single-instrument MSE of 0.248 and a composite-instrument RMSE of 0.498 on the test set, representing reductions of 9.8% and 5.0%, respectively, compared to the second-best ResBiGRU, and exhibited the smallest training-test difference, demonstrating good robustness. The CNN-DCNN-Bi-LSTM-AM model had a medium-to-high parameter count (42.7M) and computational cost (1.35 GMacs), with a significantly lower parameter count than the ResBiGRU and VGGish-CNN models. In terms of inference time, the CNN-DCNN-Bi-LSTM-AM model (31.5ms) outperformed the equally complex Conv-Transformer and ResBiGRU, indicating that while achieving superior performance, the CDBLA model incurred a certain computational cost, but its complexity remained within a manageable range. The proposed brandification model had the lowest Hamming loss (0.0623) and the highest Jaccard similarity coefficient (81.56%), indicating that its predicted label set was closest to the true label set, with the fewest mislabeling and missed labeling.

### 3.3. Music Melody Recognition Results of the CNN-Bi-LSTM-AM Algorithm

The PR curves of different models for single and compound instruments were analyzed to better assess their errors in classifying musical melodies. The outcomes are presented in Fig. 4.

Fig. 4 shows that for the two instrument recognition results, the PR curve of the fusion model was closer to the upper-right corner of the coordinate axis, with average accuracies of 86.24% and 88.23%, respectively. Other comparative models showed more significant numerical variations for individual instruments and exhibited at least a 0.73% classification error compared to the fusion model for composite instruments. The proposed fusion model was analyzed using ablation experiments, and the results are shown in Fig. 5.

In Fig. 5, compared with the model using only standard CNN as the baseline, the introduction of CNN-DCNN structure improved the accuracy of the model from about 80% to 83.21%, indicating that the standard convolution branch finely characterizes the local continuous texture of the spectrum, while the dilated convolution branch effectively captures broader spectral context information through its exponentially growing receptive field; The introduction of Bi-LSTM module significantly increased the model accuracy to 87.05%. Effectively capturing long-range dependencies at the level of musical phrases, such as the duration and decay patterns of notes, as well as the direction of melodies, fills the gap in temporal understanding of pure CNN models. The addition of a self-attention mechanism further increased the accuracy to a peak of 88.92%. This indicated that even after Bi-LSTM effectively modeled temporal information, not all time frames were equally important. The attention mechanism simulated the focusing ability of human hearing by dynamically assigning weights

based on the Bi-LSTM's hidden states at different time steps, thereby automatically amplifying the influence of key information frames and suppressing interference from irrelevant or ambiguous information.

**Table 3.** Music recognition errors and parameter calculation results of different algorithms

| Single instrument                                                   |                        |               |                     |              |                        |       |
|---------------------------------------------------------------------|------------------------|---------------|---------------------|--------------|------------------------|-------|
| Model                                                               | Training set           |               |                     | Test set     |                        |       |
|                                                                     | MSE                    | RMSE          | MAE                 | MSE          | RMSE                   | MAE   |
| MFCC-GMM                                                            | 0.285                  | 0.534         | 0.419               | 0.302        | 0.549                  | 0.435 |
| VGGish-CNN                                                          | 0.273                  | 0.522         | 0.407               | 0.290        | 0.538                  | 0.423 |
| CLNN                                                                | 0.265                  | 0.515         | 0.401               | 0.282        | 0.531                  | 0.417 |
| Conv-Transformer                                                    | 0.258                  | 0.508         | 0.394               | 0.275        | 0.524                  | 0.410 |
| ResBiGRU                                                            | 0.292                  | 0.540         | 0.425               | 0.309        | 0.556                  | 0.441 |
| CNN-DCNN-Bi-LSTM-AM                                                 | 0.231                  | 0.481         | 0.371               | 0.248        | 0.498                  | 0.387 |
| Multiple instruments                                                |                        |               |                     |              |                        |       |
| Model                                                               | Training set           |               |                     | Test set     |                        |       |
|                                                                     | MSE                    | RMSE          | MAE                 | MSE          | RMSE                   | MAE   |
| MFCC-GMM                                                            | 0.285                  | 0.534         | 0.419               | 0.302        | 0.549                  | 0.435 |
| VGGish-CNN                                                          | 0.273                  | 0.522         | 0.407               | 0.290        | 0.538                  | 0.423 |
| CLNN                                                                | 0.265                  | 0.515         | 0.401               | 0.282        | 0.531                  | 0.417 |
| Conv-Transformer                                                    | 0.258                  | 0.508         | 0.394               | 0.275        | 0.524                  | 0.410 |
| ResBiGRU                                                            | 0.292                  | 0.540         | 0.425               | 0.309        | 0.556                  | 0.441 |
| CNN-DCNN-Bi-LSTM-AM                                                 | 0.231                  | 0.481         | 0.371               | 0.248        | 0.498                  | 0.387 |
| Model complexity, computational efficiency, and multi-label results |                        |               |                     |              |                        |       |
| Model                                                               | Parameter quantity (M) | FLOPs (MMacs) | Inference time (ms) | Hamming loss | Jaccard Similarity (%) |       |
| MFCC-GMM                                                            | 0.12                   | 5.2           | 1.2                 | 0.1987       | 48.91                  |       |
| VGGish-CNN                                                          | 64.8                   | 1,256.3       | 15.6                | 0.1245       | 65.43                  |       |
| CRNN                                                                | 28.5                   | 845.7         | 22.3                | 0.0989       | 71.28                  |       |
| Conv-Transformer                                                    | 48.2                   | 1,589.4       | 35.8                | 0.0856       | 75.12                  |       |
| ResBiGRU                                                            | 68.1                   | 2,145.6       | 45.1                | 0.0889       | 74.45                  |       |

