从时域到频域：基于多分支 CNN 网络的 AI 音频检测模型

NKUMMF2025138

October 5, 2025

摘要 I

Abstract

随着深度学习与生成模型的发展，AI 音乐创作在产业与艺术领域快速普及，但其版权归属、作品质量及可识别性等问题亟待解决。

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、频域及声学统计等多角度提取 11 类特征，分别经五个并行分支建模后融合判别。针对单分支贡献难以量化的问题，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。

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。

摘要 II

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

问题重述 I

近年来，随着深度学习与生成模型的快速发展，人工智能在音乐创作领域的应用呈现爆发式增长。从 2012 年虚拟歌手“洛天依”问世，到 2024 年由 Suno AI 技术生成的作品震撼乐坛，AI 生成音乐已深度融入现代音乐产业与艺术创作当中。

与此同时，AI 音乐创作也伴生出一系列亟待解决的现实问题：一方面，作品的版权归属常常难以厘清；另一方面，由于生成模型训练策略与数据源的差异，AI 产出的音乐在质量上良莠不齐，部分低质量内容甚至影响用户体验。各大流媒体与社交平台已开始在疑似 AI 创作作品下加注标识，反映出对高效、准确的 AI 音乐鉴别与评价技术的迫切需求。

为此，本研究旨在构建一套从端到端检测到综合评分的完整数学模型，不仅要识别 AI 生成的音乐样本，更要对其质量与“AI 痕迹”程度进行量化评价，并在此基础上探索模型的鲁棒性与改进方向。

问题一：设计一个端到端的检测模型，从多种音频特征中提取有效信息，判断其生成来源，做到能够准确识别出音频是否由 AI 生成。同时，提供完整的代码实现，并打包成可一键运行的程序，以便便捷进行音频检测。

问题二：设计一套高效的 AI 音乐评分指标，用以评估 AI 生成音乐的质量。该评分体系要综合考虑音乐的结构性、节奏性、音色丰富度等艺术因素。通过多维度的评分标准，能够客观量化 AI 生成音乐的创作水平，为后续的 AI 音乐优化提供反馈。

问题重述 II

问题三：探索模型在面对扰动或对抗样本时的表现。通过在 AI 音频中插入人类创作的音频片段，分析模型在识别中出现的薄弱环节并针对这些问题提出可能的优化方案，提高模型在实际应用中的准确性和鲁棒性。

问题分析 I

在音频信号处理中，往往需要对时域、频域及声学等多方面特征进行分析，此外结合本题提供的数据规模，本文考虑使用机器学习方法；结合现有数据量和跨平台通用性这两方面的要求，考虑使用结构相对轻量的卷积神经网络 (CNN) 模型来求解。在此基础上，为增强模型对不同类型特征信息的表达与融合能力，设计了一种多分支 CNN 模型，将不同种类的音频特征划分为若干类分别建模，并在特征拼接阶段进行统一判别，从而提升模型在跨域音频识别任务中的泛化能力与鲁棒性。

针对 AI 生成音乐质量难以量化评价的问题，本文在音频检测特征的基础上，设计了一套多维度综合评分体系，用于对 AI 生成音乐进行客观评分。考虑到音乐的结构性、节奏稳定性与音色丰富度等艺术属性，结合声学分析方法提取谐噪比（HNR）、基频曲线（F0）、频谱质心、谱平坦度、频谱对比度及 MFCC 等特征，并对各维度特征进行归一化与加权处理，以减少单一特征对总分的影响。

最终通过分层加权模型将多维特征映射为统一的量化分数，实现不同生成模型、不同音乐风格间的横向比较，并为后续生成模型的改进提供数据参考。　分析模型特征后，我们发现频域信息对于判别曲目是否为 AI 生成贡献很大，于是我们考虑使用多音轨融合、丰富和声的方式改进目标音频，从而在保证艺术性和可欣赏性的同时优化 AI 生成的音乐作品。另一方面，我们的模型过于依赖频域信息，在提升鲁棒性方面，我们可以考虑调整多分支的权重比例、域对齐等方式来减轻模型对单一类别信息的依赖性。

模型假设 I

1. 人工标注的“真人/AI”标签总体正确率较高；在统一预处理后，特征对微小扰动不敏感，少量噪声标签对训练影响可忽略。
2. 假设训练集中所用的真人音频和 AI 生成音频样本，能够满足未来待检样本的多样性（如不同风格、不同生成引擎）。
3. 假设选取的声学特征能够充分表征 AI 生成音乐与人类创作音乐在本质上的差异。
4. 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。
5. 假设对于一维时域特征分支，相邻帧之间的依赖关系使用局部卷积核足以捕捉，而不需要额外地建模长距离的时序依赖。

符号说明 I

Table 1.符号说明表

符号	说明
N	帧长
$x[n]$	每一帧的时域信号
$\|X(f)\|$	每帧信号经短时傅里叶变换得到的幅度谱
$E_{\text{top},b}$	取频带 b 中上层（如 90%）能量值
s_h	谐波分量
s_n	噪声分量
x'	信号的一阶时间导数
x''	信号的二阶时间导数
S	多分支探针加权得到的综合评分
$prob\_main$	主分类头输出的分类概率

模型建立与求解 I

本文从音频信息的时域、频域、全局统计量等角度，从原始音频文件中提取了十一个特征，按类别分组后输入给多分支 CNN 网络，构建了融合多域信息的 CNN 分类器，用于判别音频是否为 AI 生成的音乐作品。

