从时域到频域：基于多分支 CNN 网络的 AI 音频检测模型

NKUMMF2025138

October 7, 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

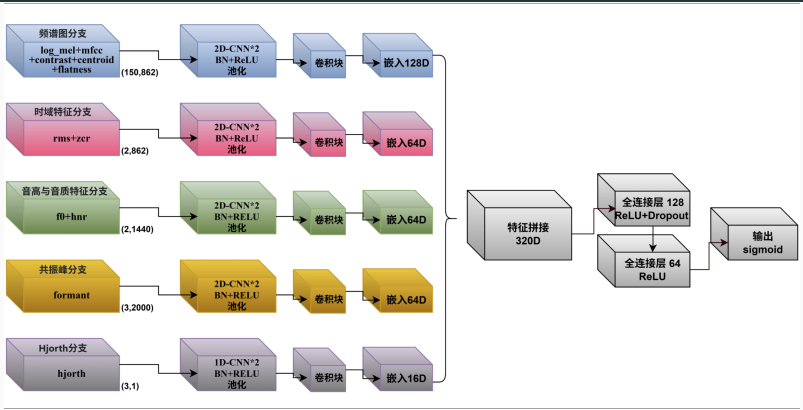


Figure 1.多分支 CNN 示意图

模型建立 i

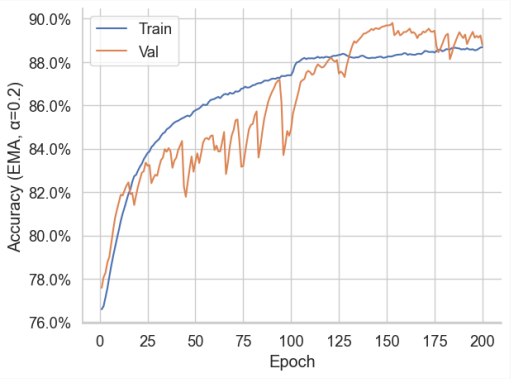
分支名称	输入特征	处理与输出
频谱图分支（Spectrogram）	log_mel, mfcc, contrast, centroid, flatness	2D CNN 提取时频模式 → GAP → 128 维表示
时域分支（Time-Domain）	rms, zcr	1D CNN 提取时间结构 → 64 维向量
音高与音质分支（Pitch & Quality）	f0, hnr	1D CNN 建模调型与音质 → 64 维向量
共振峰分支（Formant）	formant	1D CNN 处理时间序列 → 64 维向量
Hjorth 参数分支（Hjorth Param.）	hjorth (Activity,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

Table 2.多分支特征处理结构概览

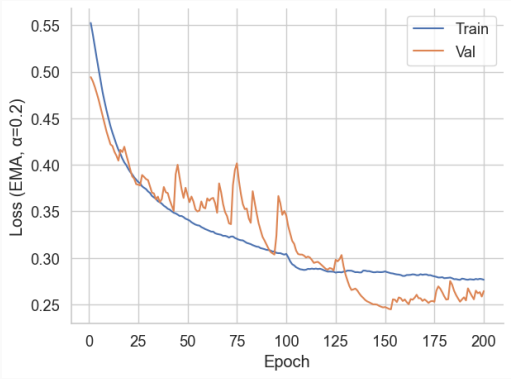
模型建立 ii

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。

鲁棒性分析

鲁棒性分析 i

- 构造扰动
- 分析扰动实验结果
- 结合**数据流**进行归因分析

扰动构造 i

从模型和数据两个角度入手，构造扰动：

1. **模型角度：**修改音频，对模型采用的是一个特征进行直接干扰
   - **时域分支：**轻度动态压缩、滤波器、相位扰动；
   - **频域分支：** EQ 频段调节；
   - **类共振峰分支：**对常见人声频段进行增益/衰减；
   - **全局统计分支：**局部加速/减速，改变节奏特征；
   - **音高与谐噪比分支：**对音高作 ± 0.15 semitone 平移，加入低幅白噪声。
2. **数据角度：**音轨合并扰动
   将原始音频与自然环境噪声片段（如 *city park, forest, rain*）按不同响度比例混合，构造更具真实感的样本。

