Advances in Applied Mathematics 应用数学进展, 2025, 14(8), 147-159
Published Online August 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.148379

Hans汉斯

基于RF-CNN-CBAM-BiLSTM的兴山民歌分类研究

白雨欣[1]，刘依林[1]，雷萌非[2]，肖维维[1*]

[1]北方工业大学理学院，北京
[2]北京市海淀外国语藤飞学校，北京

收稿日期：2025年7月12日；录用日期：2025年8月5日；发布日期：2025年8月14日

摘　要

本研究构建了RF-CNN-CBAM-BiLSTM算法，并对兴山民歌进行分类识别。该模型先利用随机森林(Random Forest, RF)降维方法替代原始的经验选择来挑选训练所用特征，再将卷积块注意力模块(Convolutional Block Attention Module, CBAM)模块融入卷积神经网络(Convolutional Neural Network, CNN)架构，增强模型特征关注与识别能力，接着利用双向长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)捕捉音频序列双向上下文信息，提升兴山民歌的识别性能。本研究所提模型的识别准确率达到91.67%，每一轮的运行时间为22 s，其准确率超过残差网络(Residual Network 50, ResNet50)，高效网络(Rethinking Model Scaling for Convolutional Neural Networks, EfficientNetV1)等四种基准模型的平均准确率，约5.24%。每一轮运行速度较四种基准模型的平均运行时间缩短68.25 s。

关键词

音频分类，注意力机制，深度学习

A Study on the Classification of Xingshan Folk Songs Using RF-CNN-CBAM-BiLSTM

Yuxin Bai[1], Yilin Liu[1], Mengfei Lei[2], Weiwei Xiao[1*]

[1]College of Science, North China University of Technology, Beijing
[2]Tengfei School of Haidian Foreign Language Academy, Beijing

Received: Jul. 12th, 2025; accepted: Aug. 5th, 2025; published: Aug. 14th, 2025

*通讯作者。

文章引用: 白雨欣, 刘依林, 雷萌非, 肖维维. 基于 RF-CNN-CBAM-BiLSTM 的兴山民歌分类研究[J]. 应用数学进展, 2025, 14(8): 147-159. DOI: 10.12677/aam.2025.148379

Abstract

This study proposes an audio classification algorithm based on the RF-CNN-CBAM-BiLSTM architecture for the classification and recognition of Xingshan folk songs. The model first employs Random Forest (RF) for feature dimensionality reduction, replacing traditional empirical feature selection. Then, the Convolutional Block Attention Module (CBAM) is integrated into the Convolutional Neural Network (CNN) architecture to enhance the model's ability to focus on and extract relevant features. Subsequently, a Bidirectional Long Short-Term Memory (BiLSTM) network is utilized to capture bidirectional contextual information in audio sequences, further improving the recognition performance of Xingshan folk songs. The proposed model achieves a recognition accuracy of 91.67%, with an average runtime of 22 seconds per training epoch. Its accuracy surpasses the average performance of four benchmark models, including the Residual Network (ResNet50) and EfficientNetV1, by approximately 5.24%. Additionally, the average runtime per epoch is reduced by 68.25 seconds compared to these benchmarks.

Keywords

Audio Classification, Attention Mechanism, Deep Learning

1. 引言

在非物质文化遗产保护领域，民歌作为非物质文化遗产的重要载体，其数字化保护需求日益迫切。以国家级非物质文化遗产“兴山民歌”为例，其特有的“兴山三度音程”是中国传统音乐体系的独特代表。但是随着经济全球化和外来文化的冲击，兴山民歌面临着日益严峻的考验，在非物质文化遗产保护的时代背景下，数字化技术为兴山民歌传承提供了创新路径，可实现传统民歌的系统性保护与动态化管理。

随着计算机与数字多媒体技术的进步，音频分类任务得到研究人员的重视[1]。人工音频分类工作主要基于专业人士利用耳朵和专业知识进行识别，然而民歌体系具有历史积淀深厚，曲目数量庞大的特点，若延续传统人工标注模式进行音频筛选与分类，将面临标注效率低下与人力资源消耗过高的双重挑战。

当下，针对兴山民歌识别的研究相对匮乏，音频分类领域的研究成果却颇为丰硕。考虑到兴山民歌本质上属于音频范畴，故而在开展其分类研究时，可借鉴音频分类的相关方法与思路。在音频分类研究领域，主要分为传统分类方式和基于深度学习的分类方式[2]。传统音频分类方法主要依赖音乐理论体系与音频结构特征的深度解析，其核心在于对音乐旋律，节奏，和声等核心元素的规律性分析。这类方法主要有两个环节，特征提取和分类。在特征提取模块，多数研究人员依据音频信号在音色及节奏方面的差异，开展时域与频域分析，从而获取特征参数[3]-[5]，并在处理结构相对简单，规律性较强的音频数据时表现出良好的性能。深度学习方法则通过建立复杂的神经网络模型，提取音频的深层特征，因此适用于处理复杂的音频类型和大规模的音频数据集[6]-[8]。利用深度神经网络实现兴山民歌特征提取与精准分类，既具技术挑战，又为非物质文化遗产保护提供新路径。

将深度学习应用于音频分类领域，基于音频特征提取和分类研究已取得阶段性进展，但是还存在着

两方面局限：其一，现有的研究多依赖于提取单一域特征，缺乏针对民歌音乐特性的特征提取机制。其二，民族音乐分类领域研究匮乏。兴山民歌作为即将消亡的非物质文化遗产，其独特音乐形态与稀缺数据对分类技术提出更高要求。值得注意的是，基于双向长短期记忆网络(BiLSTM)的音频分类框架已在通用音乐分类任务中展现出优异性能，为兴山民歌识别提供了可借鉴的技术路径。然而现有研究存在技术瓶颈：首先，多数研究直接采用梅尔频率倒谱系数(MFCC)，过零率(ZCR)等通用音频特征进行模型训练，缺乏针对民歌音乐特性的特征提取机制；其次，传统分类方法未充分捕捉音频序列的长时依赖关系，特别是在处理具有复杂节奏结构的民歌音频时，未能有效构建相邻时间帧之间的语义关联；此外，现有模型普遍未引入注意力机制对关键时间片段进行聚焦分析，导致在处理包含大量非特异性背景噪声的民歌录音时分类性能下降。基于以上研究和分析，本文提出一种针对兴山民歌的识别算法，解决已有技术的局限性，构建了包含 1800 首民歌样本的音频数据集，并详细阐述了音频预处理流程，特征提取方法以及针对 BiLSTM 音频分类模型提出的改进模型 RF-CNN-CBAM-BiLSTM。最后，介绍了音频识别的实验结果，实验部分通过准确率，F1 分数等 5 项指标，对改进模型的性能进行了量化评估，并与多种基准模型(如 ResNet50, EfficientNetV1 等)开展了兴山民歌识别效果的对比分析。