本文在特征提取阶段，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。该采样率在音乐信息检索（MIR）和语音识别等任务中广泛使用，其频率响应上限为 11025 Hz（奈奎斯特频率），能够保留音乐与人声中的主要能量成分，同时显著降低计算开销。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，确保在批量处理时每个样本具有相同的时间维度，有利于模型的稳定训练。

为了兼容多种音频格式（如 AAC、M4A、MP3 等），代码中首先将输入文件转码为 WAV 格式，然后调用 `librosa.load` 函数读取，并在加载过程中统一采样率（`sr=22050`）。短时傅里叶变换（STFT）采用窗口长度 `n_fft=2048`，帧移步长 `hop_length=512`，可在时间分辨率与频率分辨率之间取得平衡。梅尔频谱计算中使用 `n_mels=128` 个梅尔滤波器，以近似人耳在对数频率域的感知特性。

在此基础上，本文提取了表 2 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。

模型建立与求解 II

特征名称	维度	特征类型	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 2.各特征的名称、维度及类型

频谱图分支（Spectrogram Branch）输入由五类频域特征沿通道维度拼接形成的多通道频谱图：

1. `log_mel` 对数梅尔频谱：

模型建立与求解 III

$$E_m = \sum_f |X(f)|^2 \, H_m(f), \quad \log\text{-Mel}_m = \log\big(E_m + \epsilon\big), \quad m = 1, \ldots, 128. \tag{1}$$

模拟人耳对低频更敏感的特性，提供紧凑且具感知意义的频谱表示。
`mfcc` 梅尔频率倒谱系数：

$$\mathrm{MFCC}_k = \sum_{m=1}^{M} \log(E_m) \cos\Big[\frac{\pi k}{M}\big(m - \tfrac{1}{2}\big)\Big], \quad k = 0, \ldots, 12. \tag{2}$$

将对数 Mel 能量转到倒谱域，去除声道效应，更聚焦于音色和共振峰结构，常用于语音和音色分类。
`contrast` 频谱对比度：

$$\mathrm{Contrast}_b = E_{\mathrm{top},b} - E_{\mathrm{bot},b}, \quad b = 1, \ldots, 7. \tag{3}$$

反映频谱峰谷差异，突出谐波与噪声成分。
`centroid` 频谱质心：

$$\mathrm{Centroid} = \frac{\sum_f f \cdot |X(f)|}{\sum_f |X(f)|}. \tag{4}$$

模型建立与求解 IV

衡量频谱重心位置，高质心表示高频成分占优，用于区分亮音与暗音。
`flatness` 谱平坦度：

$$\text{Flatness} = \frac{\left(\prod_f |X(f)|\right)^{1/N}}{\frac{1}{N}\sum_f |X(f)|}. \tag{5}$$

度量信号是更像窄带谐波（平坦度小）还是像噪声（平坦度接近 1），常用于音乐/噪声分离。
时域分支（Time-domain Branch）输入为两类时域特征：

- `rms` 均方根能量：

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2}. \tag{6}$$

  度量信号的整体功率和响度水平，能够反映音量强弱的变化，用于区分轻音重音、检测瞬时能量突变。

模型建立与求解 V

- `zcr` 过零率：

$$\mathrm{ZCR} = \frac{1}{2N} \sum_{n=1}^{N-1} \left| \mathrm{sgn}(x[n]) - \mathrm{sgn}(x[n+1]) \right|. \tag{7}$$

描述波形频率粗略程度和有声／无声边界。

音高与音质分支（Pitch and Quality Branch）输入为两类声学特征：

- `f0` 基频：

$$f_0 = \frac{1}{\tau_{\max}}, \quad \tau_{\max} = \arg\max_{\tau} R(\tau), \tag{8}$$

追踪音高变化，是旋律、声调及说话人特征分析的核心

- `hnr` 谐噪比：

$$\mathrm{HNR} = 10 \log_{10} \frac{\mathbb{E}[s_h^2]}{\mathbb{E}[s_n^2]}. \tag{9}$$

衡量声音清晰度与噪声含量。

模型建立与求解 VI

共振峰分支（Formant Branch）输入为三通道共振峰序列：

- `formant` 共振峰：

$$\{F_1, F_2, F_3\} = \frac{\angle(\text{LPC poles})}{2\pi T_s}. \tag{10}$$

使用 LPC 法提取时间窗口内（本文设定为 10ms）音频的谐波包络，取出其前三个共振峰的频率点作为当前窗口的音频 formant 特征。该特征反映语音及共鸣腔腔的物理结构，是元音识别及说话人区分的关键特征。

Hjorth 参数分支（Hjorth Parameter Branch）输入为三个统计特征：

- `hjorth` Hjorth 参数：

$$\text{Activity} = \text{Var}[x], \quad \text{Mobility} = \sqrt{\frac{\text{Var}[x']}{\text{Var}[x]}}, \quad \text{Complexity} = \frac{\sqrt{\text{Var}[x'']/\ \text{Var}[x']}}{\text{Mobility}}. \tag{11}$$

