arXiv:2301.01578v1 [cs.SD] 4 Jan 2023

音乐信息研究实验中的效度

Bob L. T. Sturm[1] 和 Arthur Flexer[2]

[1]语音、音乐与听觉部，斯德哥尔摩皇家理工学院，瑞典，
bobs@kth.se[2]计算感知研究所，林茨约翰·开普勒大学，奥地利，
arthur.flexer@jku.at

摘要

效度是依据证据（例如在实验中收集的数据）所作推断的真实性，是科学研究的核心。鉴于音乐信息研究（MIR）这一领域已相当成熟，我们认为，效度应当比迄今为止得到更多的讨论和考量。在研究工作中考量效度，能够提升其科学与工程价值。透过效度的视角，对抗攻击和性能瓶颈等令人费解的音乐信息检索现象也会变得不那么神秘。本文回顾效度这一主题，并结合一篇重要参考文献 Shadish 等人 [2002]，探讨效度的四种主要类型。我们以一种典型的音乐信息检索实验——使用机器学习进行音乐分类——为基础，讨论这些类型。借此，可引导音乐信息检索实验者从实验收集的数据中作出有效推断。

1 引言

音乐信息研究（MIR）是一门多学科领域，致力于让各种用户能够获取音乐及有关音乐的信息。其应用范围从搜索和检索系统（如 Shazam [Wang、 2003] 或 TunePal [Duggan 和 O'Shea， 2011]，），到推荐系统（如 Spotify 或 Pandora），甚至包括生成音乐等更具创造性的应用。这些系统由许多计算组件构成，工作于声学与符号领域之间的可计算区域 [M¨uller， 2015]。对于音乐信息研究人员而言，这些复杂系统及其组件的有效性和信度至关重要，其他利益相关者也同样关注这一点。

音乐信息研究人员会开展实验，比较音乐数据建模与检索方法。研究的主要重点是用户，但与用户开展实验的成本高昂，而且这类研究难以复现。因此，人们开始采用计算机实验，以"测试

集”作为人类用户的替代，称为克兰菲尔德范式[Cleverdon，1991]：研究者将音乐信息检索系统应用于问题领域的代表性样本（例如查询和相关文档），以预测其效用。尽管这种方法成本低且可复现，但其在音乐信息检索乃至一般信息检索中的相关性和信度仍受到质疑 [Voorhees, 2001, Urbano 等人，2013]。

在克兰菲尔德范式下，最先进的音乐信息检索系统在复现某些数据集的真实标注方面表现出色，例如从音频数据中推断节奏、音乐流派或情感。然而，对音频稍作无关紧要的变换，就可能突然令这些系统失效 [Sturm，2014b，Kereliuk 等，2015，Rodr´íguez-Algarra 等，2016，Prinz 等， 2021]。有一个案例中 [Rodr´íguez-Algarra 等， 2016], 某个“音乐流派识别”系统依赖次声特征，这些特征似乎源自数据收集过程，但人类听众无法感知，且与任务无关。在另一个案例中[Sturm 等人， 2015], 某个“节奏识别”系统似乎利用速度来推断节奏，这种混淆同样源自数据收集过程。

相关讨论还涉及音乐信息检索系统的“天花板效应” [Aucouturier 和 Pachet，2004年，Pohle 等， 2008], 即观察到的系统性能提升障碍——无法达到完美水平或人类水平——是因为从音频录音中提取的可计算特征缺少音乐的心理物理因素和文化因素 [Wiggins， 2009]。音乐信息研究人员若致力于大幅提升既定问题和数据集上的这些指标，实际上可能不会给人类用户带来任何实际影响 [Urbano 等人， 2012]。Flexer 和 Grill [2016]指出了导致性能瓶颈的一个原因：用于训练和测试音乐信息检索系统的数据可能来自定义不当的任务（例如音乐相似性），因此基于这些数据追求更好的指标毫无意义。研究问题缺乏明确界定，促使 Sturm 和 Collins [2014] 提议音乐信息研究人员针对更难、定义不当的问题，解决合成但定义明确的简化问题。

MIR 实验的核心，是从实验结果得出的结论与其效度，或“真值”之间的关系 [Shadish 等人， 2002]。理想情况下，实验应经过精心设计并加以实施，以回答一个定义明确的假设。[1] 实验的各个组成部分——实验单位、处理、实验设计、观测和实验环境——都应经过操作化（将理论转化为实践），以提升质量并降低成本（金钱、时间和伦理代价）。这正是实验设计这门学科所探讨的：如何以最低成本获取最有力的证据？然而，浏览 MIR 首屈一指的会议 ISMIR[2] 的会议论文集及相关出版物，会发现音乐信息检索社群普遍缺乏对效度的认识。

最早明确将效度引入音乐信息检索（MIR）的文献是 Urbano [2011]，他结合文本信息检索（IR）文献讨论效度，并具体倡导对 MIREX 结果进行基于效度的元评估。[3] 这

[1]我们鼓励读者阅读 Chase [2001]，了解实验者为了检验哪怕是一个简单的假设，必须付出多么惊人的努力。[2]https://ismir.net/conferences/[3]“音乐信息检索评估交流”（MIREX）是一项年度评估

一观点在 Urbano 等人 [2013], 的研究中得到进一步发展，他们采用了 Shadish 等人 [2002], 提出的效度类型学，并引入信度和效率的概念。Schedl 等人 [2013] 的综述虽未明确提及效度，甚至也未提及克兰菲尔德范式，却清楚表明计算机实验对音乐信息检索至关重要，并在结论中批评人们普遍忽视用户。Sturm [2013]指出，仅报告准确率等评价指标，并不足以判断音乐信息检索系统是否真正识别了音乐信号中的“音乐流派”，或是否依赖无关的混杂因素；他随后指出，这些未受控制的因素会危及从此类实验得出的结论的效度 [Sturm，2014b， 2017]。Sturm 等人 [2014] 试图界定音乐描述（包括“用例”），以此推动以能够得出有效且相关结论的方式评估音乐描述系统。

最近，Rodr´ıguez-Algarra 等人 [2019]提出并测试了一种方法，用于检测音乐信息检索实验中是否存在混杂，从而解决由此得出的结论缺乏构念效度的问题。Liem 和 Mostert[2020] 通过让音乐信息检索系统面对与训练数据不同的数据，研究了音乐流派等高层次音乐概念的构念效度。Flexer 等人 [2021] 批评了关于一般音乐相似性的实验缺乏外部效度，并要求聚焦于音乐相似性的具体方面，同时明确具体用例。最近一次试图全面探讨音乐信息检索领域效度与信度的工作，可能是 Urbano 和 Flexer 在 ISMIR 2018 [举办的一场尚未发表的教程 2018]。

以上对音乐信息检索领域效度相关先前工作的回顾表明，相关研究分散而零碎，且仅见于相对少数的出版物中。尽管已有少数声音呼吁解决音乐信息检索实验中的重大方法论问题，以提升该领域的效度，例如 Urbano [2011], Schedl 等人 [2013], Sturm [2014b], Urbano 和 Flexer [2018], Liem 和 Mostert [2020],Flexer 等人 [2021], ，但迄今仍未有出版物系统性地批判探讨效度在音乐信息检索语境中的含义，以及在实验设计、实施和分析时应如何考量效度。

本文回顾 Shadish 等人提出的四种主要效度类型。 [2002], Shadish 等人的著作是关于因果推断和实验科学中效度问题的权威资源。其他类型学也存在，例如 Lund [2021], ，但我们聚焦于 Shadish 等人的类型学， [2002] 因为它是一个公认的参考框架，而且已由 Urbano 等人引入音乐信息检索领域。 [Urbano 等人， 2013]对于每种类型，我们都会讨论常见威胁，将其与音乐信息检索学科整体联系起来，并进一步结合一个典型的音乐信息检索实验进行具体说明，该实验见第 2 节。我们还提供了一个 Python Jupyter 笔记本，用于运行实验并计算下文讨论的各类统计数据和结果。[4] 第 3 节回顾了实验的各个组成部分，效度讨论也以此为基础。第 4–7 节分别回顾四种效度类型，并提出一些可操作的问题，帮助音乐信息研究人员审视他们从实验中得出的结论。

音乐信息检索算法宣传活动 [Downie, 2006].