2. RF-CNN-CBAM-BiLSTM 算法流程

图 1 为 RF-CNN-CBAM-BiLSTM 分类模型流程图。

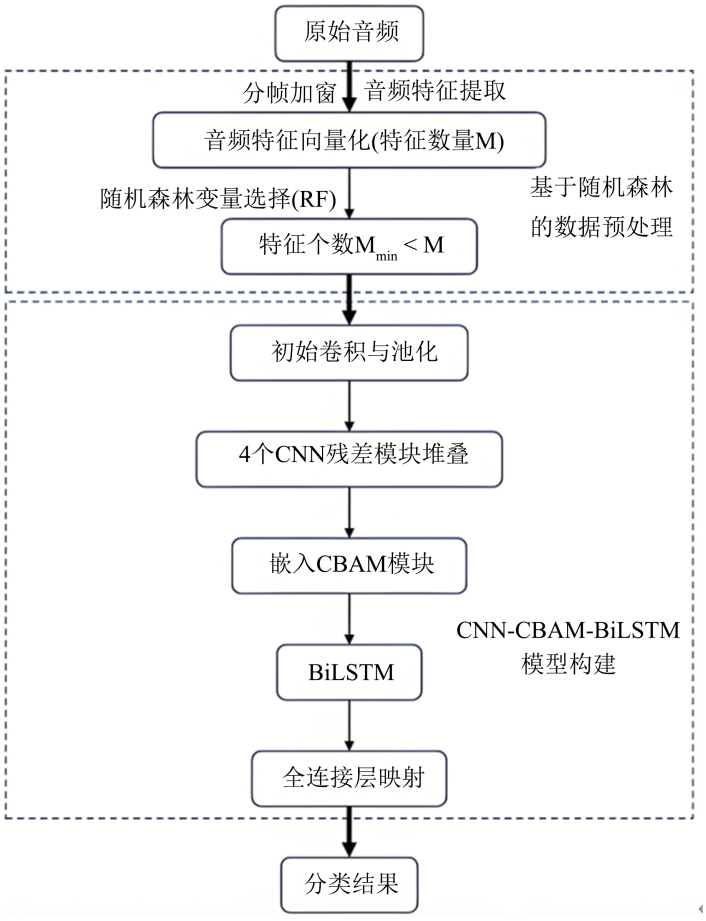

Figure 1. Flowchart of the RF-CNN-CBAM-BiLSTM classification model
图 1. 基于 RF-CNN-CBAM-BiLSTM 分类模型流程图

2.1. 数据预处理与特征提取

音频数据输入分类模型前，需要经过预处理，特征提取以及特征处理三个步骤[3]。预处理旨在对原始音频信号开展削减噪声干扰，提炼关键信息，让数据契合后续分析需求。针对长时音频数据，通常采用分帧处理与窗函数应用(如汉宁窗)实现信号的平稳化。特征提取环节则从预处理后的数据中提取多维度特征参数，涵盖时域(如过零率，短时能量)，频域(如频谱质心，带宽)及倒谱域(如梅尔频率倒谱系数，MFCC)等方面特征。特征处理旨在对提取的全部特征进行归一化操作，同时筛选出关键特征。音频分类基本流程如图 2 所示。

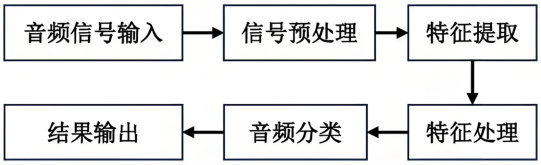

Figure 2. Basic workflow of audio classification
图 2. 音频分类基本流程

本研究收集，整理的数据集涵盖了兴山民歌与其他九种类型民歌。其中，其他类型民歌包含劳动号子，搬运号子，渔船号子，灯歌，小调，舞歌，田歌，儿歌以及生活音调这九种不同音乐流派，总计 1800 首音频数据作为实验样本。各音频类型及其对应的数量详情如表 1 所示。

Table 1. Audio types and quantities
表 1. 音频类型与数量

音乐流派	数量	合计
兴山民歌	900	900
其他类型民歌	9 * 100	900

为确保数据集具备良好的平衡性与多样性，共选取 1800 首音频作为研究样本。其中，兴山民歌 900 首，其他音乐类型，针对每种类型均严格选取 100 首。

为了保证输入模型的音频样本具有一致的长度，我们首先对原始音频数据进行了切分处理。具体而言，将每段音频切割为固定时长为 5 秒的样本，如图 3 所示。这一处理方式的选择基于多方面考虑：一方面，统一的音频长度能够确保模型输入的一致性，避免因音频长度差异过大而导致的训练不稳定问题；另一方面，5 秒的时长能够较好地保留音频中的关键特征信息，同时又不会因时长过长而增加计算成本。对于长度不足 5 秒的音频样本，我们采取了丢弃处理，而对于长度超过 5 秒的音频，则按照 5 秒的固定时长进行连续切割，确保不遗漏任何可能有用的信息。通过这种方式，我们构建了一个长度统一的音频样本数据集，为后续的特征提取和模型训练做好准备。