分别衡量信号能量、变化速度和波形复杂度，用于快速估计信号的动态特征。

模型建立与求解 VII

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。

模型建立与求解 VIII

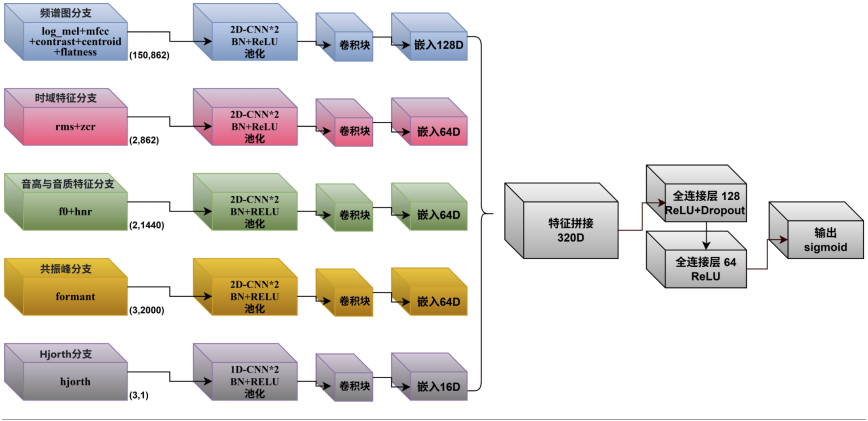


Figure 1.多分支 CNN 示意图

模型建立与求解 IX

- **频谱图分支（Spectrogram Branch）：**输入由 `log_mel` 频谱图（Log-Mel Spectrogram）、`mfcc`（Mel Frequency Cepstral Coefficients）、`contrast`（Spectral Contrast）、`centroid`（Spectral Centroid）以及 `Spectral Flatness` 五种频域特征沿通道维度拼接形成的多通道频谱图。该分支通过两层二维卷积神经网络（2D CNN）提取其在时间与频率双维度上的局部模式，随后经全局平均池化获得固定长度的 64 维特征向量，并通过全连接层映射为 128 维表示。
- **时域分支：**主要处理 `rms`（Root Mean Square Energy，均方根能量）和 `zcr`（Zero Crossing Rate，过零率）两类时域特征。输入形状为 $(2, 862)$，采用一维卷积（1D CNN）提取其时间结构，最终得到 64 维特征向量。
- **音高与音质分支：**接收 `f0`（Fundamental Frequency，基频）与 `hnr`（Harmonics-to-Noise Ratio，谐噪比）两种声学特征，拼接后输入为 $(2, 1440)$，通过 1D CNN 提取其调型与音质模式，映射为 64 维向量。
- **共振峰分支：**输入为 20s 音频片段在各个时间窗口上的 `formant` 特征构成的三维序列，通过 1D CNN 处理后压缩为 64 维特征。

模型建立与求解 X

- **Hjorth 参数分支（Hjorth Parameter Branch）：** `hjorth` 特征为 $(3, 1)$ 的全局统计量，包括 Activity（活动度，衡量信号能量）、Mobility（移动性，反映波形变化速度）与 Complexity（复杂度，描述波形不规则性），不含时间维度，因此直接使用两层全连接网络映射为 16 维表示。

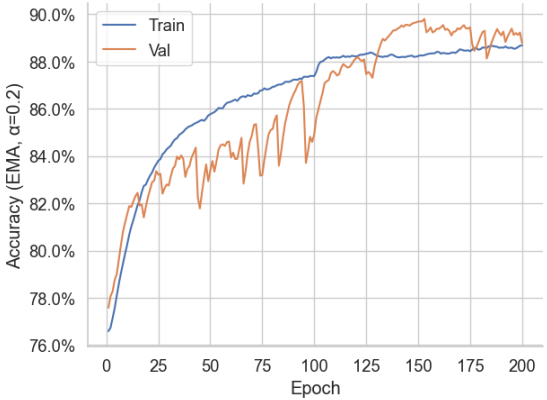
五个分支提取的特征向量分别为：128 维（频谱图）、64 维（时域）、64 维（音高音质）、64 维（formant）与 16 维（hjorth），合并后得到一个 336 维的全局音频表征。该向量随后输入至全连接分类器，包含一层隐藏层（64 单元），使用 ReLU 激活与 Dropout 防止过拟合，最终输出一个概率值，判定音频是否为 AI 生成。

数据可视化

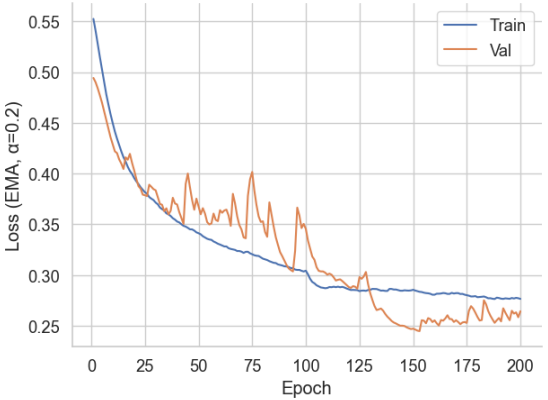
模型在 200 个 epoch 内的学习曲线如图 2 所示。随着训练代数的增加，训练准确率从 76.5% 持续上升并最终收敛于 88.5% 左右，训练损失呈单调递减趋势，最终趋近于 0.27；验证准确率和损失虽经历多次波动，但最终趋于稳定，分别在 89%-90% 和 0.25-0.26 之间。单一指标对比下，经过多次训练的模型在验证集上的指标要优于训练集。

上述结果表明模型训练稳定，未出现过拟合迹象（如验证损失反弹、验证精度持续下降）。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。