[4] https://github.com/boblsturm/mirvalidititytutorial

	准确率	精确率	召回率	F1分数
LDA	0.714	0.711	0.711	0.703
QDA	0.719	0.715	0.723	0.717
1NN	0.662	0.644	0.635	0.638
3NN	0.681	0.673	0.651	0.656
5NN	0.719	0.699	0.687	0.689
7NN	0.695	0.669	0.656	0.659
9NN	0.700	0.681	0.664	0.668
unif	0.12 ± 0.02	0.13 ± 0.03	0.12 ± 0.02	0.12 ± 0.02
freq	0.13 ± 0.02	0.13 ± 0.03	0.13 ± 0.02	0.13 ± 0.02
maj	0.16	0.02	0.12	0.03

表1：在 BALLROOM 测试分区中观察到的若干模型的准确率、宏平均精确率、召回率和 F1分数 [Dixon 等人， 2004]。表中还展示了两个随机选择标签的模型（含标准差）的表现，分别标记为：均匀 采样标签；频率按照训练数据标签频率采样标签。最后一行 多数类策略 展示了选择训练数据中最常见标签的模型的表现。

2 非典型 MIR 实验

为了以具体实例阐明我们对效度的一般讨论，我们介绍一个典型的音乐信息检索实验，它体现了相当一部分音乐信息检索研究的特点：使用机器学习和基准数据集进行音频分类。以 BALLROOM 数据集为例， [Dixon 等人， 2004], 该数据集于 2004年左右创建，数据来自一个专注于交谊舞的网站上出售的音乐 CD 片段。BALLROOM 包含 698 段音频录音，每段约 30 秒，并被标注为八种舞蹈风格或音乐节奏之一，例如“华尔兹”。BALLROOM 已出现在数十项研究中， [Sturm， 2017], 并且最常用于测量不同分类器与特征组合所复现的真实标签数量。

我们随机进行 70/30 划分，其中训练/测试数据集分别包含 488/210 个录音。我们对每段录音计算频谱通量起音强度包络，跳步大小为 1024 个采样点，即 46 毫秒 [McFee 等人， 2015]。我们计算该包络的归一化自相关，并保留与 $[0.23, 4.14]$ 秒内滞后对应的部分。最后，我们使用 12 阶自回归模型对自相关进行建模，得到一个包含 13 个维度的特征，用于描述这段 30 秒的音频录音。我们将训练特征的每个维度归一化为零均值和单位方差。我们使用与归一化训练数据特征相同的参数，对测试数据特征进行归一化。

我们使用多元高斯分布（线性判别分析和二次判别分析，即线性判别分析和二次判别分析），或 K 近邻来对特征建模。我们将模型推断出的标签与真实标注进行比较，并计算准确率、宏平均精确率、召回率和 F1分数。[5] 表1展示了

[5]宏平均评价指标是对各类别测得的评价指标取平均值，不考虑每个类别的观测数量。

本实验的结果，可通过补充材料中的 Python Jupyter 笔记本复现。[6] 最后几行展示了另外三种策略的统计数据：均匀 按照均匀先验对采样点标签进行抽取，频率按照训练数据标签频率对采样点标签进行抽取；[7] 多数类策略 则选择训练数据中频率最高的标签。这些策略可作为比较基准。

有人可能会批评说，上述实验并不能真正代表当今的音乐信息检索实验：如今的数据集规模大了好几个数量级，而且在许多情况下，深度学习已使特征工程变得过时。不过，我们仍将继续采用这个例子，原因有四。首先，BALLROOM 数据集已得到如此深入的分析，以至于基于它开展的分类实验能够以清晰易懂的方式展示四种效度。使用规模更大的数据集也能说明同样的问题，但首先需要对其构成和来源进行全面分析。其次，仅仅增加数据量或改变机器学习方法，并不能消除人们对效度的担忧。第三，上述实验的实验设计颇具典型性：收集并处理一个数据集，将其用于多种机器学习方法，再把训练好的系统应用于某个测试集并据此进行测量。最后，BALLROOM 数据集独具特色：十年后，人们利用同一网络资源创建了它的一个“扩展”版本[Marchand 和 Peeters， 2016]。因此，我们可以有意义地检验关于在 BALLROOM 上训练的系统所得结论的外部效度，例如，从 BALLROOM 中学到的概念是否适用于舞厅舞曲，而不仅仅适用于 2004年的舞厅舞曲（见第 7.2 节）。

实验的 3 个组成部分

在讨论从典型的音乐信息检索实验中得出的结论的效度之前，我们必须明确其实验单位、处理、实验设计、观测和实验环境。处理是施加于实验单位以产生（或不产生，在对照的情况下）效果的事物，实验单位是接受处理的事物，而观测则是对实验单位测量的内容。实验设计规定了对哪个实验单位施加哪种处理，而实验环境涉及时间、地点和条件。为了让这些概念更具体，我们来看一项医学实验，其中研究的是某种处理对血压的影响。研究人员从一个总体中抽取若干人，其中一些人接受该处理，其他人则接受安慰剂（对照）。实验设计说明哪些人接受处理，哪些人不接受。观测指标是一个月后某人的血压。实验环境可以包括总体的具体特征（农村或城市）、处理地点（医院或家中）等。实验研究者对比各组的血压观测结果，以得出处理是否有效的结论。

回到典型的音乐信息检索实验，我们希望确定不同机器学习（ML）算法在预测

[6]参见脚注 4。[7]这些评价指标的均值和标准差可以通过理论得出，但这里是通过实证方法获得的。参见脚注 4 中的配套代码。

测试录音数据集标签方面的效果。这里有两种可能的看法。我们可以将十种算法视为处理，将整个测试数据集视为实验单位（数据集对每种算法的反应如何？）；也可以将整个测试数据集视为一种处理，将十种算法视为实验单位（每种算法对数据集的反应如何？）。由于表1报告的是每种算法在整个数据集上的评价指标（观测），后一种解释有助于得出有关特定算法有效性的结论。在这种情况下，实验设计很简单：每个实验单位都接受相同的处理。该设置涉及BALLROOM 的划分、提取的特征、随机种子、软件库等。

4 统计结论效度

统计结论效度是指"关于两个变量之间协变的推断效度" [Shadish 等人， 2002]。这包括得出存在协变的结论，也可能包括判断其强度。这是关注统计显著性的层面，也就是说，处理与结果之间观测到的协变不太可能是偶然出现的。Shadish 等人[2002] （第45页）列出了九种对统计结论效度的不同威胁。与计算机实验相关的四种威胁是：关于响应所依据的统计数据的假设遭到违反（以及使用了错误的统计检验，即第三类错误 [Kimball， 1957]）；样本量过小，无法可靠地检测出协变（检验效能不足）；通过尝试多种分析方法和数据筛选，有意寻找显著结果（"p值操纵"，Head 等人[2015]）；以及由于实验单位的异质性，观测值的方差增加。