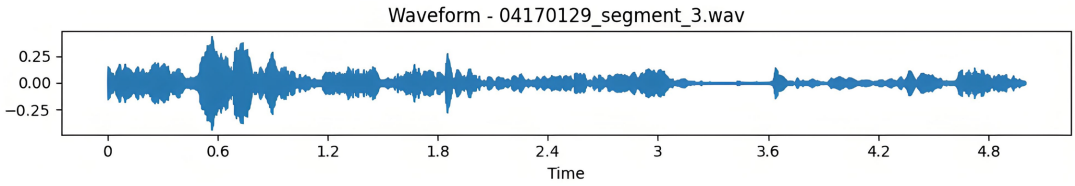

Figure 3. Mountain song audio waveform chart
图 3. 山歌音频波形图

音频波形图以可视化的方式展示了音频信号随时间变化的振幅情况。图中横轴表示时间，单位为秒；纵轴表示音频信号的振幅。波形的起伏代表了音频信号在不同时刻的强弱变化，波峰越高或波谷越低，则该时刻音频信号的振幅越大，声音就越响亮。而波形较为平缓的区域，说明音频信号振幅较小，声音较弱。通过观察音频波形图，可以大致了解音频的节奏，静音部分以及音量的变化情况等。

在音频分类任务中，分帧是特征提取前的必要预处理步骤。由于音频信号具有时变特性，直接分析整段语音较为困难，将其分帧后，每帧内信号可近似看作平稳信号，便于特征提取[9]。

本研究选取 20 ms 作为帧长，针对一段时长 5 s，采样率 16,000 Hz 的音频信号，经计算可分为 250 帧。分帧如式所示：

$$x_n(m) = x(n + mR), 0 \le m \le L - 1 \tag{1}$$

其中，$x_n(m)$ 为第 n 帧的第 m 个采样点；$x(n+mR)$ 为原始信号的第 n 帧的第 m 个采样点；R 为帧移；L 为每帧的采样点数，即帧长。

音频信号预处理阶段，加窗对准确提取音频特征至关重要。音频是时变信号，特征随时间变化大，特征提取时通常假设其短时间平稳，直接对分帧信号做频谱分析会产生频谱泄漏，导致频谱分辨率下降、无法反映真实频谱。加窗通过将帧信号与窗函数相乘，使信号边界平滑、减少不连续点，以此降低频谱泄漏。汉明窗与汉宁窗能有效抑制频谱泄漏现象[4]。

在本研究中，我们选择汉明窗(Hamming Window)作为加窗函数。汉明窗主瓣宽度适中，能够在一定程度上保证频率分辨率。旁瓣衰减较快，可以有效抑制频谱泄漏。其数学表达式为：

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), n = 0,1,\cdots,N-1 \tag{2}$$

其中，N 是窗函数的长度，也就是每帧音频信号的长度；n 是离散时间索引。

音频特征提取作为数字信号处理的关键环节，在音频分析、识别与处理等多方面有着广泛且重要的应用。在音频处理工作里，将原始音频信号转化为更具抽象性和代表性的特征是必要步骤，这些特征能够为诸如音乐分类、语音识别、情感分析等多样化的任务提供基础支撑。本研究从处理好的音频信号中提取 17 个特征，这 17 个特征涉及音频的时域、频域、倒谱域特征，可以从多维度描述音频，捕捉不同音频特性，为分类提供完整且丰富的支撑。多域特征融合可有效提升模型的泛化与区分能力，避免过拟合，使分类更精准。同时，依据不同音频分类需求灵活组合上述特征，在复杂环境中利用它们的抗噪优势，保证音频分类在多样场景下的效果。

时域特征：

1) 短时平均振幅(Short-Term Average Amplitude, STAA)反映了音频信号在短时间内的平均能量大小。其计算方法是先将音频信号分帧，然后计算每一帧信号绝对值的平均值。具体公式为：

$$STAA_i = \frac{1}{N}\sum_{n=0}^{N-1}\left|x_i(n)\right| \tag{3}$$

其中，$x_i(n)$ 表示第 i 帧的音频信号，N 为帧长。STAA 特征可以用于描述音频信号的响度变化，是语音识别和音乐分析中重要特征之一。

2) 短时能量(Short-Term Energy, STE)是衡量音频信号在短时间内能量集中程度的指标。它的计算方式与 STAA 类似，只是对每一帧信号先进行平方运算，再求和。计算公式如下：

$$STE_i = \sum_{n=0}^{N-1} x_i^2(n) \tag{4}$$

其中，$x_i(n)$ 为第 i 帧的音频信号，N 为帧长。STE 特征对于区分语音和静音部分非常有效，在语音端点检测等任务中有着广泛的应用。

频域特征：

1) 频谱质心(Spectral Centroid)描述了音频信号频谱的重心位置，反映了音频信号的明亮程度或音调高低。其计算方法是对每一帧信号的频谱幅度进行加权平均，权重为对应的频率。具体公式为：

$$SC_i = \frac{\sum_{k=0}^{N-1} k\left|X_i(k)\right|}{\sum_{k=0}^{N-1}\left|X_i(k)\right|} \tag{5}$$

其中，$X_i(k)$ 表示第 i 帧信号的频谱幅度，N 为 FFT 点数。频谱质心较高的音频信号通常听起来更加明亮、尖锐，而频谱质心较低的音频信号则更加低沉、柔和。

2) 频谱带宽(Spectral Bandwidth)衡量了音频信号频谱的分布宽度，它反映了音频信号的频率丰富程度。频谱带宽的计算基于频谱质心，其公式为：

$$SB_i = \sqrt{\frac{\sum_{k=0}^{N-1}\left(k - SC_i\right)^2\left|X_i(k)\right|}{\sum_{k=0}^{N-1}\left|X_i(k)\right|}} \tag{6}$$

其中，$X_i(k)$ 是第 i 帧信号的频谱幅度，SC_i 是第 i 帧的频谱质心，N 为 FFT 点数。频谱带宽较大的音频信号包含更多的频率成分，听起来更加丰富多样；而频谱带宽较小的音频信号则相对单一。

倒谱域特征：

梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)是音频处理中最为常用的特征之一，它模拟了人类听觉系统对不同频率声音的感知特性。提取梅尔谱倒谱系数除了分帧、加窗外还包括傅立叶变换、梅尔滤波器组、对数运算以及离散余弦变换四个步骤[10]。

快速傅里叶变换(FFT)，对加窗后的每一帧信号 $x_{i,w}(n)$ 进行 N 点快速傅里叶变换(FFT)，将时域信号转换为频域信号。FFT 的公式为：

$$X_i(k) = \sum_{n=0}^{N-1} x_{i,w}(n)\mathrm{e}^{-j\frac{2\pi}{N}kn}, k = 0,1,\cdots,N-1 \tag{7}$$