扰动实验结果及分析 i

对真实标签为"AI 生成" 的音频施加前述两类扰动，构造测试集，分别计算：

$$S = \text{加权融合评分}, \quad \texttt{prob\_main} = \text{主分类概率}$$

每类扰动设置 `mild`、`medium`、`strong` 三档强度，并统一归一化 RMS 以消除响度影响。

不同扰动类型下的鲁棒性表现 i

下表汇总了五组混淆效果较大的实验结果。

Table 5.不同扰动类型下主分类概率与综合评分均值对比

扰动方式	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	音高调整 + 白噪音注入	1	0.659	0.757
specEQ	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

不同扰动类型下的鲁棒性表现 i

分析：

- **总体表现：**多种扰动叠加（comboAll）在主分类概率和综合评分上仍保持在 0.76 以上，模型具备较强鲁棒性；
- **轻度扰动：**音高与谐噪比调整（pitchHNR）对性能影响较小，综合评分 $S = 0.757$；
- **中等扰动：**频谱均衡/倾斜（specEQ）使主分类概率下降至 0.604；
- **高风险扰动：**高频注入与环境噪声混入显著降低模型输出置信度，prob_main 分别降至 0.461 和 0.422；
- **综合结论：**环境噪声在较低响度比例下即可破坏时域与频域模式，是当前模型鲁棒性提升的关键方向。

量化分析：有针对性的微扰能显著改变判别 i

实验思路：对相同音源施加多类微扰，记录主分类概率（prob_main）变化。

主要发现：

- 时域动态扰动（`timeDyn_strong/medium/...`）与 类共振峰扰动（`formantish_strong/medium/...`）显著降低分类概率；
- 典型结果：

$$\texttt{timeDyn\_strong} \approx 0.24, \quad \texttt{formantish\_strong} \approx 0.30,$$

  均远低于原始片段；
- 高频窄带 EQ 微调（$15\,\mathrm{kHz}, \pm(0.8 \sim 1.8)\,\mathrm{dB}$）影响较弱；

量化分析：有针对性的微扰能显著改变判别 ii

- 探针结果表明，模型决策高度依赖：
  - **时域包络/动态特征**；
  - **人声共振峰布局特征**。

一旦这些特征被微调，决策边界明显向“非 AI”区域偏移。

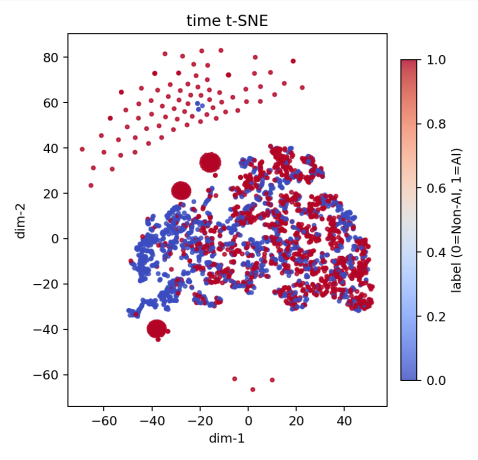
归因分析：分支数据流特征与扰动敏感性的关联 i

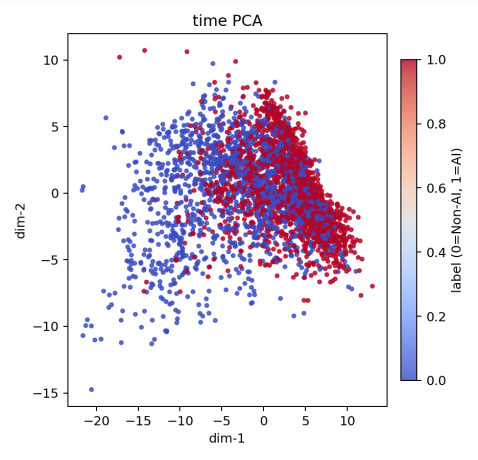
目的：探究检测器在不同分支上的数据流特征与其对扰动的敏感性之间的关系。

方法：

- 对各 CNN 分支的嵌入空间（即输入全连接层前的一维张量）进行降维可视化；
- 分析各分支在嵌入空间中的类别可分性；
- 采用 AUC、ACC、Fisher 比率与 logit 统计量量化每个分支对最终决策的贡献。

