

基于梅尔频谱与卷积神经网络的肉鸡叫声小样本识别实现与教学研究

赵永才¹ 李慢¹ 郜静茹¹ 关曼玉¹ 李广博^{1*}

(1.淮北理工学院电子与信息工程学院,安徽 淮北 235000)

摘 要: 肉鸡叫声能够直观反映其采食、应激、求救等生理与行为状态,准确识别肉鸡叫声对规模化养殖中的动物福利监测和异常预警具有重要意义。然而,肉鸡叫声标注样本获取困难,现有研究多针对单一网络模型展开,缺少在小样本条件下对不同骨干网络及迁移学习策略的系统比较,模型选型缺乏依据。为此,本文构建了包含求救呼叫、愉悦音、短促啾啾、颤音和其他发声共5类、54条原始音频的肉鸡叫声数据集,将音频统一转为128×128维梅尔频谱图作为输入,采用四种典型卷积神经网络,并分别在特征提取式与微调式两种迁移学习策略下进行对比实验,采用分层5折交叉验证评估性能。实验结果表明: AlexNet 特征提取式方案取得最优识别效果,总体准确率为59.54%、宏平均F1值为49.10%,其中求救呼叫的F1值达到0.94;轻量级网络 MobileNet-V2 在微调式策略下性能提升显著,而结构较复杂的网络在小样本下采用冻结骨干的特征提取式训练更为稳定。研究表明,在小样本肉鸡叫声识别任务中,骨干网络与迁移学习策略应根据模型复杂度协同选择,该结论可为面向养殖场景的禽声自动识别系统提供方法选型参考。

关键词: 肉鸡叫声识别; 梅尔频谱; 卷积神经网络; 迁移学习; 小样本学习

基金项目: 安徽省教育厅优秀青年教师培育项目(YQYB2026117); 安徽省智慧农业技术与装备重点实验室开放基金(AEC2026016; AEC2025009); 安徽省教育厅自然科学基金项目(重大)(2024AH040217); 省级质量工程-教学研究项目(2023jyxm1009; 2023xjzlt117; 2023sdxx145); 基于人工智能下的家禽康养与疾病预警技术研究(S202413620055); 基于自监督学习与对抗增强的小样本信号的智能感知研究(2025SK019); 基于深度学习的生猪识别与康养研究(S202513620083)

DOI: doi.org/10.65436/hssj.v1i9.64

肉鸡叫声是反映其采食、饮水、打斗、应激和求救等生理与行为状态的重要信号,及时、准确地识别这些叫声有助于养殖人员及早发现鸡群的健康异常与环境应激,对提升规模化养殖效率和保障动物福利具有重要意义。传统的肉鸡状态监测主要依赖养殖人员现场听辨和肉眼观察,不仅劳动强度大、主观性强,而且受人员经验和工作时长限制,难以实现连续、客观、可复现的监测^[1]。近年来,随着深度学习和音频识别技术的发展,利用鸡舍采集的音频自动识别肉鸡发声类型,成为智慧养殖中替代人工听辨、实现规模化连续监测的有效途径^[2]。

基于深度学习的禽声识别通常将音频转换为梅尔频谱图后输入卷积神经网络进行分类,已有研究在鸡、猪等畜禽发声识别上取得了较好效果^[3]。然而,这类研究大多围绕单一网络模型展开,且多在样本相对充足的条件下验证^[4]; 在实际养殖场景中,肉鸡各类叫声的标注样本往往十分有限,如何选择合适的卷积骨干网络、以及采用何种迁移学习策略,尚缺乏系统的对比与分析,直接套用现有结论容易出现选型不当、小样本下精度不稳定等问题^[5]。

针对上述问题,本文以自建的5类、共54条肉鸡原始音频数据集为研究对象,将音频统一预处理为128×128维梅尔频谱图,并引入时间掩码与频率掩码进行数据增强; 在此基础上,选取 MobileNet-V2、EfficientNet-B0、

作者简介: 赵永才(2007—),男,安徽蚌埠人,本科生。研究方向:人工智能;

李慢(2005—),女,安徽池州人,本科生。研究方向:人工智能;

