从时域到频域：基于多分支 CNN 网络的 AI 音频检测模型

NKUMMF2025138

October 6, 2025

Abstract

针对 AI 音频的识别问题，本文提出一种基于**多分支卷积神经网络（Multi-Branch CNN）**的端到端 AI 音频检测与评分模型。

从**时域、频域及声学常见统计量**等多角度提取 11 类特征，分别经五个并行分支建模后融合判别。针对单分支贡献的量化问题，引入**分支探针**机制，基于各分支独立判别准确率确定加权系数，构建可解释的 AI 痕迹综合评分体系。

本文设计多种扰动与对抗性处理（如频谱均衡、高频注入、环境噪声混入等）评估模型鲁棒性，并结合分支贡献分析揭示其在时域包络与共振峰布局上的依赖性。实验结果表明，该模型在验证集上准确率可达 89%–90%，在多数轻中度扰动下保持稳定性能，综合评分在强扰动下亦具较高稳健性。本文方法具有较低计算开销与良好可扩展性，可推广至语音伪造检测、环境音识别等领域。

关键词： AI 音乐检测；多分支 CNN；音频特征提取；探针机制；AI 痕迹评分

判别模型的建立与求解

模型建立与求解 i

1. 模型假设
2. 数据处理
3. 特征说明
4. 模型建立
5. 结果分析

模型假设 i

1. 人工标注的“真人/AI”标签总体正确率较高；
2. 在选定的分析窗口（20s）内，音频的统计特性近似平稳，从而帧级特征的统计聚合是有效的。

数据处理 i

1. 代码中首先将输入文件转码为 WAV 格式
2. 采用音频处理领域常用的 22050 Hz 采样率对输入音频进行统一解码与重采样
3. 在特征提取流程中，原始歌曲首先被切分为 20 秒长度的片段

在此基础上，本文提取了表 1 中所列的多种时域、频域及声学特征，涵盖能量变化、频谱形状、谐波结构及共振峰等方面的信息。

特征说明 i

特征说明 ii

特征名称	维度	特征类型
rms	(1, 862)	时域能量
zcr	(1, 862)	时域变化
hjorth	(3, 1)	全局统计
log_mel	(128, 862)	频域谱图
mfcc	(13, 862)	声学特征
centroid	(1, 862)	频谱形状
contrast	(7, 862)	频谱对比度
flatness	(1, 862)	频谱形状
f0	(1440,)	音高曲线
hnr	(1998,)	声学质量
formant	(3, 2000)	共振峰

Table 1.各特征的名称、维度及类型

模型建立 i

本研究构建了一个多分支卷积神经网络（Multi-Branch CNN），用于融合处理多种类型的音频特征，从而实现对 AI 生成音频的准确识别。整体网络结构如下图所示，模型共由五个输入分支构成，分别对应不同特征类型，并通过特征融合实现最终的二分类判别。