模型建立与求解 XI



(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

消融实验

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（Ablation Study），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：（1）全分支开启；（2）依次关闭某一分支，其余保持开启；（3）单独开启某一分支，其余全部关闭。五个分支分别对应频谱图分支（spec）、时域分支（time）、音高与音质分支（prosody）、共振峰分支（formant）以及 Hjorth 参数分支（hjorth）。

实验结果如表 3 所示，全分支开启时模型在测试集上取得了最高的准确率 89.30%，表明多特征融合能够显著提升检测性能。当关闭频谱图分支（spec）时，准确率骤降至 29.87%，说明该分支在整体识别中贡献最大；关闭时域分支（time）或音高与音质分支（prosody）时，准确率分别下降至 79.20% 和 88.77%，显示二者在细节刻画方面亦具有重要作用；关闭共振峰分支（formant）或 Hjorth 分支（hjorth）时，准确率分别为 88.40% 和 89.10%，影响相对较小，但依然对整体性能有一定增益。

在单分支测试中，仅开启频谱图分支即可达到 77.87% 的准确率，远高于其他分支单独使用的表现（时域分支 30.02%，音高与音质分支 47.76%，共振峰分支 50.67%，Hjorth 分支 25.20%）。这一结果进一步验证了频谱类特征在 AI 音频检测任务中的核心地位，同时也表明单一类型特征难以支撑全面准确的判断，融合多维度特征是提升性能的有效途径。

模型建立与求解 XIII

Table 3.消融实验结果（准确率%）

spec	time	prosody	formant	hjorth	accuracy
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%

注：1 表示该 CNN 分支传入真实的特征张量，0 表示传入对应形状的全 0 张量

模型建立与求解 XIV

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。
- **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。

模型建立与求解 XV

基于以上不足，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。
问题一中的多分支 CNN 网络将各类特征整合起来用于判断音频是否由 AI 生成，它可以很好地综合各类特征的信息，但无法单独依据某一特征进行判断，很难从多角度、多方面评价一首曲子的 AI 痕迹强度。
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，使得每个分支可以单独依据本分支的特征对音频进行评估。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。
权重确定 设各分支探针在独立判别（AI/非 AI）下的准确率 $\{a_i\}_{i\in\{\text{spec,time,prosody,formant,hjorth}\}}$。为避免随机猜测带来的偏置，先做机会校正（chance-corrected）并归一化：

$$s_i \;=\; a_i - 0.5, \qquad w_i \;=\; \frac{s_i}{\sum_j s_j}, \qquad \sum_i w_i = 1. \tag{12}$$

根据实验结果（见表 5），可得各分支的权重 w_i 如式12所示，具体数值列于表 4。

Table 4.各分支探针权重（由表 5 的准确率经机会校正归一化得到）

分支（Branch）	权重 w_i
spec（频谱图）	0.1987
time（时域）	0.2067
prosody（音高音质）	0.1982
formant（共振峰）	0.1981
hjorth（Hjorth）	0.1982

评分规则 记各分支探针头给出的“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

$$S \;=\; \sum_{i} w_i\, p_i \;\in\; [0,1], \tag{13}$$

模型建立与求解 XVII

其中 w_i 为上式确定的分支权重。S 越大表示该音频越可能为 AI 生成。
探针准确率与权重 表 5 汇总了主分类器与各分支探针的独立准确率以及依据式 (1) 计算得到的分支权重。可以看到，time 分支（时域特征）在单分支判别中的准确率最高（78.34%），因此在综合评分中获得了略高的权重（0.2067）；其余四个分支的准确率非常接近（约 $77.17\% \sim 77.25\%$），对应的权重也基本均衡（约 0.198）。这种结果与我们在问题一中对多源特征互补性的观察一致：频谱类与声学统计特征在区分 AI 音频时具有相近而可互补的判别力。

Table 5.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	–

模型建立与求解 XVIII

分析讨论

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13的线性加权后，可在推理阶段以显式、可解释的方式衡量各模态对最终结论的贡献；当需要侧重某类伪迹（例如时域突变或共振峰异常）时，也可在不改变主干网络的前提下，通过更新 w_i 实现可控的偏好调整。

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。

第三问的目标是在尽量不影响可听性的前提下，通过对目标音频进行人为扰动和对抗性处理，考察模型检测 AI 音频的稳定性，并基于分析提出可操作的改进方案，使得模型在实际复杂场景中保持较高准确度与可解释性。为此，我们可以设计以下流程：扰动构造、鲁棒性分析、归因分析、改进方向。

为了研究检测模型在微小音频变动下的鲁棒性，我们设计了三类扰动方法。第一类基于多分支特征通道的细粒度信号处理，包括：(i) **时域分支**：使用轻度动态压缩器（RMS 检测 + 软拐点曲线）配合一阶全通滤波引入非线性相位扰动；(ii) **频域分支**：施加倾斜式均衡（spectral tilt）以及窄带峰值

模型建立与求解 XIX

EQ，以在高/低频端产生微小幅度变化；(iii) **类共振峰分支**：在 700 Hz、1500 Hz 等人声相关频段执行峰值增益/衰减，模拟共振峰位置的轻微漂移；(iv) **全局统计分支**：通过微小的时间伸缩改变节奏分布；(v) **音高/谐噪比分支**：对音高作 ± 0.15 semitone 的平移并加入低幅白噪声降低 HNR。上述扰动均设定了 `mild`、`medium`、`strong` 三档强度，并在处理后统一归一化至目标 RMS。