归因分析：分支数据流特征与扰动敏感性的关联 i

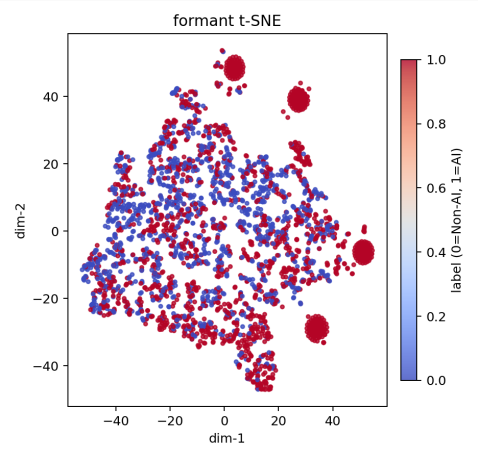


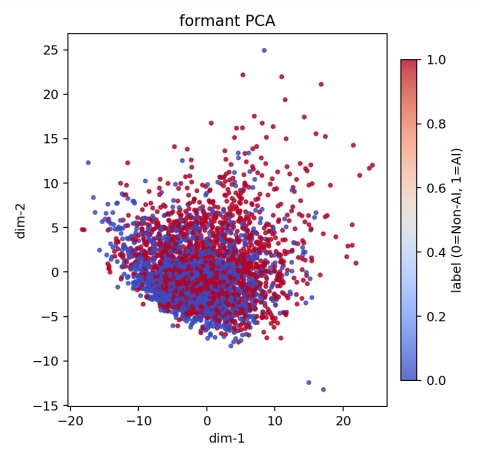


归因分析：分支数据流特征与扰动敏感性的关联 ii

Figure 3.时域分支嵌入可视化：t-SNE（左）与 PCA（右）。

归因分析：共振峰分支嵌入特征 i





归因分析：共振峰分支嵌入特征 ii

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

量化分析：logit 与 contrib 指标定义 i

为分析各分支对最终决策的影响方向，定义两个指标：

1. logit **指标：**

- 为主分类器融合所有分支嵌入后线性层的输出（未经过 Sigmoid）；
- 正值表示模型倾向判为 AI，负值表示倾向 **非** AI；
- 绝对值大小反映决策置信度；
- 可视为样本被判为 AI 的“力度”指标，用于衡量扰动前后决策信号变化。

2. contrib **指标：**

量化分析：logit 与 contrib 指标定义 ii

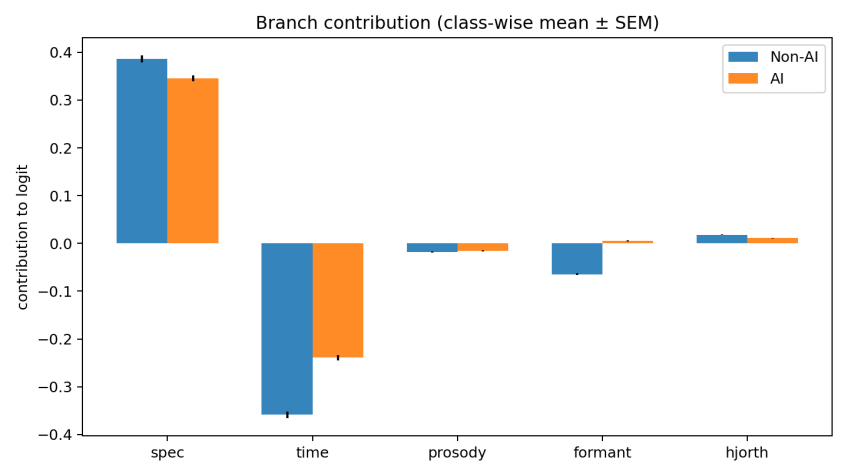
- 将最终线性层权重按分支拆分，对该分支嵌入与权重求点积：

$$\text{contrib}_k = \mathbf{w}_k^\top \mathbf{h}_k,$$

  其中 k 表示第 k 个分支；
- 若无法显式分块，可用 Grad×Input 在拼接点近似计算；
- contrib 值为正：推动决策向 AI；为负：推动决策向非 AI。