模型建立 ii

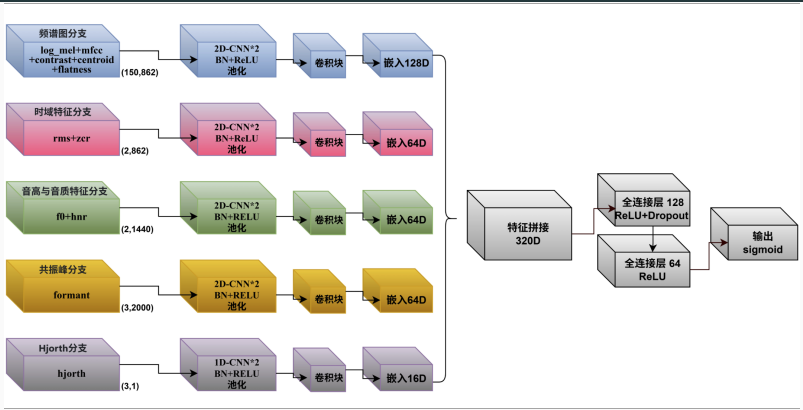


Figure 1.多分支 CNN 示意图

模型建立 i

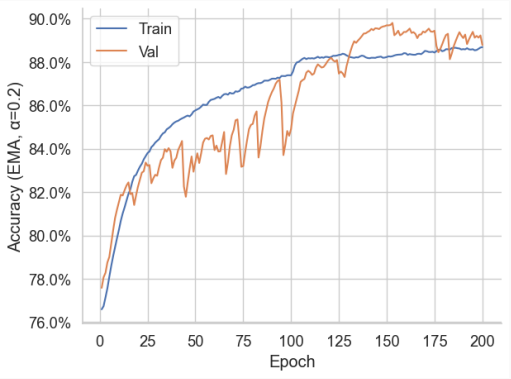
分支名称	输入特征	处理与输出
频谱图分支（Spectrogram）	log_mel, mfcc, contrast, centroid, flatness	2D CNN 提取时频模式 → GAP → 128 维表示
时域分支（Time-Domain）	rms, zcr	1D CNN 提取时间结构 → 64 维向量
音高与音质分支（Pitch & Quality）	f0, hnr	1D CNN 建模调型与音质 → 64 维向量
共振峰分支（Formant）	formant	1D CNN 处理时间序列 → 64 维向量
Hjorth 参数分支（Hjorth Param.）	hjorth (Activity, Mobility, Complexity)	两层全连接 → 16 维表示

Table 2.多分支特征处理结构概览

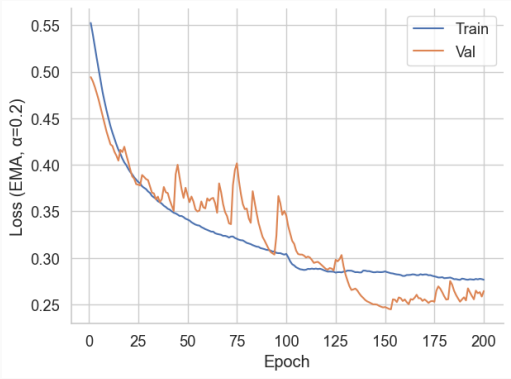
模型建立 ii

五个分支提取的特征向量分别为：128 维（频谱图）、64 维（时域）、64 维（音高音质）、64 维（formant）与 16 维（hjorth），合并后得到一个 336 维的全局音频表征。该向量随后输入至全连接分类器，包含一层隐藏层（64 单元），使用 ReLU 激活与 Dropout 防止过拟合，最终输出一个概率值，判定音频是否为 AI 生成。

结果分析



(a) 准确率（Train vs Val）



(b) 损失（Train vs Val）

Figure 2.训练过程的准确率与损失对比

结果分析 i

训练损失呈单调递减趋势，最终趋近于 0.27；

训练集准确率从 76.5% 持续上升并最终收敛于 88.5% 左右，

验证集准确率和损失虽经历多次波动，但最终趋于稳定，分别在 89%-90% 和 0.25-0.26 之间。

上述结果表明模型训练稳定，未出现过拟合迹象。这说明多分支特征融合结构在提取不同维度音频信息时具有较好的**泛化能力**，能够有效捕捉 AI 生成音乐与人类创作音乐的差异特征。

消融实验 i

为探究各输入分支在多分支 CNN 模型中的贡献度，本研究设计了消融实验，通过控制不同分支的启用/关闭状态，评估其对模型分类精度的影响。具体而言，我们分别在以下三种设置下进行测试：

- 全分支开启；
- 单独关闭某一分支，其余保持开启；
- 单独开启某一分支，其余全部关闭。

消融实验 i

Table 3.消融实验结果（准确率%）

频谱图分支	时域分支	音高与音质分支	共振峰分支	Hjorth 参数分支	准确率
1	1	1	1	1	89.30%
0	1	1	1	1	29.87%
1	0	1	1	1	79.20%
1	1	0	1	1	88.77%
1	1	1	0	1	88.40%
1	1	1	1	0	89.10%
1	0	0	0	0	77.87%
0	1	0	0	0	30.02%
0	0	1	0	0	47.76%
0	0	0	1	0	50.67%
0	0	0	0	1	25.20%