第二类为频率域参数均衡器扰动。我们预先设定若干组窄带或宽带 *peaking* EQ 参数（中心频率 f_0、增益 G、品质因数 Q），例如 $(f_0 = 12\,\text{kHz}, G = +1.5\,\text{dB}, Q = 4.0)$、$(f_0 = 150\,\text{Hz}, G = -1.0\,\text{dB}, Q = 1.0)$ 等，并利用 RBJ cookbook 公式设计二阶 IIR 滤波器逐通道施加。所有增益幅度控制在 ± 2 dB 内，以保证主观听感基本不变，但在频谱结构和特征空间中引入可测差异。

第三类是音轨合并扰动，我们寻找了常见的白噪音音乐片段（city park、forest、rain）与原始音频按照不同的响度比例合并以构造更具自然色彩的音频。

为评估所提出多分支 CNN 检测模型在不同扰动条件下的稳健性，我们基于前述三类扰动方法构造测试集，并分别计算主分类概率（`prob_main`）与多分支探针加权得到的综合评分 S。分析流程如下：首先，将原始测试集样本按扰动类型与强度划分为多个子集，涵盖：(i) **细粒度特征通道扰动**（包括 RMS/相位扰动、倾斜式 EQ、窄带峰值 EQ、共振峰微调、节奏微变及音高/谐噪比调整）；(ii) **频率域参数均衡器扰动**（多组中心频率、增益、品质因数组合的窄带或宽带 *peaking* EQ）；(iii) **音轨合并**

模型建立与求解 XX

扰动（与环境音如城市、公园、雨声按不同比例混合）。各扰动均设有 mild、medium、strong 三档强度，处理后统一归一化 RMS，以消除整体响度差异的影响。
随后，对每个扰动子集输入模型进行推理，统计其 prob_main 与 S 的均值及分布范围，并与原始样本结果对照，计算相对降幅和标准差变化，从而量化不同扰动对模型检测性能的冲击程度。
最后，将结果按扰动家族汇总，比较主分类头与融合评分在抗扰动能力上的差异，重点关注高频注入、环境噪声混入、共振峰微调等高风险扰动下，模型预测稳定性与置信度的变化规律。

Table 6.不同扰动类型下主分类概率与综合评分均值对比

扰动家族（Family）	示例	样本数	prob_main 均值	S 均值
comboAll	多种扰动叠加版本	1	**0.764**	**0.776**
pitchHNR	微转调 + 谐噪比调整	1	0.659	0.757
specEQ	频谱均衡/倾斜（EQ）	7	0.604	0.727
highFreqInject	高频注入（HI）	8	0.461	0.667
ambientNoise	环境噪声混入（雨声、森林、城市）	4	0.422	0.671
other	其他扰动方式	13	0.411	0.719

模型建立与求解 XXI

由表 6 可见，不同扰动类型对模型的影响程度差异显著。多种扰动叠加版本（comboAll）在主分类概率和综合评分上均维持在 0.76 以上，表明模型在面对复合扰动时依然具备较强的鲁棒性。轻微的音高平移与谐噪比调整（pitchHNR）同样表现出较好适应性，S 值达到 0.757。相比之下，频谱均衡/倾斜（specEQ）使主分类概率下降至 0.604，而高频注入（highFreqInject）与环境噪声混入（ambientNoise）对模型影响更为显著，prob_main 分别降至 0.461 和 0.422，综合评分亦明显降低。尤其是 ambientNoise，在较低响度比例下即可造成显著干扰，表明其在特征空间中对时域与频域模式均有破坏性。此外，“other”类扰动虽然类型多样，但平均表现略优于 ambientNoise，提示高频噪声与环境混入等扰动情形仍是提升模型鲁棒性的重点方向。

为识别检测器对不同声学特征的依赖，我们在分支嵌入空间（即各 CNN 分支输入到全连接层的一维张量）进行了降维可视化，分析数据点在各个分支上的区域可分性，同时使用 AUC、ACC、Fisher 比率和 logit 等指标来量化分析各个分支对最终决策的贡献程度。

(1) 区域可分性：时域分支更“可分”，formant 分支更“缠绕”。在图 3 中，时域嵌入出现了与主簇明显分离的红色带状簇，显示 AI 音频在能量包络/瞬态结构维度存在系统性轨迹；PCA（图 3）虽线性可分性有限，但仍呈现出沿第一主轴的标签渐变。

模型建立与求解 XXII

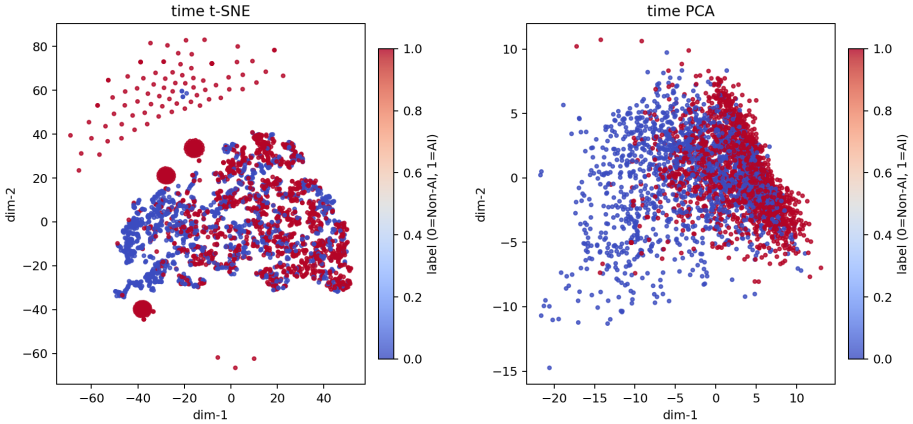


Figure 3.时域分支嵌入：t-SNE（左）与 PCA（右）。

模型建立与求解 XXIII

相对地，formant 分支（图 4）总体上红蓝高度交织，仅在若干局部“岛状”区域聚集，表明整体频谱包络并不足以稳定区分全部 AI 样式，但对特定风格（如固定声道/共振峰模板）的 AI 样本存在局部可分性。