郜静茹(2005—),女,安徽阜阳人,本科生。研究方向:人工智能;

关曼玉(1998—),女,安徽淮北,硕士,助教。研究方向:人工智能与大数据;

李广博(1997—),男,安徽淮北人,讲师,博士。研究方向:人工智能、智能信号感知、模式识别、计算机视觉等。

通讯作者: 李广博

ResNet-18 和 AlexNet 四种典型卷积神经网络,分别在特征提取式与微调式两种迁移学习策略下开展对比实验,采用分层 5 折交叉验证,从总体准确率、各类别精确率、召回率、F1 值和宏平均 F1 等多个维度系统评估不同方案的识别性能。

针对上述问题,本文以自建的 5 类、共 54 条肉鸡原始音频数据集为研究对象,将音频统一预处理为 128×128 维梅尔频谱图,并引入时间掩码与频率掩码进行数据增强;在此基础上,选取 MobileNet-V2、EfficientNet-B0、ResNet-18 和 AlexNet 四种典型卷积神经网络,分别在特征提取式与微调式两种迁移学习策略下开展对比实验,采用分层 5 折交叉验证,从总体准确率、各类别精确率、召回率、F1 值和宏平均 F1 等多个维度系统评估不同方案的识别性能。

本文通过以下两方面进行展开:

第一方面为方法理论。本文方法理论部分按“总体框架—数据制备—数据增强—网络与迁移策略”逐层展开。首先给出基于迁移学习的肉鸡叫声识别总体框架,按预训练、微调与推理三个阶段,阐述从 ImageNet 预训练初始化、目标域数据处理到五类发声自动输出的完整流程;随后介绍自建肉鸡叫声数据集的采集与预处理,包括求救呼叫、愉悦音、短促啾啾、颤音和其他发声共 5 类的划分依据与样本规模,以及重采样、预加重、汉明窗分帧、短时能量与过零率双阈值端点检测、0.5s 固定长度切片和对数梅尔频谱转换等操作,并说明按原始音频进行分层 5 折划分、以训练集统计量做 Z-score 归一化的防泄露设计;在此基础上介绍仅对训练集实施的 SpecAugment 数据增强策略,通过时间掩码与频率掩码扩充样本多样性;接着阐述特征提取式与微调式两种迁移学习策略的原理,以及预训练权重在小样本条件下的作用机制;最后以 AlexNet 为例说明卷积运算、Softmax 分类与交叉熵损失的前向传播和反向传播机制。

第二方面为实验验证。本文实验验证部分按“参数设置—定量对比—可视化分析”逐层展开。首先给出实验环境与超参数设置,包括 GPU 与 PyTorch 框架、 $128 \times 128 \times 3$ 输入尺寸、训练轮次、批大小、SGD 优化器、余弦退火学习率策略及早停机制,以及分层 5 折、8:1:1 的数据划分方式;其次在完全统一的实验条件下,将 MobileNet-V2、EfficientNet-B0、ResNet-18 和 AlexNet 四种卷积骨干网络分别在特征提取式与微调式两种迁移学习策略下进行横向对比,从总体识别准确率、宏平均 F1、逐类精确率、召回率与 F1,以及参数数量和计算量等维度,分析不同方案的性能差异与适配规律;最后以最佳模型 AlexNet 特征提取式为例,通过测试集混淆矩阵与高层特征 t-SNE 二维可视化对识别结果进行分析,验证迁移学习策略的选择及本文方法在肉鸡叫声识别任务中的可行性。