4.1 音乐信息检索对统计结论效度的关注

我的结果具有统计显著性吗？零假设显著性检验（NHST）量化处理对响应产生的观测效应是否纯属偶然，同时也量化效应的方向和大小。这回答了这样一个问题：结果是否具有统计显著性？Flexer [2006], 讨论了音乐信息检索中的统计检验基础，Cohen [1995], 讨论了人工智能领域的一般统计检验基础，Japkowicz 和 Shah [2011]则讨论了机器学习领域的相关基础。选择要使用的统计检验时必须谨慎；每种检验都有可能被违反的强假设。零假设显著性检验最适用于完全随机化实验设计 [Bailey， 2008], 因为这样可以降低实验单位和处理中的结构干扰响应的可能性（这会导致混杂，下一节将对此进行讨论）。大多数音乐信息检索实验无法采用完全随机化，因为采样点所属的目标总体并不明确（"悲伤"音乐的随机样本是什么？"悲伤"这个词本身就相当含糊），因此，音乐信息检索领域采用零假设显著性检验所能得出的结论也受到限制。[8]

[8]无法完全随机化的实验设计称为*quasi-experimental designs* [2002]，这是 Shadish 等人讨论的另一个重要主题。

观测到的统计显著性对用户有实际意义吗？在音乐信息检索领域，即使发现了统计显著性，这也未必能转化为实际用户与“改进后”的音乐信息检索系统交互时可感知的差异。以音乐信息检索领域的一个例子来说，一项众包用户评估 [Urbano 等人， 2012] 表明，用户对音乐推荐系统的满意度上限约为 80%，因为认同两个系统“同样好”的用户比例最高就是这一数值。此外，针对 MIREX 的音频音乐相似性与检索任务，Urbano 等人 [2012] 表明，算法之间具有统计显著性的差异可能小到对用户没有任何实际影响。

4.2 典型音乐信息检索实验中的统计结论效度

现在让我们考察第 2 节中的典型音乐信息检索实验，并推理我们能从中得出哪些具有统计结论效度的结论。表1清楚地显示，机器学习（ML）对该数据集的每项响应都高于随机方法$unif$、$freq$ 和maj。ML 的任何响应由偶然因素导致的可能性有多大？也就是说，任何一种 ML 方法实际上都不比某种随机方法更好吗？由于我们有$unif$ 和$freq$的经验分布，因此可以估算它们得到例如至少为 0.6 的宏平均召回率的概率： $p < e^{-200}$.[9] 因此，一个有效的统计结论是：我们观察到，在 BALLROOM 的某个特定划分上，使用带有这些特定特征的 ML 与测得的响应之间存在显著协变。等价地说，这些经过训练的 ML 系统在该 BALLROOM 划分上的每项响应，都与随机选择不一致。

上述结论涉及机器学习的使用，而人们也可以考虑与机器学习类型相关的统计结论，即高斯建模（线性判别分析和二次判别分析）与最近邻建模（K 近邻）的对比，或线性判别分析与二次判别分析的对比。我们可能希望得出结论：对于 BALLROOM 上的这些特征，高斯建模优于最近邻建模；甚至更确切地说，使用这些特征时，二次判别分析在 BALLROOM 上的表现最佳。这需要考察响应之间的差异，即所谓的对比。虽然没有太多理由怀疑所用的 BALLROOM 70/30 划分对某一种机器学习模型特别有利，但我们的实验并未揭示机器学习所导致的任何响应的分布，因此也未揭示由机器学习导致的响应差异在 BALLROOM 随机划分下的分布。我们有两项高斯模型的测量结果，以及五项最近邻模型的测量结果，但不能在不引入实验中未包含的辅助信息的情况下，直接计算并比较这些统计数据

[9]这是模型随机选择标签时，观察到不低于 0.6 的宏平均召回率的概率。具体来说，这是从均值为 0。125、标准差为 0。02 的高斯分布随机变量在域 $[0.6, \infty)$ 中抽取一个数值的概率。请参阅随附的 Jupyter 笔记本。另请注意，0.6 的宏平均召回率是一个悲观的选择。采用表1中的较高数值会得到更低的概率。

，例如机器学习模型的决策边界如何随训练数据的变化而变化。

事实上，如果我们根据表1得出结论，认为在 BALLROOM 的70/30划分中，使用这些特征时，高斯建模优于最近邻建模，那么我们就错了。这属于“第一类错误”，即在实际上不存在显著差异时，得出存在显著差异的结论。对 BALLROOM 进行随机70/30划分并重复这一实验1000次后，我们发现，高斯模型的最佳响应与最近邻模型的最佳响应之差服从高斯分布；对于任一评价指标，观察到差值为零或更小的概率为 $p > 0.41$ （参见配套的 Jupyter 笔记本）。也就是说，更具统计有效性的结论是：随机划分 BALLROOM 时，高斯模型的最佳表现优于最近邻模型最佳表现的情况约占五分之三。有了这些新数据，我们可以看出，表1中的结果不足以支持这样一个统计上有效的结论：对于任意显著性水平 $\alpha < 0.4$，使用这些特征对 BALLROOM 进行建模时，某一种建模方法显著优于或劣于其他任何方法。另一方面，如果我们得出结论，认为在 BALLROOM 的随机70/30划分中，使用这些特征时，二次判别分析优于最近邻建模，那么这个结论就是正确的——但仅限于宏平均召回率。将实验重复1000次后，我们发现，二次判别分析的响应与其他所有模型中的最佳响应之差，其分布使得差值为负的概率仅在召回率这一指标上为 $p < 0.046$ （参见配套的 Jupyter 笔记本）。换言之，在召回率方面，二次判别分析表现不如其他模型的情况不到二十分之一。

总之，从表1中我们能得出的最普遍的统计结论是，我们观察到的机器学习响应与随机选择的模型的响应高度不一致。每个机器学习模型都了解 BALLROOM 的一些信息，知道如何将从音乐录音中计算出的特征与其标签关联起来。由于我们不知道表1中任何响应因划分而产生的变异程度，因此无法对哪种类型的机器学习模型最优，或哪一个具体实现最优，针对这个特定数据集得出任何有效的统计结论。为进一步探究，我们必须多次运行实验，以获得对比值分布。然而，即便如此，我们仍无法说明显著差异的原因。这时，内部效度的概念便显得尤为重要。

5 内部效度

内部效度是指“关于两个变量之间观察到的协变是否具有因果关系的推断效度” [Shadish 等人， 2002]。统计结论效度只关注处理与响应之间协变的强度，而内部效度则聚焦于特定响应由处理引起的原因 。Shadish 等人 [2002] （第 55 页）列出了一张表，其中包含九种

种对因果结论的不同威胁。其中有几种涉及混杂，即在实验中，由于操作化不当，处理与其他因素相互混淆。

举一个具体的例子：假设有一项实验，测量两种不同药物降低血压的效果，但一种药物给年轻患者服用，另一种给老年患者服用。这种实验设计混淆了药物种类与患者年龄，因此无法区分这两个因素各自造成的影响。根据这项实验得出的关于药物效果的任何结论，都缺乏内部效度。