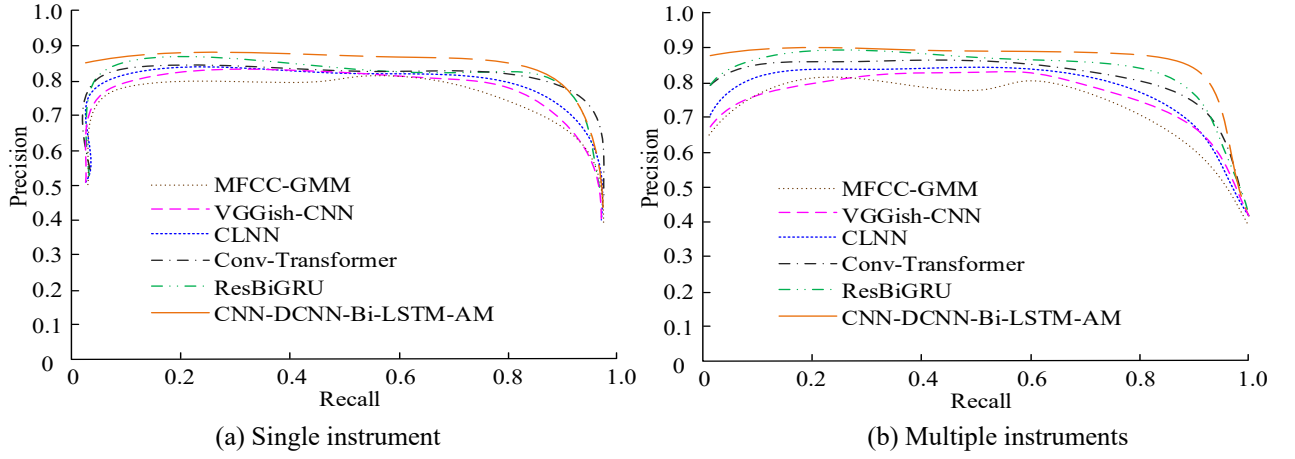


Fig. 4. PR curves of different models on single and composite instruments

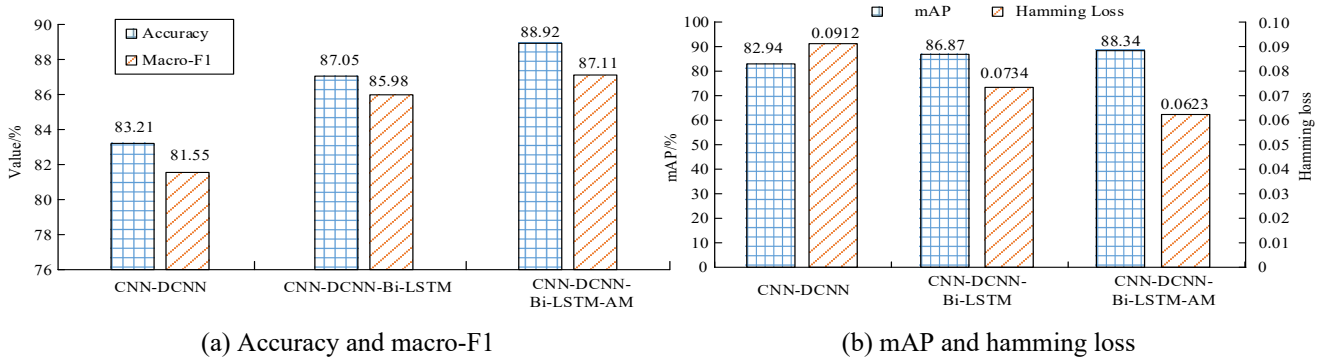


Fig. 5. Ablation results

#### 4. Discussion

To address the challenges of instrument recognition in multi-instrument mixing scenarios in digital music, a CNN-DCNN Bi-LSTM-AM recognition model was designed. The research results showed that the accuracy of the traditional MFCC-GMM model was 58.34%, primarily relying on manually designed statistical acoustic features that cannot capture the complex two-dimensional textures and dynamic evolution patterns of instrument timbre in the time spectrum. The VGGish CNN model demonstrated the powerful ability of deep convolutional networks to automatically learn hierarchical features from data, but it irreversibly reduced the resolution of feature maps when processing audio spectrograms, which may result in blurring or loss of subtle harmonic structural information that is crucial for distinguishing similar musical instruments (Luo et al., 2023). Models such as CLNN, Conv-Transformer, and ResBiGRU improved performance to 81%-84% by incorporating recurrent or self-attention mechanisms to capture temporal dependencies. The proposed CNN-DCNN-Bi-LSTM-AM model achieved an accuracy of 88.92% and a Macro-F1 score of 87.11%, outperforming other comparative models. Its CNN-DCNN module constructed a more powerful spatial feature extractor, and then utilized Bi-LSTM and Attention mechanisms for more refined temporal analysis.

The focus of the standard convolution branch and the extended convolution branch on features is mainly on local and generalized aspects, with the latter capturing information via the dilation rate without losing resolution. Wang et al. (2023) also suggest that dilated convolution can effectively integrate multi-scale information in tasks such as semantic segmentation. The parallel fusion design avoided the "grid effect" that pure dilated convolution may cause due to sparse sampling, while retaining the fine characterization of local details by standard convolution, thus generating more robust and discriminative multi-scale feature representations (Lekshmi and Dandamudi, 2025). After adding Bi-LSTM to CNN-DCNN, the model's accuracy improved significantly to 87.05%. This showed that even when CNN-DCNN extracted high-quality frame-level features, ignoring temporal order and dependencies between them still led to a significant bottleneck in recognition performance. The attention mechanism could automatically amplify the influence of key information frames by weighing the Bi-LSTM outputs at different