一、方法

(一) 总体框架

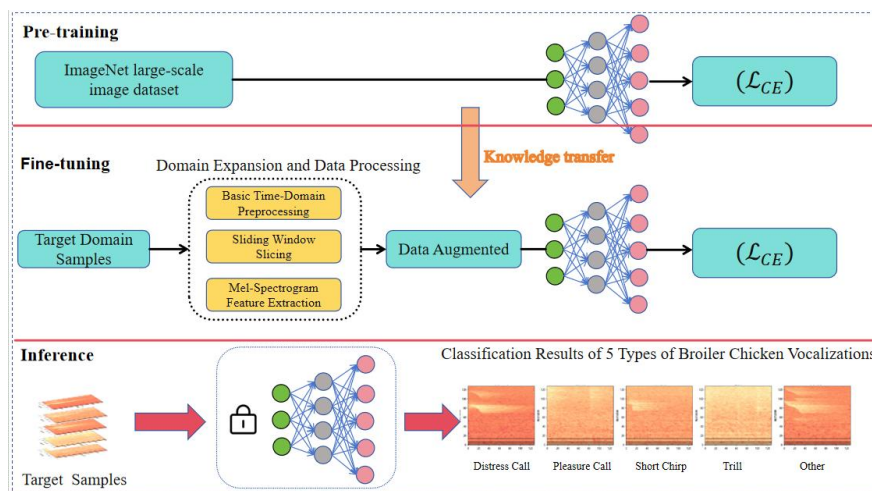


图 1: 基于迁移学习的肉鸡声音信号小样本分类总体框架流程图

本研究采用基于迁移学习的三阶段模型框架,如图 1 所示。第一阶段为预训练阶段:以 ImageNet 大规模图像数据集作为源域数据,在大规模自然图像上完成模型主干网络的初始化训练,通过交叉熵损失 (Cross-Entropy Loss, L_{CE}) 学习通用的视觉特征表征能力,为后续目标域任务提供良好的参数起点。第二阶段为

微调阶段：将预训练模型应用于肉鸡发声分类任务,首先对目标域样本进行域扩展与数据处理,依次经过基础时域预处理、滑动窗口切片与梅尔频谱（Mel-Spectrogram）特征提取三个步骤,将原始音频波形转换为时频谱图输入；同时引入数据增强策略扩充目标域样本规模,再以交叉熵损失在目标域数据上进行微调,使模型参数逐步适配养殖场场景下的肉鸡发声分布。第三阶段为推理阶段：将待识别的目标样本输入训练完成的模型,经过网络前向计算后输出五类发声的分类结果,包括求救呼叫（Distress Call）、愉悦声（Pleasure Call）、短促鸣叫（Short Chirp）、颤音（Trill）及其他发声（Other）,从而实现肉鸡发声类型的自动判别。

（二）数据采集与预处理

2.1 数据集采集

本文所用肉鸡叫声数据采集于商业化密闭式肉鸡养殖鸡舍。采集过程中,录音设备悬挂于鸡群活动区上方,尽量贴近鸡群发声位置,同时避开风机与料线等强噪声源,以保证采集到的叫声信号清晰、可辨。原始音频采用商用录音设备录制,采样率为 48kHz,单声道,量化位数 16bit,单条音频时长约为 2.4~17 秒,共采集标注音频 54 条。采集过程覆盖肉鸡不同日龄和不同活动时段,以增强样本在发声强度和环境背景上的多样性,为后续模型训练与评估提供基础。

2.2 类别标注

表 1 肉鸡叫声数据集类别与样本规模

类别	行为含义	样本数	占比
求救呼叫	鸡群遭遇应激、危险或不适时发出的高声鸣叫	10	18.5%
愉悦音	采食、聚集等舒适状态下发出的平缓叫声	10	18.5%
短促啾啾	日常活动中的短促联系叫声	15	27.8%
颤音	带有明显颤动感的警示或情绪叫声	14	25.9%
其他发声	背景杂声及难以归入上述类别的发声	5	9.3%

数据标注由养殖与音频分析人员人工听辨完成,依据叫声的发声场景与行为含义划分为上述 5 类。其中"求救呼叫"多出现于鸡群受惊吓、踩踏或环境突变时,具有较高的养殖预警价值；"愉悦音""短促啾啾"和"颤音"对应鸡群不同的日常行为状态；"其他发声"为少量难以明确归类的混合样本。各类样本规模如表 1 所示,数据集整体规模较小,属于典型的小样本场景