量化分析结果：分支贡献对比 i

量化分析结果：分支贡献对比 ii



量化分析结果：分支贡献对比 iii

Figure 5.各分支在 AI 与非 AI 样本上的平均贡献值（均值 ± 标准误差）。

量化分析结果：分支贡献对比 i

结果概要：

- 图 5 展示了不同分支在 AI 与非 AI 样本上的平均贡献值（均值 ± 标准误差）；
- `spec` 分支在两类样本中均呈显著**正贡献**，推高 AI 判别概率；
- `time` 分支整体为**负贡献**，对 AI 判别起抑制作用，且幅度较大；
- `prosody`、`formant` 与 `hjorth` 分支的贡献幅度较小，表明其对最终判决的直接推动作用有限。

量化分析：分支贡献的相关性与对冲关系

平均贡献分析：

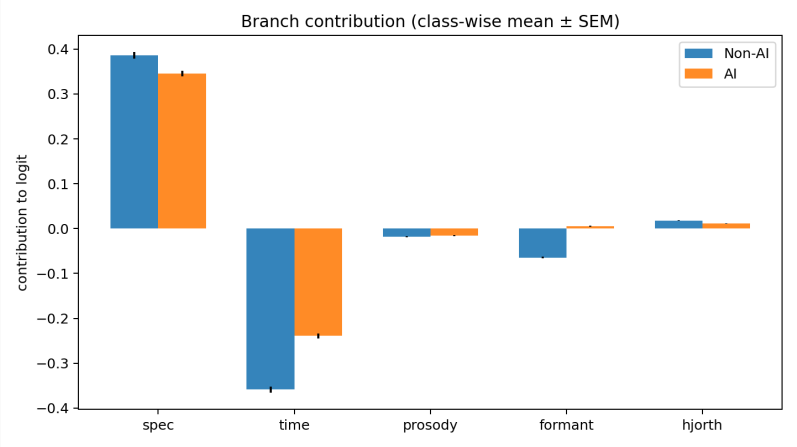
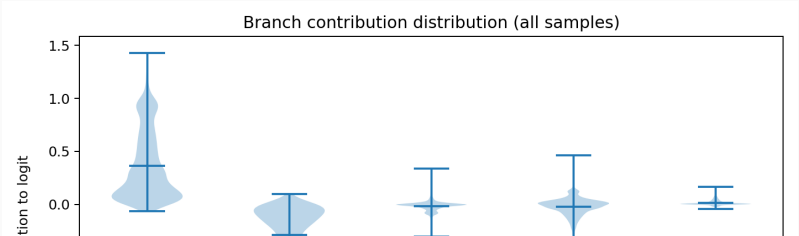


Figure 6.不同分支在 AI 与非 AI 样本上的平均 logit 贡献（均值 ± SEM）。

量化分析：分支贡献的分布特性

样本级分布：

- 图 8 展示各分支在所有样本上的 logit 贡献分布；
- `time` 分支：分布跨度最大，从强负到强正均有覆盖；说明其在不同样本中权重差异显著；
- `spec` 分支：分布集中且整体偏正，表明其稳定推高 AI 判别；
- `formant` 分支：分布较分散、极端值多，可能在特定声学风格下起关键作用；
- 其他分支（`prosody`、`hjorth`）贡献较弱，说明其在多数样本中的直接作用有限。



量化分析：分支判别能力指标定义

目的：评估各分支嵌入的判别能力，分析最终决策对不同分支的依赖性。

主要指标：

- AUC（Area Under the ROC Curve）
  - 表示在所有可能阈值下模型正确区分正负样本的概率；
  - ROC 曲线越靠近左上角，说明召回率高、误报率低；
  - AUC 值越接近 1 → 判别能力越强；AUC = 0.5 → 等价于随机猜测；
  - 图 9 展示各分支的 ROC 曲线。
- ACC（Accuracy）
  - 在固定分类阈值（本文取 0.5）下的分类准确率；
  - 反映分支在常规决策条件下的直接预测性能；
  - 强调“硬判别”能力。
- Fisher **比率**（Fisher_mean）
  - 衡量类间分离度与类内紧凑度的比值：