其中，$X_i(k)$ 是第 i 帧信号的频域表示，k 为频域索引。通常我们只关注频谱的幅度，即 $\left|X_i(k)\right|$。

梅尔滤波将频谱幅度 $\left|X_i(k)\right|$ 通过一组梅尔滤波器组。梅尔频率尺度是一种基于人类听觉感知的频率尺度，它与线性频率 f (单位：Hz)的转换关系为：

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right), f = 700\left(10^{\frac{m}{2595}}-1\right) \tag{8}$$

梅尔滤波器组通常由 M 个三角形滤波器组成(一般 M 取值在 20 到 40 之间)，每个滤波器对频谱幅度 $\left|X_i(k)\right|$ 进行加权求和，得到梅尔频谱 $S_i(m)$：

$$S_i(m) = \sum_{k=0}^{N-1}\left|X_i(k)\right|H_m(k), m = 0,1,\cdots,M-1 \tag{9}$$

对数运算与离散余弦变换(DCT)对梅尔频谱 $S_i(m)$ 取对数，得到对数梅尔频谱 $\log\left(S_i(m)\right)$。这一步是为了模拟人类听觉系统对声音强度的对数响应。然后对对数梅尔频谱进行离散余弦变换(DCT)，得到 MFCC 系数 $c_i(n)$：

$$c_i(n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log\left(S_i(m)\right)\cos\left[\frac{\pi n}{M}\left(m+\frac{1}{2}\right)\right],n=0,1,\cdots,L-1 \tag{10}$$

其中，L 是所取的 MFCC 的数量，在本研究中取前 13 个系数作为特征。

综上所述，从时域、频域和倒谱域三个方面对预处理后的每帧信号提取特征，时域提取 2 个能反映信号时间变化特性的特征，频域提取 2 个展示频率组成情况的特征，倒谱域提取 13 个有助于分离音频不同成分的子特征，每帧共提取 17 个特征。对于任意一个音频样本 S，分帧加窗后形成包含 250 个时间帧的时间向量 $S(t_1,t_2,\cdots,t_{250})$，经特征提取得到对应 17 个特征的特征向量 $S(a_1,a_2,\cdots,a_{17})$，二者组合形成 17 行 250 列的特征矩阵，1800 个样本处理后得到目标特征矩阵。

2.2. 特征降维

由于上述方法提取的特征维度较高，会增加训练复杂度和时间消耗，为此采用随机森林(RF)算法对所有特征打分，依据结果筛选重要特征实现降维，以降低训练复杂度、减少时间，使分类模型更高效地训练和预测。

RF 模型作为一种非线性降维手段，在降低特征维度过程中可有效留存原始特征信息。RF 算法能够通过其独特的决策树构建与投票机制，有效评估各个特征的重要性。通过 RF 进行特征选择，成功减少了特征数量，这使得模型在训练和预测过程中，需要处理的数据量大幅降低，从而显著减少了系统运行时间。

经过 RF 筛选后的特征子集，去除了对分类贡献较小的冗余特征，保留了与兴山民歌分类最为紧密相关的关键特征。这些精选特征使得模型能够更加专注于学习有效信息，避免了因无关特征干扰而导致的学习偏差。从实验结果来看，虽然特征数量减少，但模型的准确率较之前反而更高。这表明 RF 不仅实现了数据降维，还优化了模型输入，使得模型能够更高效地捕捉兴山民歌音频中的关键模式与特征，进而提升了分类性能。

本实验采用 5 秒时长的音频片段进行分析，每个音频片段被划分为 250 个等长时间帧，帧长 20 毫秒。针对每个时间帧，提取时域、频域倒谱频域这三方面特征，共获得 17 维数据。利用 RF 方法进行降维处理后，生成一个尺寸为 250 行 12 列的矩阵，该矩阵将作为后续模型的输入。特征重要性得分如图 4 所示。

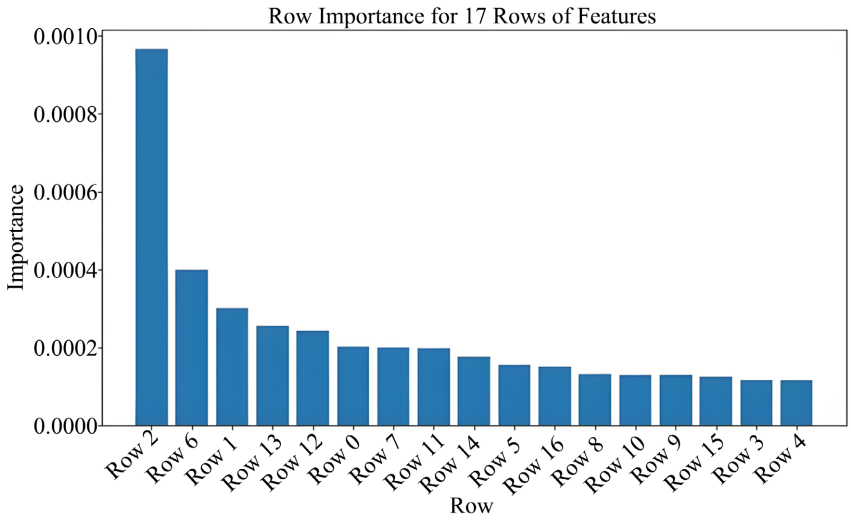

Figure 4. Importance scores of the 17 features obtained via RF processing
图 4. 通过 RF 处理得到的 17 个特征的重要性得分

Row0-Row12 是 MFCC 的子特征，Row13 是短时平均振幅，Row14 是短时能量，Row15 是频谱质心，Row16 为频谱带宽。根据图中显示的结果，Row8 以后的特征重要性得分低于 0.0001，故取前 12 个特征，进行后续的分类，这 12 个特征中，有 75%的特征属于梅尔谱频谱特征，这是因为梅尔谱频特征共 13 维，占所有特征的 76.4%，能够提供丰富的频域信息与时域变化，且涵盖声音特性信息。