time steps, while suppressing interference from irrelevant or ambiguous information. The performance improvement in research methods had clear operational significance, with a 4-5 percentage-point increase in the macro-average F1 score and a significant reduction in Hamming loss. This means that designers can further optimize their designs based on user results from instrument-based search and automatic playlist generation, thereby increasing user stickiness, satisfaction, and the platform's core competitiveness. Research has shown that the fusion architecture combining parallel multi-scale convolution and temporal modeling can achieve significant performance improvements while incurring only an acceptable increase in computational cost, which is highly advantageous for music retrieval platform managers seeking to upgrade music analysis systems and achieve high-precision instrument recognition.

#### 5. Conclusion

The fusion model designed in this study could extract robust, information-rich multi-scale spatial features from Mel spectrograms, demonstrating strong performance in overall accuracy, instrument recognition, and model efficiency. The research results can provide technical selection references for decision-makers, and they will be more inclined to support and invest in the development of hybrid deep learning models that can simultaneously handle multi-scale spatial features and time-series dependencies. Moreover, it can better achieve resource allocation and arrangement, promoting the team to scientifically measure the true value of algorithms to the business. However, some limitations remain. For example, the scalability of the model's generalization ability needs further testing, and its performance in real-time applications requiring extremely low latency (such as real-time track recognition for live performances and interactive music applications) needs improvement. Therefore, future research should develop a lighter version while maintaining high accuracy. Simultaneously, source separation models will be used to recognize sound source streams, further reducing the difficulty of the recognition task. Although MedleyDB is a high-quality and widely used multi-track dataset, it may have inherent preferences in instrument combinations, music styles, and mixing engineering. This means that the excellent performance of the research model on MedleyDB requires further validation of its ability to generalize to other datasets with different feature distributions. Future work should cross-validate models on more diverse datasets to comprehensively evaluate their robustness and generalization ability.