5.1 音乐信息检索中对内部效度的关注

我的数据收集会引入混淆因素吗？收集音乐数据的方法可能会无意中引入某些结构。例如，Sturm[2017] 讨论了 BALLROOM 数据集的构建方式：研究者从一家销售舞厅舞蹈比赛用音乐的网站下载了音乐 CD 片段。舞厅舞蹈比赛受到各类组织的规范，例如世界体育舞蹈联合会（WDSF），[10] 这些组织旨在确保世界各地参赛者所参与的赛事具有一致性。它们对每种舞蹈的速度设定了严格要求，因此舞者需要具备高超的技巧。由此，BALLROOM 数据集中的标签可能表示：1）音乐的节奏；2）舞蹈类型；3）舞蹈对速度的严格要求。

我的数据划分是否引入了混杂因素？数据集划分也可能引入混杂因素，例如“真实标签泄漏”。一个例子是，先将录音切分成较短的时间帧（例如 40 毫秒），再将这些帧划分到训练集和测试集中，从而使高度相关的特征分散到这两个集合中。在基于音频的流派分类中，已有研究表明，如果训练数据和测试数据中都包含来自同一艺术家或专辑的歌曲，性能就会被人为抬高 [Pampalk等人，2005年，Flexer和Schnitzer，2010]。Flexer [2007] 指出，与基于节奏内容（波动模式）等更抽象特征的分类相比，采用非常直接的频谱内容表征进行基于音频的流派分类时，使用艺术家/专辑筛选器会导致性能下降得更多。另一项元研究也证实了这一观点：数据划分问题对音乐信息检索系统的影响可能大不相同，因此会改变性能排名 [Sturm，2014a]。

关于可解释和可诠释的音乐信息检索的研究或许有助于识别混杂因素，例如 [Mishra等人，2018], 但最新结果表明，许多广受欢迎的解释器在面对刻意施加的混杂变换时表现不佳 [Praher等人，2021年，Hoedt等人，2022b]。Sturm 提出了一套通用框架，通过干预来检测和刻画音乐信息检索实验中的混杂效应 [2014b]，Rodr´ıguez-Algarra等人进一步完善了该框架 [2019]。音乐信息检索领域的因果推断还有许多工作有待开展，但这并不容易，因为必须先将“相似性”等核心概念转化为可操作的定义。否则，就无法解开变量之间的关系

[10] https://www.worlddancesport.org/

不可能测量可能产生的影响。

5.2 典型音乐信息检索实验中的内部效度

值得关注的是，我们在第 2 节训练的机器学习模型中，究竟是什么使其响应与随机选择不一致。鉴于线性判别分析和二次判别分析中使用的高斯模型是如何构建的——根据训练数据估计均值和协方差参数——一个内部有效的结论是，这些模型在 BALLROOM 中表现良好，是因为根据训练数据估计的似然分布也很好地拟合了测试数据。同样，鉴于最近邻分类器的工作原理，一个内部有效的结论是，在 BALLROOM 中，训练数据中带标签特征的邻域在很大程度上包含了标签相同的测试数据特征。另一个内部有效的结论是，这些机器学习模型在 BALLROOM 中的高性能，是由特征与模型的表达能力共同作用，使模型捕捉到了与 BALLROOM 中标签相关的信息所致。我们的机器学习模型从 BALLROOM 中学到了某些东西，因此其性能显著优于随机选择。

结合音乐信息检索研究的目标，我们希望得出更具体的结论，例如，我们的机器学习模型学会了识别 BALLROOM 中的节奏。这当然是对系统性能的一种合理解释，但它是唯一的解释吗？这一结论的内部效度取决于一个关键假设：推断出 BALLROOM 的正确标签只能是学习区分并识别 BALLROOM 中各种节奏的结果。换言之，我们必须假设，除了感知节奏之外，没有其他方法能够准确推断 BALLROOM 中的标签。

考虑到音乐录音中存在的众多变量，不难想象还有其他方法可以推断 BALLROOM 中的标签。人们可以听出各类别独有的乐器编制，例如，许多标记为“探戈”的录音中都有弦乐器、钢琴和手风琴。许多标记为“华尔兹”和“维也纳华尔兹”的录音中都有弦乐器。许多标记为“快步舞”和“吉特巴”的录音中都有铜管乐器、钢琴、人声和鼓。因此，感知乐器编制是比随机选择更有效地推断 BALLROOM 标签的一种方法。我们知道，所用特征的时域特性对音色并不敏感，因此无法清楚地表达乐器的声音特征，故可以排除这种解释。那么，除了节奏之外，是否还有其他与 BALLROOM 标签密切相关的时域变量？

一种可能的变量是速度。如果在 BALL- ROOM 中，速度与节奏相关，那么速度估计就是机器学习模型再现真实标签的另一种方式。通过聆听标注为“华尔兹”“快步舞”或“恰恰恰”的 BALLROOM 录音，可以听出它们具有不同的节奏和不同的速度，但仅凭速度就可以将它们区分开来。速度和节奏是相关的音乐特征，但并不是同一回事 [Sethares, 2007]。

让我们进行一项实验 验实验，以测试我们训练的机器学习模型的敏感度 L 模型

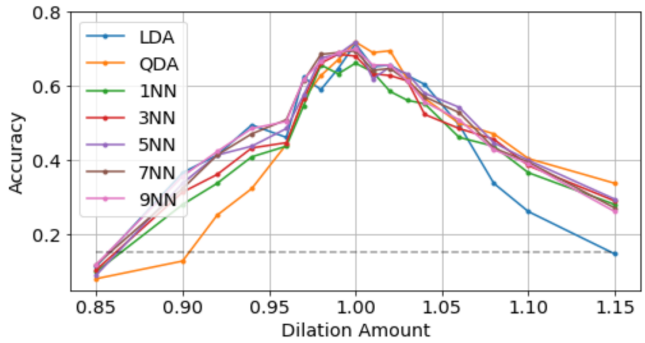

图1：表1中机器学习模型在经过保音高时间伸缩处理的测试录音上的准确率。灰色水平虚线表示表现最佳的随机系统的平均准确率加上两倍标准差。

对速度的敏感度。我们进行了一项干预，将所有测试录音进行一定程度的保音高时间伸缩，然后测量模型对这些新“处理”的响应。将录音以 1.1 的倍数伸缩，会使其时长增加 10%——或者说，使录音中的音乐速度减慢 10%。将录音以 0.9 的倍数伸缩，会使其时长减少 10%——或者说，使录音中的音乐速度加快 10%。图1展示了表1中的所有机器学习模型的准确率如何随伸缩程度而变化。我们的干预清楚地揭示了机器学习模型在多大程度上依赖测试数据中的速度；考虑到上一小节中关于 BALLROOM 的背景信息，这并不令人意外。

要得出我们的机器学习模型的响应是由其识别 BALLROOM 中的节奏所导致这一结论，其效度取决于对 BALLROOM 过于强烈的假设，即在 BALLROOM 中推断标签的唯一方式是节奏识别。实验设计未考虑数据集中存在的结构；我们没有对推断 BALLROOM 标签的其他方式进行对照，而数据集的构建方式保证了这些方式必然存在。因此，仅凭表1和我们的实验设计，我们无法对因果推断作出更具体的结论，只能说：我们的机器学习模型的响应是由它们学会了关于 BALLROOM 的某些东西 所导致的。充其量，它们至少依赖于速度和节奏。这也让人质疑，将预测与 BALLROOM 中的真实标注进行比较，究竟与我们实际可能想要测量的能力——即节奏识别——有何关联。此时，构念效度的概念便与此相关。