2.3. CNN-CBAM-BiLSTM 模型

CBAM 是一种用于增强卷积神经网络(CNN)性能的注意力机制模块[11]，如图 5 所示。CBAM 旨在提升卷积神经网络(CNN)的性能，具体通过引入通道注意力与空间注意力机制，且不增加网络复杂度来实现。该模块包含通道注意力模块(C-channel)与空间注意力模块(S-channel)两大核心组件。这两个模块能够分别嵌入 CNN 的不同层级，以此强化特征表达能力，助力模型更好地捕捉关键信息。

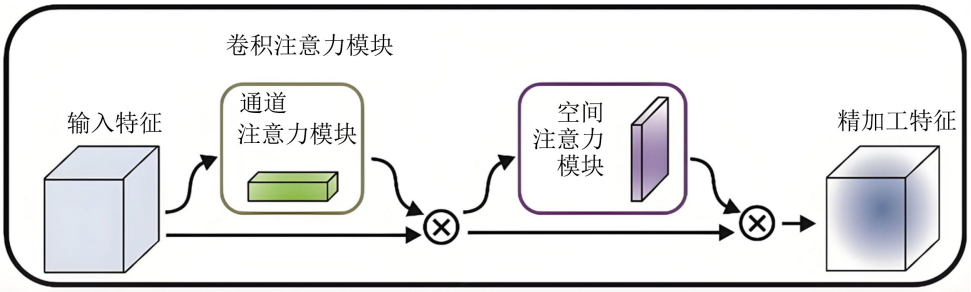

Figure 5. CBAM hybrid attention module
图 5. CBAM 混合注意力模块

在兴山民歌的识别领域，存在多重技术瓶颈。一方面，兴山民歌音频样本普遍存在模糊、嘈杂的特性，这使得传统模型在识别过程中遭遇严重阻碍，有效特征易被噪声干扰和掩盖，导致识别准确率较低。另一方面，在构建模型的过程中，未能充分关注音频数据在时间维度上的前后关联性，无法满足对兴山民歌深入研究与保护的需求。尤为关键的是，截至目前，尚未有针对兴山民歌在信号处理方面的专门研究，这无疑加剧了兴山民歌音频处理工作的难度。

鉴于此，本研究创新性地提出了一种融合注意力机制的卷积和双向长短期记忆网络。兴山民歌音频在时间序列上呈现出复杂且紧密的依赖关系，其丰富的音乐内涵与情感表达蕴含于连续的音频片段之中。为充分挖掘这些信息，模型将 CBAM 模块融入 CNN 架构。CBAM 的通道注意力机制能够依据音频特征的重要程度，自适应地为不同通道分配权重，使模型重点关注对分类起关键作用的特征通道；空间注意力机制则聚焦于关键空间区域，强化对重要特征的感知。这种融合显著提升了模型对模糊、嘈杂音频中有效特征的关注度与表达能力。

同时，考虑到兴山民歌音频在时间维度上双向信息的重要性，模型引入 BiLSTM。BiLSTM 能够同时处理音频序列的过去与未来信息，充分捕捉兴山民歌音频在时间维度上的双向上下文关联，精准提取其中蕴含的复杂特征。

CNN-CBAM-BiLSTM 网络的具体流程是，首先将降维后特征矩阵经历初始的卷积和池化操作，这一步相当于对数据进行降采样。卷积操作通过特定的卷积核在特征矩阵上滑动，提取局部特征；池化操作则对卷积后的特征图进行下采样，减少数据量的同时保留重要的特征信息，为后续的处理减轻计算负担。接下来，特征矩阵会通过 4 个基于 CNN 的残差模块进行堆叠处理。残差模块的加入，成功应对了深度神经网络训练时面临的梯度消失难题。借助这一模块，模型得以挖掘更复杂、更具表征力的特征，显著提升了对数据的学习能力。并且，在这些残差模块中嵌入了注意力模块(CBAM)。通道注意力模块会计算每

个通道的重要性权重，对特征图的通道维度进行加权，让模型关注到更关键的特征通道；经过 CNN 和注意力机制处理后的特征矩阵，被输入到 BiLSTM 层进行序列信息的处理。由于民歌音频具有一定的时序特性，BiLSTM 能够同时考虑序列的过去和未来信息，对特征进行更深入的分析和挖掘，捕捉到民歌在时间维度上的变化规律。最后，将 BiLSTM 输出的特征通过全连接层进行映射。全连接层将高维的特征向量映射到一个二维的输出空间，对应山歌和其他类型民歌这两个类别。通过对输出结果进行判断，我们就可以得到每一个民歌样本属于山歌或者其他类型民歌的分类预测结果。该模型能够充分利用民歌特征矩阵中的信息，实现对兴山民歌和其他类型民歌的有效区分。

综上所述，本研究提出的 RF-CNN-CBAM-BiLSTM 模型，紧密结合兴山民歌音频特点，有效弥补了传统模型以及现有深度学习算法的不足，且作为首次针对兴山民歌信号处理的探索，给出了对兴山民歌音频的特征提取的有效方法，在兴山民歌音频分类任务中具有兼具高准确性和高效率的性能，为兴山民歌的研究与保护工作提供了有力的技术支持。

3. 实验结果及对比分析

本节将呈现基于 RF-CNN-CBAM-BiLSTM 的音频分类模型在兴山民歌识别实验中的结果。为公正且全面地评测该模型处理音频分类问题的效能，本文按照 8:2 把数据集划分为训练集与测试集。其中，训练集包含 1440 首歌曲，测试集包含 360 首。

3.1. 实验结果

在兴山民歌识别研究中，本研究构建了 RF-CNN-CBAM-BiLSTM 模型，以实现兴山民歌与其他类型民歌的二分类任务。模型训练后，得到混淆矩阵，如图 6 所示。