2.3 音频预处理

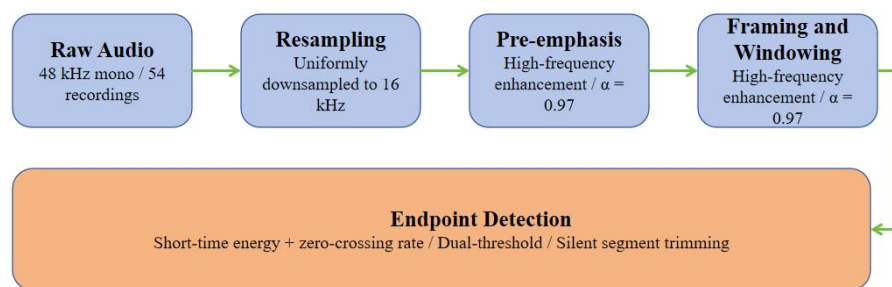


图 2：音频预处理流程图

为消除不同录音条件差异并突出有效叫声信号,本文对原始音频统一进行时域预处理,流程如图 2 所示。首先将 48kHz 原始音频重采样至 16kHz,以降低数据量并统一采样率；随后进行预加重处理,补偿发声器官造成的高频衰减,其计算式为：

$$(y(n)=x(n)-\alpha x(n-1), \quad \alpha=0.97)(1)$$

式中, $x(n)$ 为重采样后信号在第 n 个采样点的幅值, $y(n)$ 为预加重后的输出信号, α 为预加重系数,取 0.97。预加重后,将信号按帧长 25ms、帧移 10ms 进行分帧,并对每帧施加汉明窗以抑制分帧带来的截断效应,汉明窗定义为：

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

式中, $w(n)$ 为汉明窗函数, n 为帧内采样点序号, N 为单帧内的采样点数。最后, 采用短时能量与过零率相结合的双阈值法进行端点检测 (VAD), 计算每帧短时能量 E 与过零率 Z :

$$E = \sum_{n=0}^{N-1} |x(n)|^2, \quad Z = \frac{1}{2} \sum_{n=0}^{N-1} |\operatorname{sgn}(x(n+1)) - \operatorname{sgn}(x(n))| \quad (3)$$

式中, E 反映一帧内信号能量的大小, Z 表示信号在一帧内过零的比例, $\operatorname{sgn}(\cdot)$ 为符号函数, 当输入为正时取 1、为负时取 -1。当某帧的 E 与 Z 同时低于设定阈值时, 判定为静音段并予以裁剪, 仅保留有效发声片段, 从而去除录音中的静默与无关背景噪声, 为后续梅尔频谱特征提取提供干净的输入。

2.4 片段切分与特征提取

经过预处理的音频信号首先被切分为固定长度的片段。本文统一取片段时长为 0.5s (对应 16kHz 下 8000 个采样点); 为扩充训练样本数量, 支持集采用步长 0.25s (即 50% 重叠) 的滑动窗口切片, 而验证集与查询集采用无重叠的等长切片, 以保证评估输入长度一致且不引入重叠样本。每个 0.5s 片段随后被转换为对数梅尔频谱图, 转换流程如图 3 所示。

对每帧信号先作短时傅里叶变换 (STFT) 得到功率谱:

$$X(k, l) = \sum_{n=0}^{N-1} w(n) x(n) e^{-j2\pi kn/N} \quad (4)$$

式中, $x(n)$ 为帧内信号, $w(n)$ 为汉明窗, N 为帧长, l 为帧序号, k 为频点序号。随后将线性频率映射到更符合人耳听觉感知的梅尔频率, 梅尔频率 (f_{mel}) 与实际频率 f 的关系为:

$$f_{\text{mel}} = 2595 \lg\left(1 + \frac{f}{700}\right) \quad (5)$$

在梅尔刻度上布置一组三角滤波器组 ($H_m(k)$), 对功率谱进行滤波求和:

$$M_l(m) = \sum_{k=0}^{N/2} |X(k, l)|^2 H_m(k) \quad (6)$$

式中, m 为梅尔滤波器序号。为压缩动态范围并提升特征稳定性, 对滤波结果取对数, 得到对数梅尔频谱:

$$S(m, l) = 10 \lg(1 + M_l(m)) \quad (7)$$

最终将得到的频谱图插值缩放为 128×128 的单通道图像。由于本文采用的 ImageNet 预训练网络要求三通道输入, 故将单通道梅尔频谱在通道维复制为 3 通道。5 类叫声的梅尔频谱样例如图 4 所示, 可见不同类别叫声在时频分布上具有明显差异, 具备可分性。