6 构念效度

构念效度是“关于代表抽样细节的高阶构念所作推断的效度” [Shadish 等人， 2002]。这涉及实验研究者打算从实验中推断的内容与实际测量的内容之间的关-

系，也就是实验研究者意图的操作化。例如，直接测量一个人的血压，需要通过侵入性操作，将测量设备插入其静脉。也可以采用侵入性较低但间接的方法测量血压：从外部向静脉施加已知压力，并监听血流何时停止。了解封闭系统中液体不可压缩这一特性后，便可知测量气球中的压力与测量血压有关。Shadish 等人 [2002] （第 73 页）列出了十四种对构念效度的不同威胁，但其中有几种与计算机实验无关。主要威胁在于测量内容与预期测量内容之间的关系值得质疑。仅因方便而非相关性选择测量指标、从便利人群中抽样，以及未界定预期测量的内容，都是对构念效度的威胁。构念效度不仅涉及如何测量某事物，也涉及测量什么以及在何种实验环境下进行测量。

6.1 音乐信息检索对构念效度的关注

带标签的音乐数据集中，分类准确率或任何评价指标与X有何关联？音乐信息检索中的两个例子是：将“音乐流派”分类准确率用作衡量音乐相似性的间接指标 [Pohle 等人， 2008], 或用作信息检索用户满意度的指标（例如，参见 [Schedl 等人， 2013]的讨论）。这些指标之间的关系十分微妙，尤其考虑到准确率本身并不能可靠地衡量系统是否学到了与音乐相关的内容 [Sturm, 2013年，2014b]。在这方面，一个重要的参考案例是 Pfungst[1911] 对一系列实验的描述：这些实验试图可靠地测量一匹只能用敲击来作答的马的算术能力。无论提出多少问题，只计算马敲出的正确答案数量都无济于事，因为还必须考虑每个问题是如何提出的（即实验设置）。Pfungst 发现这匹马表面上具备算术能力的原因，关键在于改变实验设置：问题保持不变，也测量正确回答的准确率，但改变了提问方式，以便对实验中的不同因素进行控制。音乐信息检索也是如此。

待测试系统的“用例”是什么？为应对构念效度方面的威胁，音乐信息检索实验研究者必须尽可能明确地界定所构建和测试系统的用例。Sturm 等人 [2014], 尝试为音乐描述界定用例，并强调了定义成功标准的必要性。实验研究者必须确定其测量方法与成功标准之间的关系，例如，将 BALLROOM 中的分类准确率与特定用户的满意度联系起来。

我们如何检验结论的构念效度？一种可能的方法是评估不同实验的结果，这些实验旨在测量相同的高阶构念。音乐信息检索领域的一个例子是，研究不同音乐流派分类器在接收相同输入时的相关性 [Liem 和 Mostert, 2020]。

分类器之间的低相关性表明构念效度存在问题。另一个相关主题是对抗样本，它使人们对这样一种结论产生怀疑：音乐信息检索系统在某个数据集上的高准确率，反映了它对波形中音乐的“感知”。对抗样本最早在图像分析领域被描述 [Szegedy et al., 2014], ，其中对输入数据施加不可感知扰动会显著降低分类准确率。

对于音乐流派分类系统，Sturm [2014b] 使用与音乐信号无关的音频过滤变换，将分类准确率降低或提高到不超过随机水平，或达到完美的100%。这种无关性通过听音测试和真实的人类受试者加以确认：变换虽然可听见，但不会改变人们对某一特定音乐流派的清晰印象。关于音乐情感分类，也记录到了同样的问题行为 [Sturm, 2014b]。针对深度音乐嵌入，有人提出了一种基于不可感知音频变换的测试 [Kim等，2019], 本质上是在输入音频空间及其对应的深度潜在空间中验证距离一致性。

继这些试图将预测改为任意目标的所谓非定向攻击之后，人们又研究了旨在将预测改为特定类别的定向攻击。已有研究报告了一种针对音乐流派识别的定向攻击 [Kereliuk等， 2015], 其中，将从音频计算得到的幅度谱帧视为图像，并采用图像目标识别中的方法发起攻击。针对音乐乐器分类，也有人提出了一种定向攻击，允许直接向音频波形而非谱图中添加扰动 [Prinz等， 2021]。与 [Kereliuk等人的结果非常相似， 2015], 这些攻击能够将准确率降低到接近随机基线的水平，并造成向任意指定乐器的错误分类。除了一些高频偏差外，信号扰动几乎不可察觉。作者还通过攻击现实世界中的音乐推荐系统，人为提高了播放次数，从而证明此类攻击可能构成音乐信息检索中的安全问题。后续工作提出了针对这类恶意攻击的防御措施 [Hoedt等，2022a]。

6.2 典型音乐信息检索实验中的构念效度

谈到第 2 节中的典型音乐信息检索实验，我们关注的是围绕所测量的节奏识别潜在能力进行构念推断。例如，一种构念推断是，我们的特征能够衡量录制音乐中与节奏相关的方面。从某种意义上说，根据基本信号处理组件的定义，我们的特征来自与节奏确实相关的时间特征。我们的特征也依赖于声学信息，尤其依赖于谱通量所捕捉到的起音之间存在高对比度差异——因此，它们与由具有尖锐起音的特定类型乐器演奏的节奏的关联有限。不过，我们已经看到，这些特征也能反映速度，而速度并非节奏 [Sethares， 2007], 且速度是机器学习算法得出节奏标签的另一条途径。因此，我们仍需质疑

我们的特征与我们试图操作化的概念（即节奏）之间的关系。

BALLROOM 数据集旨在反映不同的舞厅舞蹈节奏，其构建方式与使用该数据集所作构念推断的效度密切相关。“华尔兹”片段如何体现“华尔兹”节奏？是否只有一种“华尔兹”节奏？在 BALLROOM 中，华尔兹实际上有两个不同的标签：“华尔兹”和“维也纳华尔兹”。根据世界体育舞蹈联合会的规定，两者的区别部分取决于速度：维也纳华尔兹的表演速度应为每分钟174至180拍 Federation [2014]。

如果不改变实验设置以对照其他因素，让系统为 BALLROOM 数据集的任意划分打标签，就无法可靠地衡量系统识别节奏的能力。这并非只需选择不同的特征、度量指标、交叉验证方法或采用某种统计检验那么简单。必须改变实验本身，使节奏识别 成为实际测量的对象。这意味着 BALLROOM 仍可用于衡量机器学习系统的节奏识别能力。事实上，在上一节中，我们用它反驳了这样的因果主张：表1中机器学习系统的良好表现是由其识别节奏的能力所致。那么，系统在 BALLROOM 上的表现也可能表明其在其他专注于节奏的数据集上的表现吗？这就涉及外部效度的概念。

7 外部效度

外部效度是“关于因果关系在实验单位、实验环境、处理变量和测量变量发生变化时能否成立的推断效度” [Shadish 等人， 2002]。更广义地说，外部效度是从实验中得出的概括性因果推断是否真实。例如，推断一种被发现能降低德国患者血压的药物，也能降低墨西哥居民的血压——由于饮食、生活和工作条件等方面存在差异，这一结论可能缺乏效度。另一个例子是，在研究人群中增加药物剂量会进一步降低血压。如果我们从实验中得出的因果推断缺乏内部效度，那么将这一推断推广到未经检验的变化情形时，就不具备外部效度。Shadish 等人 [2002] （第 87 页）列出了一张表，其中包含五种不同的外部效度威胁，这些威胁有别于内部效度威胁。主要威胁在于，实验组成部分的变化可能会破坏实验中成立的因果推断。例如，某种药物可能对受试的疾病类型有效，但这种疾病类型未必能推广到其他密切相关的疾病。