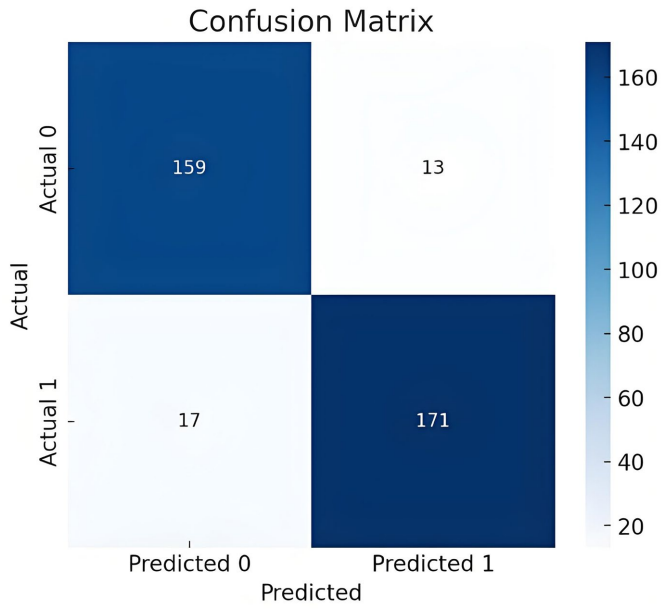

Figure 6. Confusion matrix of the RF-CNN-CBAM-BiLSTM classification model
图 6. RF-CNN-CBAM-BiLSTM 分类模型的混淆矩阵

图中 Actual 0 表示实际类别为兴山民歌的样本，Actual 1 表示实际类别为其他类型民歌的样本。Predicted 0 表示预测类别为兴山民歌的样本，Predicted 1 表示预测类别为其他类型民歌的样本。整体准确率约为 91.67%，这表明模型在整体上具备良好的识别能力，能够对大部分音频样本进行正确分类。

从表 2 中的指标来看，当以兴山民歌为正类时，精确率为 92.44%，意味着模型预测为兴山民歌的样本中，实际确为兴山民歌的比例较高；召回率达 90.34%，说明模型能有效捕捉到大部分实际的兴山民歌样本，对兴山民歌的覆盖较为全面。F1 值为 91.38%，综合体现了精确率与召回率的平衡，显示模型在兴山民歌的判别上性能较为良好。特异度为 92.93%，表明模型对其他类型民歌的正确识别能力也处于一定水平，有助于区分兴山民歌与其他类型民歌。

Table 2. Precision, recall, F1-Score, and specificity of Xingshan folk song recognition
表 2. 兴山民歌识别的精确率，召回率，F1 值，特异度

类别	精确率	召回率	F1 值	特异度
兴山民歌	92.44%	90.34%	91.38%	-
其他类型民歌	90.96%	92.93%	91.94%	92.93%

总体而言，该模型在兴山民歌与其他类型民歌的判别中表现出了一定的优势，能够在不同类别上保持相对较高的精确率、召回率和 F1 值，且特异度也较为可观，为兴山民歌的判别研究提供了较为可靠的依据。但模型仍存在一定的误判情况，后续可通过进一步优化模型、调整参数或扩充数据集等方式，不断提升其判别性能。

RF-CNN-BiLSTM-CBAM 分类模型的损失函数曲线如图 7 所示。基于交叉熵损失函数绘制的损失曲线呈现出如下特征。训练损失曲线(蓝色)在训练第一轮处于较高位置，随后迅速下降，于约 50 轮训练后趋于平稳，最终稳定在曲线的最低值，表明模型对训练数据的特征学习效果显著，随着训练推进，模型预测与真实标签间差距不断缩小，在训练集上的性能持续优化。测试损失曲线(橙色)在开始阶段同样较高且波动较大，尽管下降速度相对较慢，但也体现出模型在测试集上随训练的进行性能也有所提升，能够对新数据进行一定程度的有效识别。两条曲线的变化趋势共同显示出该模型在训练过程中不断优化，对训练数据和新数据均具备一定的处理能力，在山歌与其他类型民歌的二分类任务中展现出了良好的识别性能与应用潜力。

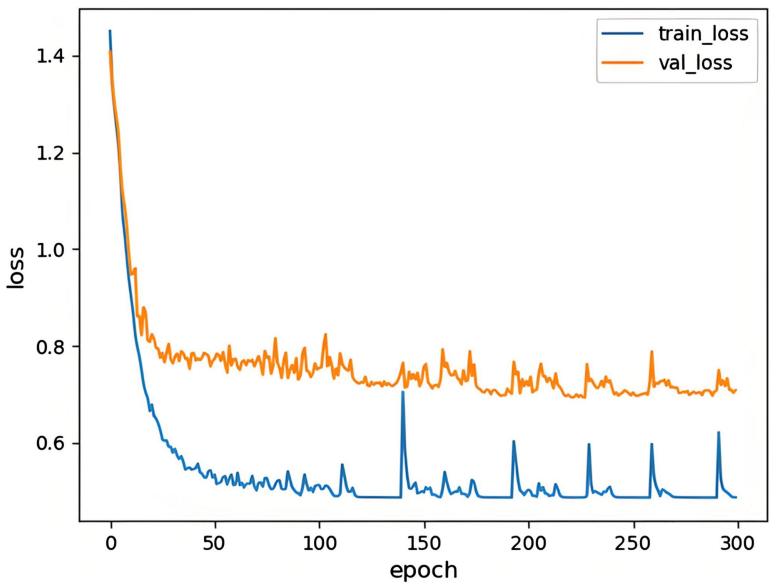

Figure 7. Loss function curve of the RF-CNN-BiLSTM-CBAM classification model
图 7. RF-CNN-BiLSTM-CBAM 分类模型的损失函数曲线

RF-CNN-BiLSTM-CBAM 分类模型的准确率曲线如图 8 所示。该曲线识别性能较好，鲁棒性较高。训练准确率(蓝色曲线)与测试准确率(橙色曲线)在训练初期虽有波动，但随着训练轮数增加，曲线均呈显著上升趋势。训练准确率最终接近 0.98，且趋于平稳，表明模型对训练数据的学习效果较好，能够准确地对训练集中的山歌与其他类型民歌进行分类。测试准确率也稳定在 0.91 左右，这显示模型在处理测试集数据时，也具备较强的分类能力，能够较好地泛化到新数据上。同时，在训练前期，两条曲线上升速度较快，说明模型收敛迅速，能够在较短的训练轮数内有效学习到数据特征，展现出较高的学习效率。从整体来看，该模型在山歌与其他类型民歌的二分类任务中表现出良好的性能，具备实际应用和进一步优化的价值。