消融实验 i

虽然消融实验能够衡量各分支在整体判别任务中的贡献度，但这种方法仍存在一定局限性：

- **无法独立量化单分支判别能力**：在消融实验中，关闭某一分支时其特征被替换为零向量，其余分支仍参与决策，因此所得精度下降值只是间接反映该分支的重要性，而非其独立完成任务的真实表现。
- **分支间存在互补与冗余效应**：不同分支特征可能存在高度相关性，关闭一个分支后，其信息可能部分被其他分支补偿，导致重要性被低估；反之，如果多个分支特征冗余，关闭其中一个分支对结果影响可能被夸大。

消融实验 ii

- **评估结果依赖当前模型权重分布**：消融实验是在固定训练好的多分支模型上进行的，权重分布已经针对多特征融合进行了优化，单分支输入下的表现可能不能代表该分支在独立训练时的潜力。
- **未能刻画特征在不同类型样本上的贡献差异**：消融实验只给出整体平均精度，而没有分析分支在不同类别、不同难度样本上的表现差异，限制了对特征适用性的理解。

针对以上不足，我们在后续模型中引入了分支探针机制，使每个分支能够在保持原有多分支协作的同时，单独完成二分类任务，从而获得更客观、细粒度的特征贡献评估结果。

评分机制

评分公式的定义

受消融实验结果的启发，为了细化模型的评价指标，我们改进了多分支 CNN 网络，为每个分支添加了一个由线性层和 Sigmoid 函数构成的**探针头结构**。

对于一段音频，每个分支的探针都可以独立判断该音频是否为 AI 生成，并输出**分类概率**。

评分规则

记各分支探针头给出的“音频为 AI 生成”的概率为 $p_i \in [0,1]$，则作品的综合评分（AI 痕迹强度）定义为

$$S \;=\; \sum_{i} \omega_i \, p_i \;\in\; [0,1], \tag{1}$$

其中 ω_i 为分支的评分权重。S 越大表示该音频越可能为 AI 生成。

分支权重的确定

在综合各分支结果为音乐作品打分时，我们以各探针头单独决策时的**准确率**为依据，确定了各分支的评分权重。

设各分支探针在独立判别（AI/非 AI）下的准确率 $acc_i, i \in \{\text{spec,time,prosody,formant,hjorth}\}$。

为避免随机猜测带来的偏置，先做机会校正（chance-corrected）并归一化：

$$\omega_i = \frac{acc_i - 0.5}{\sum_j (acc_j - 0.5)}, \qquad \sum_i \omega_i = 1. \tag{2}$$

Table 4.分支探针头独立预测准确率及由此推导出的归一化权重

分支（Branch）	准确率 a_i	权重 w_i
spec（频谱图）	0.7725	0.1987
time（时域）	0.7834	0.2067
prosody（音高音质）	0.7718	0.1982
formant（共振峰）	0.7717	0.1981
hjorth（Hjorth）	0.7718	0.1982
主分类器（参考）	0.8924	–

探针准确率与权重

从独立判别能力看，各分支探针准确率集中在 $77\% \sim 78\%$ 区间，说明不同模态特征（时域、频域、声学统计与共振峰）在本任务上均具较强的区分信息，且不存在明显的“短板”分支。各分支探针的准确率相比于消融实验结果（表 3) 多数分支有较高提升，进一步证实了加入探针头对模型进行微调的必要性。

综合评分采用式1的线性加权后，可在推理阶段以显式、可解释的方式衡量各分支对最终结论的贡献。

主分类器的总体准确率为 89.24%，显著高于任一单分支探针，验证了多特征融合的有效性；而探针权重的近似均衡分布则为后续的可解释分析提供了稳定基线。

谢谢大家！