7.1 音乐信息检索对外部效度的关注

我的模型能否推广到样本外数据？评估音乐信息检索分类系统的标准方法，是在交叉验证实验中使用独立的训练集和测试集，以获得看似无偏的性能估计。然而，如果将此类音乐信息检索系统应用于独立的样本外数据，通常会观察到性能严重下降。一个例子是音乐流派识别实验：当在训练集之外的不同数据集上进行评估时，准确率结果并不成立 [Bogdanov 等人，2016，2019]。这些结果无法推广到不同数据集中被认为相同的音乐流派标签，这反映出外部效度不足。不同数据集的不同标注者对诸如“摇滚”之类的音乐流派标签会有不同用法——这构成了构念效度威胁。另一个例子是，不同的音频编码会如何影响后续描述符计算和分类结果 [Urbano 等人， 2014], ；又或者，软件实现之间的差异通常会如何降低可复现性 [McFee 等人， 2018]。

不同评分者对真实标注的看法是否一致？人类对音乐的感知具有很强的主观性，因此评定者间一致性可能较低。因此，如果要求多名人类受试者根据感知相似度对同一组歌曲对进行评分，那么预期只能达到一定程度的一致性，这取决于多种主观因素 [Schedl 等人，2013年，Flexer 和 Grill，2016] 例如个人品味、聆听经历、对音乐的熟悉程度、当前心情等。关于音乐标注，Seyerlehner 等人 [2010]指出，人类将歌曲分为 19 种音乐流派时，准确率介于较低的 26% 至 71% 之间。对日常音乐术语进行基于音频的语义锚定，也呈现出同样的问题 [Aucouturier， 2009]。甚至有人认为 [Wiggins， 2009]在音乐语境中并不存在一成不变的“根基”，因为音乐本身具有主观性、高度依赖语境且不断变化。

评定者间一致性缺失会造成外部效度问题，因为从实验中得出的推断无法从参与实验的用户或标注者推广到任意用户/标注者这一预期目标总体。这也会造成信度问题，因为持有不同主观意见的不同用户组或标注者组会妨碍实验结果的可重复性。评定者间一致性缺失为音乐信息检索方法设定了上限，因为计算模型超越人类一致性水平并无意义。Jones 等人，2007年、Schedl 等人，2013年、Flexer 和 Grill 报告了此类上限，涉及 MIREX 的“音频音乐相似性与检索”（AMS）和“音乐结构分段”（MSS）任务。AMS 的上限在 2009年 已经达到，而 MSS 的上限至少对某些音乐流派而言已触手可及。音乐结构分析方面也有类似结果 [Nieto 等人， 2014]和弦估计方面亦然 [Ni 等人，2013年、Koops 等人， 2019]。

评分者在不同时间点对自己的判断是否一致？除了探讨不同标注者是否对真实标注达成一致，我们还可以考察同一个人在不同时间点面对相同标注任务时的一致程度。较高的评分者内部一致性有助于解决音乐信息检索系统中的上限问题，

因为这将使模型个性化变得有意义，也就是说，为不同的个体分别建立模型。然而，至少就一般音乐相似性任务而言，研究表明，评分者内部一致性仅略高于评分者之间的一致性 [Flexer 等人， 2021], ；其绝对水平还取决于音乐素材以及评分者在测试时的心情。通过深度学习为个别标注者个性化和弦标签的方法则取得了更好的效果[Koops 等人， 2020]。

回到无关变换的影响 [Sturm，2014b，Kere- liuk 等人，2015，Rodr´ ıguez-Algarra 等人， 2016]以及对抗样本的存在 [Prinz 等人， 2021], 我们可以问：我的模型能否泛化到这类攻击？更具建设性地，我们可以寻找减轻这些攻击影响的方法，从而可能提高模型的泛化能力 [Hoedt 等，2022a]。

7.2 典型音乐信息检索实验中的外部效度

考虑第 2 节中的典型音乐信息检索实验，我们不能有效地得出结论，认为我们的任何模型都能识别一般意义上的节奏，因为我们不知道它们是否识别了 BALLROOM 中的节奏。我们在第 5.2 节进行的伸缩干预实验表明，所有模型都失去了其所谓的识别 BALLROOM 中节奏的能力，因此没有理由推断它们能在其他地方识别节奏。我们或许可以做出这样一个因果结论：我们所有的模型在 BALLROOM 上表现良好，是因为它们学到了一些关于 BALLROOM 的信息——BALLROOM 是一个精选录音集，这些录音于 2004 年 从某个特定网站下载。我们的模型是否也学到了关于同一网站上其他录音的信息，而这些录音是在多年后收集的?

扩展 BALLROOM 数据集（X-BALLROOM） [Marchand 和 Peeters ，2016] 由 3,484 段音频录音组成，涵盖与 BALLROOM 相同的八种舞蹈风格或音乐节奏，但这些录音是十多年后从同一网站下载的。这让我们有机会检验我们的假设。使用 BALLROOM 数据集训练的模型应用于全部 X-BALLROOM 数据后测得的评价指标如表2所示。我们仍然看到响应与使用基于我们特征的机器学习之间存在显著协变。总的来说，我们的机器学习模型从 BALLROOM 中学到的无论是什么概念，都能迁移到 X-BALLROOM；但我们仍然不知道这些概念是否与节奏有关。

我们可以借此机会检验表1中关于高斯建模优于最近邻建模这一结论的普适性。使用 X-BALLROOM 的 70/30 随机划分训练和测试相同的机器学习模型，得到表3中的结果。我们仍然看到二次判别分析表现最佳，但最近邻模型的性能有了大幅提升。应当调查这种性能变化的原因，并将其与已知存在于 BALLROOM 中的混杂联系起来。表2中的结果是否表明 X-BALLROOM 中也存在混杂？这表明，在典型的音乐信息检索实验中所作的观测，是开启各种探索途径的起点

	准确率	精确率	召回率	F1分数
LDA	0.659	0.647	0.643	0.643
QDA	0.682	0.678	0.672	0.673
1NN	0.622	0.616	0.602	0.604
3NN	0.636	0.629	0.610	0.613
5NN	0.644	0.643	0.617	0.619
7NN	0.647	0.646	0.619	0.621
9NN	0.645	0.643	0.615	0.618
unif	0.12 ± 0.01	0.13 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
freq	0.13 ± 0.01	0.13 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
maj	0.13	0.02	0.12	0.03

表1所示，在 BALLROOM 上训练并在整个 X-BALLROOM 上测试的模型 [Marchand 和 Peeters， 2016]。

Mod	准确率	精确率	召回率	F1分数
LDA	0.726	0.721	0.722	0.717
QDA	0.763	0.764	0.773	0.758
1NN	0.702	0.691	0.693	0.690
3NN	0.741	0.731	0.727	0.727
5NN	0.750	0.744	0.736	0.738
7NN	0.753	0.746	0.736	0.738
9NN	0.751	0.746	0.734	0.737
unif	0.12 ± 0.01	0.13 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
freq	0.13 ± 0.01	0.13 ± 0.01	0.13 ± 0.01	0.12 ± 0.01
maj	0.13	0.02	0.12	0.03

表3：与表1相同，但模型在 X-BALLROOM [Marc- Hand 和 Peeters, 2016] 中进行训练和测试。

的实验途径 定位、假设来源，以及不 t he 最终结果。

8 结论

本文基于 Shadish 等人 [2002]提出的类型学，对效度这一概念进行了综述。本文汇集了音乐信息检索领域中少数提及效度相关内容的资料，以及若干虽未直接提及效度但与之相关的资料。本文无意规定应如何设计和开展实验，以便能够从中得出有效的结论。相反，本文旨在阐明效度的含义、其重要性，以及效度可能受到何种威胁，并将这些内容纳入音乐信息检索的研究范畴。