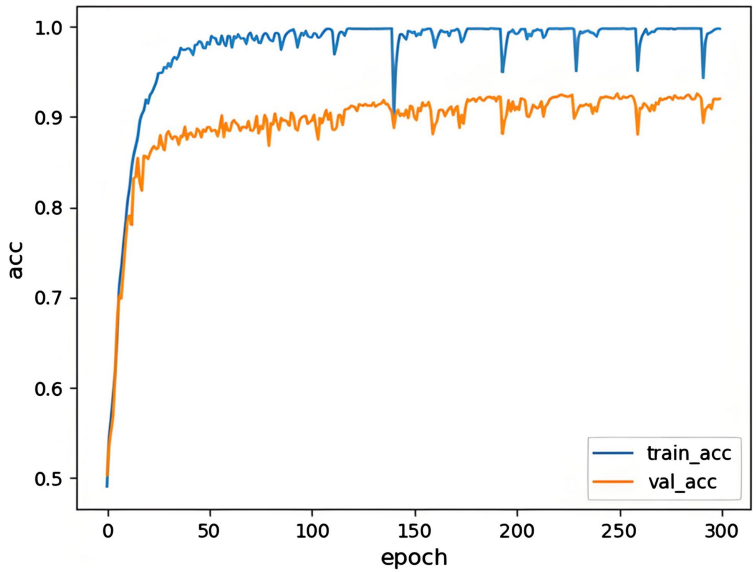

Figure 8. Accuracy curve of the RF-CNN-BiLSTM-CBAM classification model
图 8. RF-CNN-BiLSTM-CBAM 分类模型的准确率曲线

3.2. 对比分析

将本文提出的模型 RF-CNN-CBAM-BiLSTM 分类模型与 CNN-CBAM-BiLSTM、ResNet50、EfficientNetV1、BiLSTM、LSTM 五个神经网络模型进行对比实验，给出所有模型的分类准确率和每一轮运行时间。在兴山民歌识别任务中，本文提出的 RF-CNN-CBAM-BiLSTM 模型在分类性能与计算效率上均展现出显著优势。如图 9 所示，该模型以 91.67%的准确率超越所有对比模型，较未利用 RF 降维的 CNN-CBAM-BiLSTM 模型的准确率 89.97%，提升 1.7%，较 ResNet50 的准确率 87.36%，提升 4.31%。这一性能提升得益于特征降维与模型架构的协同优化：通过随机森林(RF)对原始 17 * 250 维特征进行筛选，保留 12 * 250 维高区分性特征，剔除冗余信息，使模型更聚焦于关键声学模式；同时，CNN-CBAM 模块通过通道与空间注意力机制强化局部特征提取能力，BiLSTM 捕捉音频序列的长期依赖关系，二者结合有效弥补降维带来的信息损失，确保模型在低维特征空间中仍能精准建模兴山民歌的独特性。

在运行效率方面，本文模型也呈现优势。如图 10，分类模型的每一轮运行时间对比。本文模型每一轮运行时间仅需 22 秒，较未降维的 CNN-CBAM-BiLSTM 模型的每一轮运行时间 27 秒，缩短 18.5%，且显著优于参数量庞大的 ResNet50 模型的每一轮运行时间 257 秒与 EfficientNetV1 模型的每一轮运行时间 51 秒。效率提升的核心在于特征降维对计算复杂度的优化：输入特征维度从 17 * 250 压缩至 12 * 250 (减少 29.4%)。尽管 BiLSTM 模型的每一轮运行时间(18 秒)与 LSTM 模型的每一轮运行时间(15 秒)耗时更