模型建立与求解 XXIV

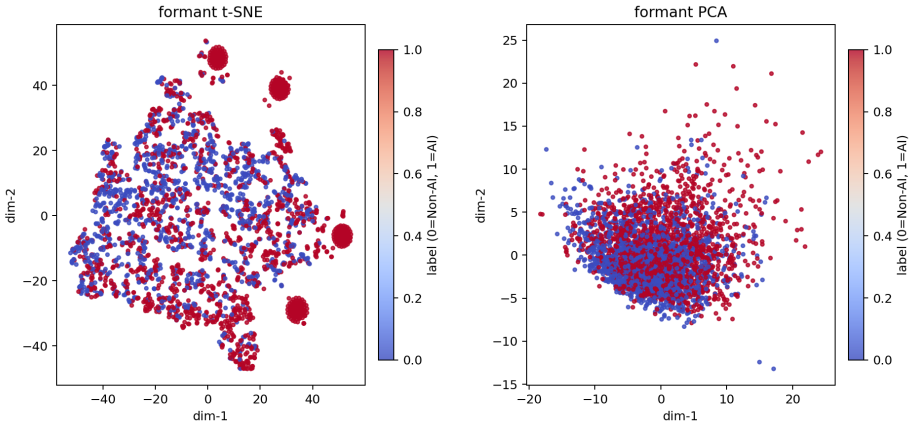


Figure 4.formant 分支嵌入：t-SNE（左）与 PCA（右）。

模型建立与求解 XXV

(2) 量化分析：有针对性的微扰能显著改变判别。我们对相同音源施加多类微扰并记录主分类概率 ($\mathrm{prob\_main}$) 变化。结果表明，时域动态扰动 (timeDyn_strong/medium/…) 与 类共振峰扰动 (formantish_strong/medium/…) 可将判别概率稳定压低（例如 timeDyn_strong ≈ 0.24，formantish_strong ≈ 0.30），显著低于原始片段；而单纯的高频窄带 EQ 微调（如 $15\,\mathrm{kHz}$、$\pm(0.8\sim1.8)\,\mathrm{dB}$）影响较弱。结合探针头输出可见，时域与 formant 分支单独判断时的 AI 概率降幅最大，表明模型决策高度依赖 **时域包络/动态**与**人声共振峰布局**两类特征；一旦这些特征被微调，决策边界即向“非 AI”区域偏移。

为了初步分析各个分支对最终决策的决策方向的影响，我们构造了如下两个指标来进行量化分析：

- **logit**：模型主分类器在融合所有分支嵌入后的线性层输出（未经过 Sigmoid）。logit 值的正负反映了模型决策的方向（正值偏向 AI，负值偏向非 AI），绝对值大小反映了决策的置信度。在分析中，logit 可作为样本被判定为 AI 或非 AI 的“力度”指标，用于衡量扰动前后决策信号的变化。

模型建立与求解 XXVI

- **contrib**：按分支拆分最终线性层权重，对该分支嵌入向量与权重做点积，得到该分支对 logit 的加和贡献。若无法直接分块，可用 Grad×Input 在分支拼接点近似计算。贡献值为正表示该分支推动决策向 AI 方向，为负则推动向非 AI 方向。通过统计 contrib 值的分布与类别差异，可以识别在整体判决中起主导作用的分支，并结合扰动实验分析其易受攻击性。

图 5 展示了不同分支在 AI 与非 AI 样本上的平均贡献值及标准误差。可以看到，spec 分支在两类样本中均呈显著正贡献（推高 AI 判别概率），time 分支则整体呈负贡献（抑制 AI 判别），且幅度较大；prosody、formant 与 hjorth 分支的贡献幅度较小，说明在多数样本中，它们对最终判别的直接推动作用有限。

模型建立与求解 XXVII

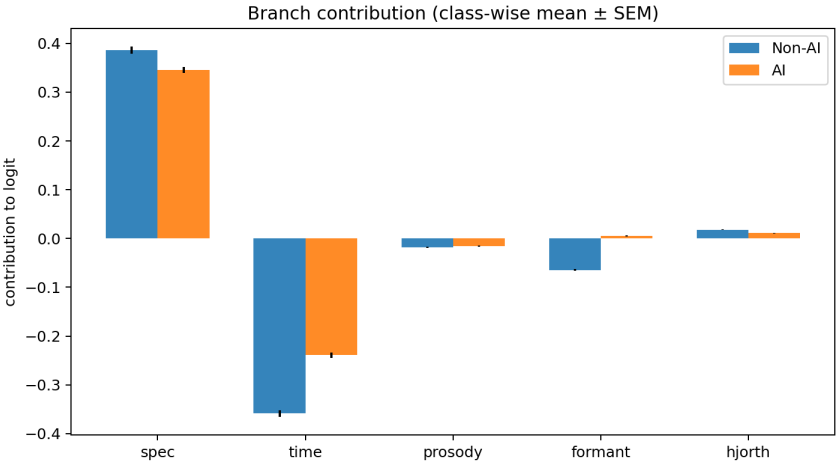


Figure 5.不同分支在 AI 与非 AI 样本上的平均 logit 贡献（均值 ± SEM）。

模型建立与求解 XXVIII

为了考察分支贡献的相关性，我们计算了样本层面的皮尔逊相关系数矩阵（图 6）。结果显示，spec 与 time 分支呈显著负相关，表明它们在多数样本的决策中存在一定的“对冲”关系——当一个分支强烈推动向 AI 方向时，另一个分支往往抑制该判别。此外，prosody 与 time、formant 的正相关，说明这些分支在部分样本中可能协同作用。