在音乐信息检索领域，占主导地位的实验方法论采用的是 Cranfield 范式：在数据集的一个划分上训练模型，再统计它在另一个划分上的正确答案数量。这类实验利用手头现成的数据，实施成本低廉，而且能提供可供比较的数字，让同行评审相信研究取得了进展 [Hand, 2006]。尽管有人多次呼吁改变现状 [Peeters 等人, 2012, Schedl 等人, 2013] ，也有人再三恳请 [Urbano 等人, 2012, 2013, Stur m, 2013, 2014b, Flexer 和 Grill, 2016, Flexer 等人, 2021], ，这种实验方法

仍是该领域的标准做法，其严重缺陷却遭到忽视。从这类实验得出的任何超出“系统从这个数据集中学到了一些东西”的推断，都缺乏内部效度、构念效度和外部效度。这并不意味着所有这类推断都是错误的，只是说，这些推断并非实验按其设计和实施方式所能得出的结论。复现数据集的真实标注只是起点，之后还必须探寻观察到的行为背后的原因。我们必须抵制这种冲动，不要断定系统必然在做我们所希望它做的事。

Shadish 等人 [2002] 为音乐信息检索提供了一个成熟的起点，但效度还有其他类型。例如，Lund [2021] 修订了 Shadish 等人 [2002] 的类型学，以处理原因与处理之间的歧义，更好地界定实验环境的各个方面，并确立五种效度的层级顺序：统计结论效度、因果效度、构念效度、泛化效度和理论效度。该类型学与 Shadish 等人 [2002] 的类型学之间的一个重要区别，在于它强调基础研究的一个主要目标：对理论作出贡献。其他类型的效度还包括生态效度、聚合效度和效标效度[Urbano， 2011]；但这些效度仍然涉及如何从以某种方式收集的证据中得出结论。

最后需要指出的是，作为一名工程师，在阅读 Shadish 等人 [2002] 的著作时，一个令人沮丧之处在于：全书 623 页中，至少包含一个方程的页面只有五页。相反，Shadish 等人 [2002] 描述了实验，以及每种效度如何体现在所得结论中，并指出了对这些结论的推理构成具体威胁的因素。实验更不用说实验研究者了，都是如此复杂的组合体，以至于用形式化方式表达它们，并据此计算出与每种效度相关的数值，可能几乎没有实际适用性，而且即使能够实现，也只有有限的读者群能够理解。效度的语言就是推理，我们希望本手稿能够激励音乐信息研究人员创造性地思考他们所观察到的现象，以发现其原因。

致谢

感谢 Juli´an Urbano 和 Hugo Maruri-Aguilar 在本文撰写过程中进行的有益讨论，也感谢《音乐信息检索学会汇刊》审稿人提出的建设性批评。Sturm 的贡献得到一项研究项目的支持，该项目获得了欧盟地平线2020研究与创新计划下欧洲研究委员会（ERC）的资助（资助协议编号：864189，MUSAiC：人工创造力与批评前沿的音乐）。Flexer 的贡献得到奥地利科学基金（FWF，项目编号 P 31988）的资助。为实现开放获取，作者已对因本次投稿而产生的任何经作者接受的稿件版本采用知识共享署名许可（CC BY）公共版权许可。

参考文献

J.-J. Aucouturier。《听起来像青少年精神：日常音乐术语基础的计算洞见》。语言、进化与大脑，第 35–64 页，2009 年。J.-J. Aucouturier 和 F. Pachet。《改进音色相似度：天空有多高？》《言语与音频科学负结果》，1(1):1–13，2004年。R. A. Bailey。比较实验的设计。剑桥大学出版社，2008年。D. Bogdanov、A. Porter、H. Boyer、X. Serra 等。《音乐分类任务的跨集合评估》。载于第*17*届国际音乐信息检索学会会议论文集，第 379–385 页，2016。D. Bogdanov、A. Porter、H. Schreiber、J. Urbano 和 S. Oramas。《AcousticBrainz 音乐流派数据集：多源、多层级、多标签、大规模》。载于第*20*届国际音乐信息检索学会会议论文集，2019年。A. Chase。《鲤鱼（Cyprinus carpio）的音乐辨别能力》。《style id='11'>《动物学习与行为》，29(4):336–353，2001。C. W. Cleverdon。《克兰菲尔德测试对索引语言的重要意义》。载于第*14*届 *ACM SIGIR* 信息检索研究与发展年度国际会议论文集，第 3–12 页，1991。P. R. Cohen。人工智能的实证方法，第139卷。麻省理工学院出版社，马萨诸塞州剑桥，1995。S. Dixon、F. Gouyon 和 G. Widmer。《通过节奏模式表征音乐》。载于第*5*届国际音乐信息检索学会会议论文集，第 509–517 页，2004年。J. S. Downie。《音乐信息检索评估交流（MIREX）》。《style id='19'>《D-Lib 杂志》，12(12)，2006。B. Duggan 和 B. O’Shea。《TunePal：搜索传统音乐乐谱数字图书馆》。《style id='21'>《OCLC 系统与服务》，27:284–297，2011。W. S. D. 联合会。*WSDF* 竞赛规则。罗马尼亚布加勒斯特，2014。A. Flexer。《音乐信息检索实验的统计评估》。《style id='25'>《新音乐研究期刊》，35(2):113–120，2006。A. Flexer。《深入探究艺人过滤器对音乐流派分类的作用》。载于第*8*届音乐信息检索国际会议论文集，第 341–344 页，2007。

A. Flexer 和 T. Grill。《音乐相似性建模中评定者间一致性有限的问题》。《新音乐研究期刊》，45(3):239–251，2016。A. Flexer 和 D. Schnitzer。《在超大型音乐数据库中计算音频相似性时专辑与艺人过滤器的影响》。《计算机音乐期刊》，34(3):20–28，2010。A. Flexer、T. Lallai 和 K. Raˇsl。《音乐推荐中评定者间及评定者内一致性的评估》。国际音乐信息检索学会汇刊，4(1):182–194，2021。D. J. Hand。《分类器技术与进步假象》。《统计科学》，21(1):1–15，2006。M. L. Head、L. Holman、R. Lanfear、A. T. Kahn 和 M. D. Jennions。《科学研究中p值操纵的程度及其后果》。《*PLOS*生物学》，13(3):1–15，2015。K. Hoedt、A. Flexer 和 G. Widmer。《防御音乐推荐系统免受基于枢纽性的对抗攻击》。载于第*19*届声音与音乐计算会议论文集，第385–390页，2022a。K. Hoedt、V. Praher、A. Flexer 和 G. Widmer。《构造对抗样本以考察深度音频与图像分类器中解释的可信度》。《神经计算与应用》，2022b。N. Japkowicz 和 M. Shah。《评估学习算法：分类视角》。剑桥大学出版社，纽约，纽约州，美国，2011。M. C. Jones、J. S. Downie 和 A. F. Ehmann。《人类相似性判断：对正式评估设计的启示》。载于第*8*届国际音乐信息检索学会会议论文集，第539–542页，2007。C. Kereliuk、B. L. Sturm 和 J. Larsen。《深度学习与音乐对抗样本》。《*IEEE*多媒体汇刊》，17(11):2059–2071，2015。J. Kim、J. Urbano、C. Liem 和 A. Hanjalic。《近邻是亲属吗？深度音乐嵌入测试》。《应用数学与统计学前沿》，第53页，2019。A. W. Kimball。《统计咨询中的第三类错误》。《美国统计学会杂志》，52(278):133–142，1957年6月。H. V. Koops、W. B. De Haas、J. A. Burgoyne、J. Bransen、A. Kent-Muller 和 A. Volk。《流行音乐和声标注中的标注者主观性》。《新音乐研究期刊》，48(3):232–252，2019。H. V. Koops、W. B. de Haas、J. Bransen 和 A. Volk。《通过深度学习共享和声区间轮廓实现自动和弦标签个性化》。《神经计算与应用》，32(4):929–939，2020。