#### Author Contributions

Junhui Zhao contributed to conceptualization, methodology, investigation, data collection, draft preparation, and project administration. Xiaohang Jia contributed to methodology, software, validation, analysis, visualization, and manuscript editing. Yuxuan Zhou contributed to validation, formal analysis, investigation, data collection, and draft preparation.

#### Funding

This work was sponsored in part by the department-level, Hebei Province Youth Development Research "Research on the Generation and Alleviation of Occupational Anxiety Among Z-Generation Online Streamers--A Case Study of Graduates from Art Colleges" (No.: 2025HBQN74).

#### Institutional Review Board Statement

Not applicable.

#### Declaration of Artificial Intelligence (AI) Tools

The authors confirm that no AI tools were used in the preparation of this manuscript.

#### References

- Gelbukh, A., Pérez Alvarez, D. A., Kolesnikova, O., Chanona-Hernández, L., and Sidorov, G. (2024). Multi-instrument based N-grams for composer classification task. *COMPUT SIST*, 28, 1, 85–98. doi: 10.13053/cys-28-1-4903.
- Germen, E. and Karadoğan, C. (2024). End-to-end automatic music transcription of polyphonic qanun and oud music using deep neural network. *Eskisehir Tech Univ J Sci Technol A Appl Sci Eng*, 25, 3, 442–455. doi: 10.18038/estubtda.1467350.
- Green, G., Kennedy, P., and McGown, A. (2002). Management of multi-method engineering design research: A case study. *Journal of Engineering and Technology Management*, 19, 2, 131–140. doi: 10.1016/S0923-4748(02)00006-1.
- Guo, R. (2024). Research on Neural Network-based Automatic Music Multi-Instrument Classification Approach. *IJACSA*, 15, 1, 1318–1324. doi: 10.14569/IJACSA.2024.0150179.
- He, J. (2024). Research on the Construction of Music Recognition Model Based on Improved CNN Convolutional Neural Network. *APPL COMPUT ENG*, 96, 1, 195–204. doi: 10.54254/2755-2721/96/20241394.
- Ji, S., Yang, X., and Luo, J. (2023). A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM COMPUT SURV*, 56, 1, 1–39. doi: 10.1145/3597493.
- Lekshmi, C. R. and Rajan, R. (2025). Beyond transformers: hierarchical contextualization and gated aggregation for multiple predominant instrument recognition in polyphonic music. *J SUPERCOMPUT*, 81, 8, 921. doi: 10.1007/s11227-025-07279-7.
- Lekshmi, C. R., and Dandamudi, J. T. (2025). Dynamic Feature Learning with Involution and Convolution for Predominant Instrument Recognition in Polyphonic Music. *CIRC SYST SIGNAL PR*, 44, 9, 6698–6715. doi: 10.1007/s00034-025-03111-y.
- Li, L. and Han, Z. (2023). Design and innovation of audio iot technology using music teaching intelligent mode. *Neural Comput Applic*, 35, 6, 4383–4396. doi: 10.1007/s00521-022-07025-7.
- Li, X., Yan, Y., Soraghan, J., Wang, Z., and Ren, J. (2023). A music cognition-guided framework for multi-pitch estimation. *Cogn Comput*, 15, 1, 23–35. doi: 10.1007/s12559-022-10031-5.
- Li, P. and Wang, B. (2024). Artificial intelligence in music education. *Int J Hum Comput Interact*, 40, 16, 4183–4192. doi: 10.1080/10447318.2023.2209984.
- Liu, Y. (2025). Deep Neural Network and SVM-Based Multiclass Musical Instruments Classification Using MFCC, STFT, and Related Acoustic Features. *Informatica*, 49, 33, 197–212. doi: 10.31449/inf.v49i33.8730.
- Luo, N., Yu, H., You, Z., Li, Y., Zhou, T., and Han, N. (2023). Fuzzy logic and neural network-based risk assessment model for import and export enterprises: A review. *J Data Sci Intell Syst*, 1, 1, 2–11. doi: 10.47852/bonviewJDSIS32021078.
- McCarthy, J. F., Nguyen, D. H., Rashid, A. M., and Soroczak, S. (2003). Proactive Displays & The Experience UbiComp Project. *Proceedings of the Fifth International Conference on Ubiquitous Computing, Seattle, Washington* 78–81.
- Reghunath, L. C. and Rajan, R. (2023). Predominant audio source separation in polyphonic music. *J AUDIO SPEECH MUSIC PROC*, 2023, 1, 49. doi: 10.1186/s13636-023-00316-4.