模型建立与求解 XXIX

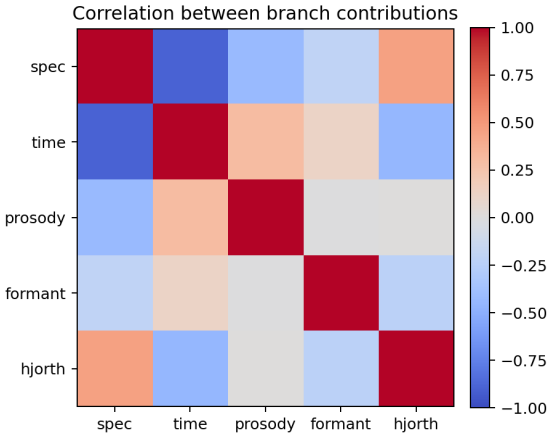


Figure 6.各分支 logit 贡献的相关性矩阵（皮尔逊相关系数）。

图 7 给出了所有样本的分支贡献分布。可以看出，time 分支的贡献分布跨度最大，从强负到强正均有覆盖，说明它在部分样本中具有极高的判别权重，而在另一些样本中几乎不参与判别；spec 分支则分布相对集中且整体偏正，体现了其在多数样本中稳定推高 AI 判别的趋势；formant 分支分布较分散，但极端值较多，可能在某些特定声学风格下起关键作用。

模型建立与求解 XXXI

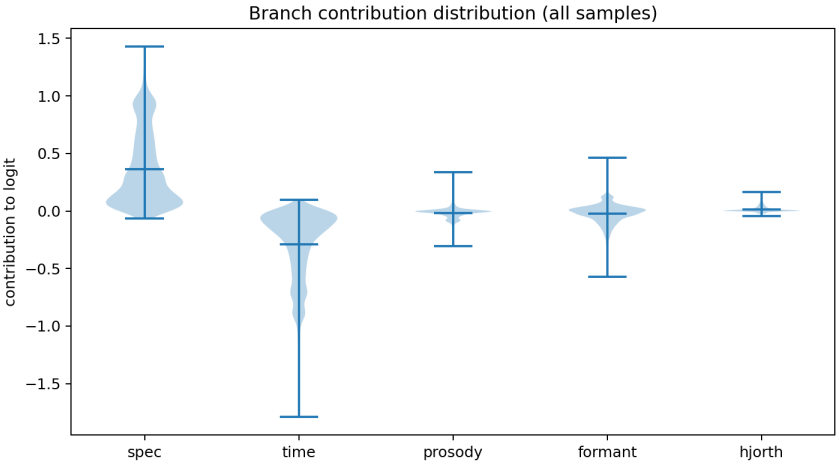


Figure 7.各分支 logit 贡献的分布（所有样本）。

模型建立与求解 XXXII

我们为每个分支计算了如下指标，表征其决策能力，分析最终决策对各个分支的依赖性：

- **AUC（Area Under the ROC Curve）**：表示在所有可能的分类阈值下，模型正确区分正负样本的概率, 由此绘制出图 8。ROC 曲线数据点越集中在左上角，则特定阈值下召回率与误报率之比越高，AUC 值越接近 1，说明该分支嵌入在全范围阈值下的判别能力越强；AUC $= 0.5$ 表示无判别能力，相当于随机猜测。
- **ACC（Accuracy）**：表示在固定分类阈值（本文取 0.5）下，分支嵌入在验证集上的分类准确率。ACC 关注的是单一阈值下的硬判别性能，反映该分支在常规决策条件下的直接预测效果。
- **Fisher_mean**：衡量类间分离度与类内紧凑度的比值。其定义为：

$$\text{Fisher Ratio} = \frac{\sum_d (\mu_d^+ - \mu_d^-)^2}{\sum_d (\sigma_d^{+2} + \sigma_d^{-2})}$$

其中 μ_d^+ 和 μ_d^- 分别为第 d 维在正、负样本上的均值，σ_d^{+2} 和 σ_d^{-2} 为对应方差。Fisher 比率越大，说明该分支的嵌入向量在不同类别间分布差异更明显，且类内方差更小，有利于判别。

模型建立与求解 XXXIII

这种扰动敏感性与分支的单独判别能力高度一致：定量评估显示，时域分支在 5 折交叉验证下的 AUC 达到 0.8251 ± 0.0076，ACC 为 0.7627 ± 0.0025，Fisher 比率 0.1324；formant 分支的 AUC 为 0.8161 ± 0.0077，ACC 为 0.7413 ± 0.0101，Fisher 比率 0.0912。两者均显著高于 prosody（AUC 0.7191，Fisher 0.0071）与 hjorth（AUC 0.6963，Fisher 0.0054）等弱分支，说明模型在融合决策时，确实更依赖时域包络/动态与人声共振峰布局等特征；当这两类特征被结构性微调时，主分类概率便会大幅偏向“非 AI”区域。