C. C. Liem 和 C. Mostert。《感觉可靠吗？开放数据揭示高层音乐描述符的意外行为》。载于第*21*届国际音乐信息检索学会会议论文集，第240–247页，2020。T. Lund。《坎贝尔效度体系修订》。《斯堪的纳维亚教育研究杂志》，65(3):523–535，2021。U. Marchand 和 G. Peeters。《利用听觉统计实现尺度与平移不变的时频表示：在节奏描述中的应用》。载于*IEEE*信号处理机器学习国际研讨会，2016。B. McFee、C. Raffel、D. Liang、D. P. W. Ellis、M. McVicar、E. Battenberg 和 O. Nieto。《librosa：用于音频与音乐信号分析的 Python 库》。载于*Python*科学会议论文集，第18–25页，2015。B. McFee、J. W. Kim、M. Cartwright、J. Salamon、R. M. Bittner 和 J. P. Bello。《音乐信号处理研究的开源实践：关于透明、可持续且可复现的音频研究的建议》。《*IEEE*信号处理杂志》，36(1):128–137，2018。S. Mishra、B. L. Sturm 和 S. Dixon。《通过特征反演理解深度机器聆听模型》。载于第*19*届国际音乐信息检索学会会议论文集，第755–762页，2018。M. M¨uller。《音乐处理基础：音频、分析、算法与应用》。施普林格，2015。Y. Ni、M. McVicar、R. Santos-Rodriguez 和 T. De Bie。《理解主观性对和弦估计准确率测量的影响》。《*IEEE*音频、语音与语言处理汇刊》，21(12):2607–2615，2013。O. Nieto、M. M. Farbood、T. Jehan 和 J. P. Bello。《用于评估音乐段落边界的F值的感知分析》。载于第*15*届国际音乐信息检索学会会议论文集，第265–270页，2014。E. Pampalk、A. Flexer、G. Widmer 等。《基于音频的音乐相似性和音乐流派分类的改进》。载于第*6*届国际音乐信息检索学会会议论文集，第634–637页，2005。G. Peeters、J. Urbano 和 G. J. F. Jones。《ISMIR 2012音乐信息检索评估专题迟到报告会记录》。载于第*13*届国际音乐信息检索学会会议论文集，2012。O. Pfungst。《聪明的汉斯（冯·奥斯滕先生的马）：对实验动物与人类心理学的贡献》。亨利·霍尔特，纽约，1911。

T. Pohle、E. Pampalk 和 G. Widmer。用于将音乐分类为感知类别的常用音频特征评估。载于国际内容型多媒体索引研讨会，2008年。V. Praher、K. Prinz、A. Flexer 和 G. Widmer。音频分类中局部模型无关解释的真实性：基于对抗样本的针对性研究。载于第*22*届国际音乐信息检索学会会议论文集，第531–538页，2021年。K. Prinz、A. Flexer 和 G. Widmer。音乐信息检索中的端到端白盒对抗攻击。国际音乐信息检索学会汇刊，4(1):93–104，2021年。F. Rodr´ıguez-Algarra、B. L. Sturm 和 H. Maruri-Aguilar。分析基于散射的音乐内容分析系统：音乐在哪里？载于第*17*届国际音乐信息检索学会会议论文集，第344–350页，2016。F. Rodr´ıguez-Algarra、B. Sturm 和 S. Dixon。通过干预分析音乐分类实验中的混杂效应。国际音乐信息检索学会汇刊，第52–66页，2019年。M. Schedl、A. Flexer 和 J. Urbano。音乐信息检索研究中被忽视的用户。《智能信息系统期刊》，41(3):523–539，2013年。W. A. Sethares。《节奏与变换》。施普林格，2007年。K. Seyerlehner、G. Widmer 和 P. Knees。人工、自动和协作式音乐流派分类的比较，以及以用户为中心的流派分类系统评估。载于自适应多媒体检索国际研讨会，第118–131页。施普林格，2010年。W. R. Shadish、T. D. Cook 和 D. T. Campbell。广义因果推断的实验与准实验设计。波士顿：Houghton Mifflin，2002年。B. L. Sturm。分类准确率还不够。《智能信息系统期刊》，41(3):371–406，2013年。B. L. Sturm。最新技术水平问世十年后的最新技术水平：音乐信息检索的未来研究。《新音乐研究期刊》，43 (2):147–172，2014a。B. L. Sturm。一种简单的方法，用于判断音乐信息检索系统是否是一匹“马”。《*IEEE*多媒体汇刊》，16(6):1636–1644，2014b。

B. L. Sturm。“马”的内在本质：探寻音乐内容分析系统行为背后的原因。《娱乐计算》（*CIE*），14(2):1–32，2017年。B. L. Sturm 和 N. Collins。kiki-bouba 挑战：面向基于内容的音乐信息检索研发的算法作曲。载于第*15*届国际音乐信息检索学会会议论文集，第21–26页，2014。B. L. Sturm、R. Bardeli、T. Langlois 和 V. Emiya。音乐描述问题的形式化。载于第*15*届国际音乐信息检索学会会议论文集，第89–94页，2014。B. L. Sturm、C. Kereliuk 和 J. Larsen。¿El Caballo Viejo? 深度学习与频谱周期性在拉丁音乐流派识别中的应用。载于数学与音乐计算国际会议论文集，第335–346页，2015年。C. Szegedy、W. Zaremba、I. Sutskever、J. Bruna、D. Erhan、I. J. Goodfellow 和 R. Fergus。神经网络的有趣特性。载于第*2*届国际学习表征会议（*ICLR*）论文集，2014。J. Urbano。信息检索元评估：音乐领域中的挑战与机遇。载于第*12*届国际音乐信息检索学会会议论文集，第609–614页，2011。J. Urbano 和 A. Flexer。音乐信息检索结果的统计分析：原因与方法（摘要）。载于国际音乐信息检索学会会议论文集，第xli–xlii页，2018年。J. Urbano、J. S. Downie、B. Mcfee 和 M. Schedl。统计显著性究竟有多显著? 以音频音乐相似性与检索为例。载于第*13*届国际音乐信息检索学会会议论文集，第181–186页，2012。J. Urbano、M. Schedl 和 X. Serra。音乐信息检索中的评估。《智能信息系统期刊》，41(3):345–369，2013年。J. Urbano、D. Bogdanov、H. Boyer、E. G´omez Guti´errez、X. Serra 等。音频质量对梅尔频率倒谱系数和色度特征鲁棒性的影响是什么？载于第*15*届国际音乐信息检索学会会议论文集，第573–578页，2014。E. M. Voorhees。信息检索评估的哲学。载于跨语言评估论坛论文集，2001。A. Wang。一种工业级音频搜索算法。载于第*4*届国际音乐信息检索学会会议论文集，第7–13页，2003年10月。

G. A. Wiggins。语义鸿沟？？在音乐科学研究中的语义鸿沟！！方法学考量。载于《第十一届 *IEEE* 多媒体国际研讨会论文集》，第 477–482 页。IEEE，2009 年。