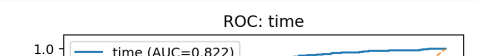
$$\sum_i (\mu_i^+ - \mu_i^-)^2$$

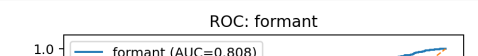
量化分析结果：主要分支的判别能力

结论概览：扰动敏感性与各分支的单独判别能力高度一致。

定量结果（5 折交叉验证）：

- **时域分支（time）：** AUC $= 0.8251 \pm 0.0076$，ACC $= 0.7627 \pm 0.0025$，Fisher 比率 $= 0.1324$；
- **formant 分支：** AUC $= 0.8161 \pm 0.0077$，ACC $= 0.7413 \pm 0.0101$，Fisher 比率 $= 0.0912$；
- 均显著高于 `prosody`（AUC 0.7191，Fisher 0.0071）与 `hjorth`（AUC 0.6963，Fisher 0.0054）；
- 表明模型决策更依赖：
  - **时域包络 / 动态结构**；
  - **人声共振峰布局特征**。





量化分析结果：分支可分性指标对比

对比分析：

- 图 10 展示各分支的 AUC、ACC 与 Fisher 比率；
- 时域与 formant 分支在三个指标上均占优；
- prosody 与 hjorth 分支指标普遍偏低，表明其独立判别能力有限；
- 说明融合决策阶段模型主要依赖结构性强、时频特征稳定的分支。

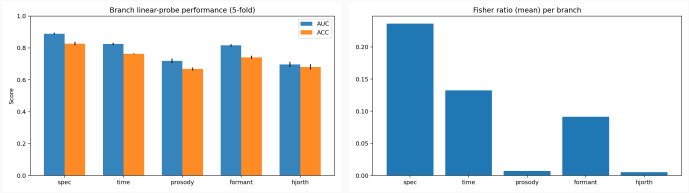


Figure 10.分支可分性定量指标：AUC/ACC（左）、Fisher 比率（右）。

总结：当时域包络或共振峰结构被扰动时，模型主分类概率显著向“非 AI”区域偏

偏差与失败模式

观察：

- 自然环境纹理（雨声、林声、城市公园等）在排序表中获得较低 AI 概率；
- 模型对非乐音 / 非稳态素材存在天然“真值”偏好；
- 该偏差可能源于训练集分布不均：音乐性与稳态谐波成分占比更高。

启示：

- 通过**轻度压缩** + **全通相位**改变瞬态，或**小幅峰值** EQ 微移 formant，可在不改变主观听感的前提下显著提高“过检”概率；
- 说明模型在**稳态性**与**谐波结构**上的依赖导致偏差。

改进方向

特征归因排序：

时域动态（首要） ＞ 频谱包络 / 共振峰（次要） ＞ 高频细窄带均衡（较弱）

改进建议：

1. **数据增强：**在训练中加入动态压缩、全通相位、微小时间伸缩等扰动；
2. **特征随机化：**引入 formant 随机漂移与声道随机化；
3. **结构改进：**在多分支架构中加入一致性正则或分支 dropout，降低单一分支主导性；
4. **数据再平衡：**对自然环境纹理补充反例样本，缓解训练数据偏差。

模型优缺点

模型优点

- 多分支 CNN 可并行提取时域、频域、音高与音质特征；
- 融合层充分利用多源信息的互补性；
- 验证集准确率稳定在 89%–90%，对多数轻中度扰动具有良好鲁棒性；
- 五分支探针头融合的 AI 痕迹评分在强扰动下仍保持稳定；
- 模型轻量、计算成本低、可实时检测并跨平台部署。

模型不足

- 在高频噪声注入（`highFreqInject`）与环境噪声混入（`ambientNoise`）下性能下降明显；
- 特征融合层为固定权重，缺乏对分支质量的动态自适应；
- 训练数据覆盖有限，对新风格或新算法生成的音乐适配性不足。

模型推广与应用前景

框架扩展性

- 多分支结构可灵活替换或新增特征分支；
- 语音伪造检测：加入相位一致性特征；
- 环境音识别：引入空间声学特征；
- 可结合自编码器进行自监督预训练，提高低资源场景性能。

混合检测思路

- 将本模型与传统机器学习（如 XGBoost、SVM）特征融合；
- 构建多模态、多尺度检测框架；
- 推广至更多 AI 音频内容安全与生成溯源领域。

→ 面向开放域、低资源与实时检测的统一声学特征框架

谢谢大家！