短，但其准确率较本文模型存在显著差距 5.34%~6.47%，表明单一模块模型难以兼顾效率与精度，而本文模型通过多模块融合实现了二者均衡。

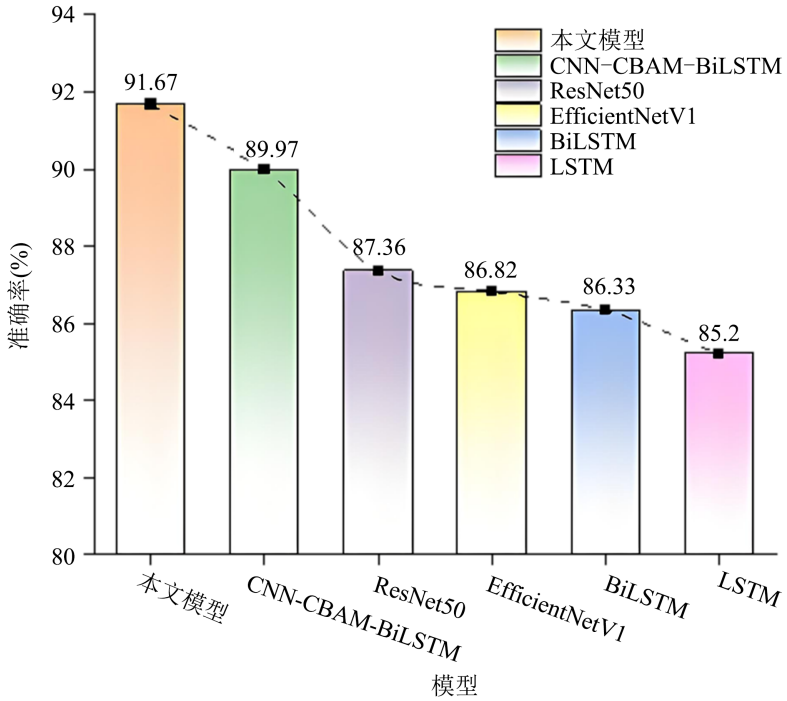

Figure 9. Comparison of accuracy of classification models
图 9. 分类模型的准确率对比

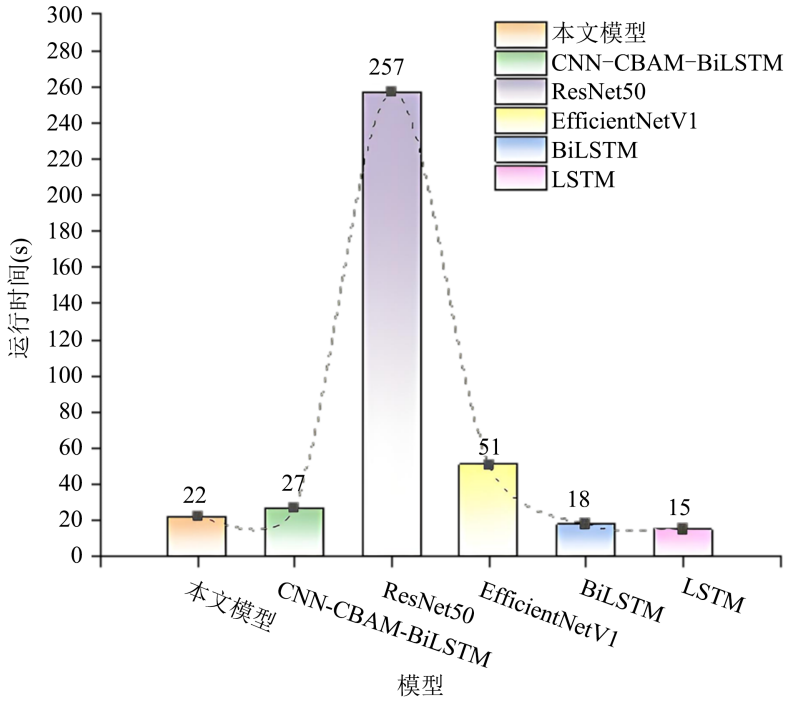

Figure 10. Comparison of running time of classification models
图 10. 分类模型的运行时间对比

RF-CNN-CBAM-BiLSTM 模型无论是在硬件存储方面(占用更少的存储空间来存储特征数据和模型参数)，还是在计算资源消耗上，都具有明显的优势。这使得该模型在资源有限的环境中，依然能够稳定运行，展现出良好的适应性。对于研究者而言，RF-CNN-BiLSTM-CBAM 模型在部署和应用过程中更加便捷，能够以较低的成本实现音频分类功能，具有广阔的应用前景和较高的工程实用价值，为兴山民歌分类以及其他音频分类相关的工程实践提供了更具性价比的解决方案。

从应用领域看，可将其推广至更为复杂的民歌分类体系，实现对众多民歌子类的精准识别，还能迁移至音乐流派分类、语音内容识别等多元音频分类场景。该模型在民歌传承与保护领域具有潜在价值，通过对海量民歌的高效分类，能够助力构建完备的民歌数据库，为音乐文化的传承、研究及创新提供坚实的数据支撑。

为进一步提升 RF-CNN-CBAM-BiLSTM 模型的性能，可从特征、模型结构、数据及学习策略等多方面着手。在特征选择上，尽管 RF 降维已优化了特征维度，但仍有提升空间。后续可探索多算法融合的特征选择方法，通过优势互补实现更精准的特征筛选；同时，精细调整 RF 算法内部参数，得以平衡特征维度与信息保留量，提升分类准确率。在模型结构方面，深入剖析 RF-CNN-CBAM-BiLSTM 架构中各组件的作用机制，通过调整卷积核数量、循环单元类型及 CBAM 模块的权重分配等方式，优化模型对民歌特征的提取与学习过程。此外，扩充训练数据集的规模与多样性，纳入更多具有地域特色、演唱风格差异的民歌样本，提升模型的泛化能力和特征判别能力。最后，运用集成学习策略，融合多个性能各异的模型，利用模型间的差异互补，有望显著提升模型的分类稳定性与准确性。

参考文献

[1] Elbir, A. and Aydin, N. (2020) Music Genre Classification and Music Recommendation by Using Deep Learning. *Electronics Letters*, **56**, 627-629. https://doi.org/10.1049/el.2019.4202

[2] Fu, Z.Y., Lu, G.J., Ting, K.M. and Zhang, D.S. (2011) A Survey of Audio-Based Music Classification and Annotation. *IEEE Transactions on Multimedia*, **13**, 303-319. https://doi.org/10.1109/TMM.2010.2098858

[3] Zaman, K., Sah, M., Direkoglu, C. and Unoki, M. (2023) A Survey of Audio Classification Using Deep Learning. *IEEE Access*, **11**, 106620-106649. https://doi.org/10.1109/access.2023.3318015

[4] Zahid, S., Hussain, F., Rashid, M., Yousaf, M.H. and Habib, H.A. (2015) Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods. *Mathematical Problems in Engineering*, **2015**, Article ID: 209814. https://doi.org/10.1155/2015/209814

[5] Breebaart, J. and Mckinney, M.F. (2004) Features for Audio Classification. In: Verhaegh, W.F.J., Aarts, E. and Korst, J., Eds., *Algorithms in Ambient Intelligence*, Springer, 113-129. https://doi.org/10.1007/978-94-017-0703-9_6

[6] Cances, L., Labbé, E. and Pellegrini, T. (2022) Comparison of Semi-Supervised Deep Learning Algorithms for Audio Classification. *EURASIP Journal on Audio*, *Speech*, *and Music Processing*, **2022**, Article No. 23. https://doi.org/10.1186/s13636-022-00255-6

[7] Nanni, L., Costa, Y.M.G., Aguiar, R.L., Mangolin, R.B., Brahnam, S. and Silla, C.N. (2020) Ensemble of Convolutional Neural Networks to Improve Animal Audio Classification. *EURASIP Journal on Audio*, *Speech*, *and Music Processing*, **2020**, Article No. 8. https://doi.org/10.1186/s13636-020-00175-3

[8] Nam, J., Choi, K., Lee, J., Chou, S. and Yang, Y. (2019) Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach. *IEEE Signal Processing Magazine*, **36**, 41-51. https://doi.org/10.1109/msp.2018.2874383

[9] Matityaho, B. and Furst, M. (1994) Classification of Music Type by a Multilayer Neural Network. *The Journal of the Acoustical Society of America*, **95**, 2959-2959. https://doi.org/10.1121/1.409056

[10] Nanni, L., Costa, Y.M.G., Lucio, D.R., Silla, C.N. and Brahnam, S. (2017) Combining Visual and Acoustic Features for Audio Classification Tasks. *Pattern Recognition Letters*, **88**, 49-56. https://doi.org/10.1016/j.patrec.2017.01.013

[11] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1