- Richard, G., Smaragdis, P., Gannot, S., Naylor, P. A., Makino, S., Kellermann, W., and Sugiyama, A. (2023). Audio signal processing in the 21st century: The important outcomes of the past 25 years. *IEEE SIGNAL PROC MAG*, 40, 5, 12–26. doi: 10.1109/MSP.2023.3276171.
- Sun, Y. (2023). Timbre-Based Portable Musical Instrument Recognition Using LVQ Learning Algorithm. *MOBILE NETW APPL*, 28, 6, 2171–2181. doi: 10.1007/s11036-023-02174-y.
- Tomar, H., Deshwal, D., and Trivedi, N. (2025). Convolutional neural network based language identification system: A spectrogram based approach. *MULTIMED TOOLS APPL*, 84, 24, 28951–28976. doi: 10.1007/s11042-024-20283-y.
- Vaithianathan, M. (2024). Digital Signal Processing for Noise Suppression in Voice Signals. *IJARISE*, 1, 4, 198–208. doi: 10.61359/11.2206-2417.
- Verzuh, E. (2008). The Fast Forward MBA in Project Management. Hoboken, New Jersey: Wiley and Sons, pp. 55–60.
- Wang, P. and Dai, N. (2025). Processing piano audio: research on an automatic transcription model for sound signals. *J MEAS ENG*, 13, 1, 130–139. doi: 10.21595/jme.2024.24345.
- Wang, J., Liu, H., Ying, H., Qiu, C., Li, J., and Anwar, M. S. (2023). Attention-based neural network for end-to-end music separation. *CAAI Transactions on Intelligence Technology*, 8, 2, 355–363. doi: 10.1049/cit2.12163.
- Wheeler, D. P. and Bragin, M. (2007). Bringing it all back home: Social work and the challenge of returning veterans. *Health and Social Work*, 32, 297–300. Retrieved from <http://www.naswpressonline.org> on February 1, 2011.
- Xiao, S., Xiong, S. L., Cai, C., Song, L. M., Zheng, S. J., Peng, W. X., Wang, P., Qiao, R., Guo, D. Y., and Wang, J. (2025). Energetic Transients Joint Analysis System for Multi-INstrument (ETJASMIN) for GECAM. II. Search, Verification, and Classification of Bursts. *ASTROPHYS J SUPPL S*, 277, 1, 9. doi: 10.3847/1538-4365/adacdf.
- Yang, Y., Qiu, Z., Shi, Y., Sun, S., Wang, C., Pan, D., Zhao, J., Liang, Z., Wang, Y., Li, X., Yu, F., Yu, T., and Dai, Q. (2024). Audio matters too! enhancing markerless motion capture with audio signals for string performance capture. *ACM T GRAPHIC*, 43, 4, 1–10. doi: 10.1145/3658235.



Junhui Zhao was born in Hebei Province in May 1981 and is of Han nationality. He holds an associate professor position. He earned his bachelor's degree in music performance from the China Conservatory of Music in 2005 and his Master of Arts in Vocal Music and Opera from the same institution in 2009. Since 2009, he has been a faculty member at the College of Art, Hebei University of Economics and Business. His research focuses on opera performance. He has led two provincial and ministerial-level research projects and participated in over ten projects at provincial, ministerial, and university levels. He has edited one monograph and published or co-authored more than 20 academic papers, including one CSSCI core journal paper, one EI conference paper, and two SSCI-indexed papers as first author.



Xiaohang Jia was born in Xingtai, Hebei, in October 2001. He is of Han nationality. He received his bachelor's degree in musicology from Hebei North University in 2023 and has been a postgraduate student in Music Performance at Hebei University of Economics and Business since 2025.



Yuxuan Zhou was born in Nanjing, Jiangsu, in August 2003. Yuxuan Zhou is of Han nationality. She received her bachelor's degree in music education from Taizhou University in 2025 and has been a postgraduate student in Vocal Performance at Hebei University of Economics and Business since 2025.