模型建立与求解 XXXIV

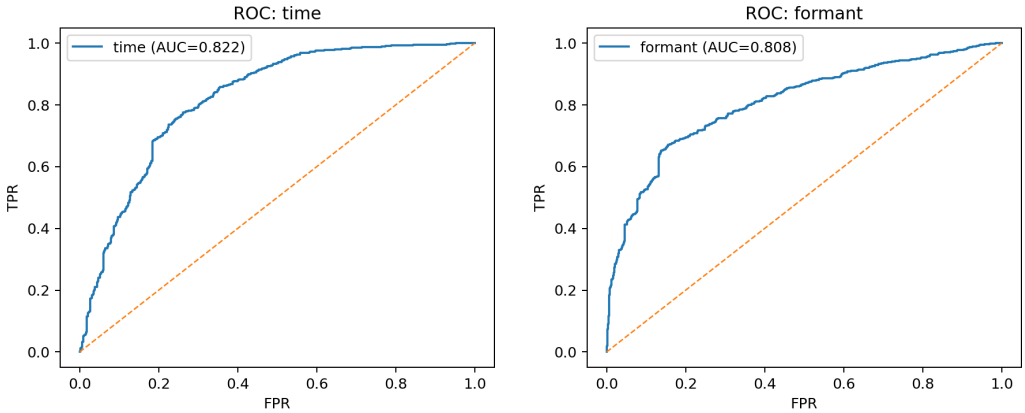


Figure 8.时域分支与 formant 分支 ROC 曲线。

模型建立与求解 XXXV

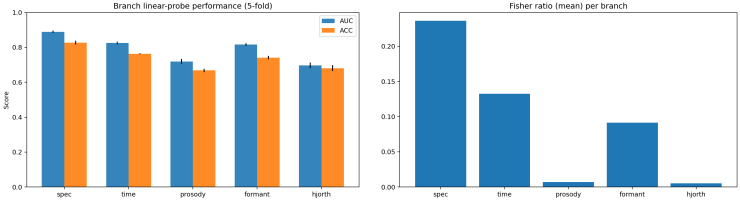


Figure 9.分支可分性定量指标：AUC/ACC（中）、Fisher 比率（右）。

(3) 偏差与失败模式。在排序表中，自然环境纹理（如雨声、林声、城市公园环境）也获得较低的 AI 概率，提示模型对非乐音/非稳态素材存在天然“真值”偏好，可能源于训练集分布差异（音乐性与稳态谐波成分占比更高)。这类偏差意味着：通过轻度压缩 + 全通相位改变瞬态，以及小幅峰值 *EQ* 微移 formant，即可在不破坏主观听感的前提下，提高“过检”概率。

综合嵌入结构与扰动响应，我们将特征归因排序为：**时域动态（首要）** > **频谱包络/共振峰（次要）** > **高频细窄带均衡（较弱）**。为缓解脆弱性，未来可以考虑在训练中加入：(i) 动态压缩、全通相位、

模型建立与求解 XXXVI

微小时间伸缩的强数据增强；（ii）formant 随机漂移/声道随机化；（iii）在多分支架构上引入一致性正则或分支 dropout，以降低单一分支的主导性，并对自然环境纹理补充“反例”样本以纠正数据偏差。

模型优缺点与推广 I

模型优点 本文提出的多分支 CNN 检测模型能够同时处理时域、频域、音高与音质等多种声学特征，结构上实现了多源信息的并行提取与融合，充分利用了不同特征维度之间的互补性。实验结果表明，该模型在常规测试集上取得了较高的分类准确率（验证集稳定在 89%–90%），且在多数轻中度扰动条件下保持了良好的检测性能；在较强扰动下，全连阶层输出的 AI 判别概率会有明显扰动，但是，综合五个分支探针头的概率计算出的 AI 痕迹评分（式 13），在强扰动下仍然坚挺；由此可见模型具备一定的泛化能力与鲁棒性。此外，模型规模相对轻量，计算开销低，易于在多平台部署，并可支持实时检测需求。

模型不足 尽管整体性能优良，但鲁棒性分析发现模型在高频噪声注入（highFreqInject）与环境噪声混入（ambientNoise）等特定扰动下的检测概率显著下降，表明在面对宽带噪声和高频干扰时的特征表达仍存在脆弱性。此外，当前特征融合层为固定权重，未能根据输入特征的质量动态调整各分支贡献，这在某些特征受扰较大的情况下可能限制了模型的自适应能力。训练数据主要集中于有限风格与生成引擎，对极端风格、全新算法生成的音乐仍可能存在适配不足。

模型推广 该多分支 CNN 框架具备良好的可扩展性，可根据不同检测任务灵活替换或新增特征分支。例如，在语音深度伪造检测中，可增加语音特有的相位一致性特征分支；在环境音场景识别中，可引入空间声学特征分支。模型也可与自编码器异常检测方法结合，用自监督方式预训练分支特征提取器，再进行有监督分类，从而提升在低资源或开放域场景下的表现。进一步地，可将该方法与传

模型优缺点与推广 II

统机器学习特征融合（如 XGBoost、SVM）进行混合，形成多模态、多尺度的综合检测框架，推广到更多 AI 音频内容安全相关领域。