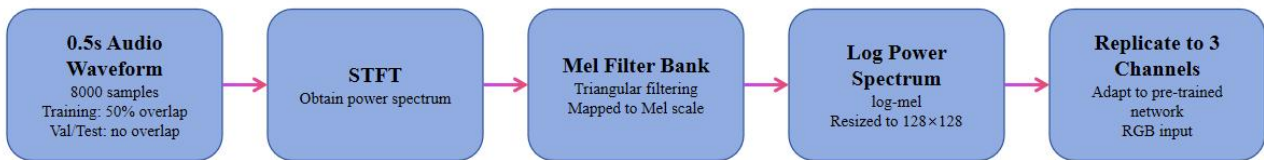


图 3: 梅尔频谱转换流程

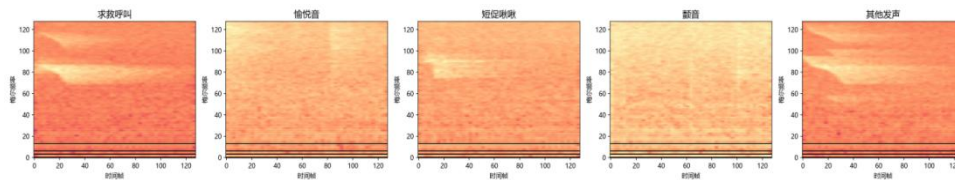


图 4: 5 类梅尔频谱样例

2.5 数据集划分

本文采用分层 5 折交叉验证评估模型性能, 划分的最小单位为原始音频文件。将全部 54 条音频按类别分层后均分为 5 折, 训练时轮流取其中 1 折作为测试集、其余 4 折作为训练集; 同一原始音频生成的所有切片与梅尔频谱图仅归属同一折, 绝不跨折拆分, 以避免数据泄露。每一折中, 训练集内部再按 8:1:1 划分为训练集、验证集与测试集, 其中训练集用于模型训练与数据增强, 验证集用于训练过程监控与早停, 测试集仅在每折训练结束后用于最

终性能评估。对梅尔频谱特征采用 Z-score 标准化,其均值与标准差仅由该折训练集统计得到,验证集与测试集沿用该统计量进行变换,最终所有性能指标均以 5 折交叉验证的均值与标准差报告。

(三) 数据增强策略

为缓解小样本条件下的过拟合并提升模型对信号缺失与频域干扰的鲁棒性,本文仅在每一折的训练集上采用 SpecAugment 策略进行数据增强,验证集与测试集保持原始状态不做任何增强。具体地,在梅尔频谱图上随机施加时间掩码与频率掩码,即随机选择 1~2 个连续时间区间和 1~2 个连续频率区间并将其置零,掩码宽度不超过对应维度总长的 10%,用以模拟真实养殖场景中可能出现的短暂信号丢失与频域干扰。该操作在不改变样本类别标签的前提下增加了训练样本的多样性,且由于增强仅作用于训练集,不会引入数据泄露。

(四) 骨干网络与迁移学习策略

本文以 AlexNet 为例,作为卷积骨干网络。AlexNet 由 5 个卷积层和 3 个全连接层组成,并在卷积层后串联最大池化层,采用 ReLU 激活函数缓解梯度消失。其卷积层对输入的梅尔频谱图进行局部特征提取,卷积运算表示为:

$$y_j^{(l)}(x) = \sigma(\sum_i K_{ij}^{(l)} * x_i^{(l-1)}(x) + b_j^{(l)})(8)$$

式中, $x_i^{(l-1)}$ 为第 l-1 层的第 i 个输入特征图, $K_{ij}^{(l)}$ 为卷积核, $b_j^{(l)}$ 为偏置, $\sigma(\cdot)$ 为 ReLU 激活函数,即 $\sigma(z) = \max(0, z)$ 。卷积与池化逐层提取从局部到全局的时频特征,最终经全局平均池化与全连接层映射到类别空间。分类阶段采用 Softmax 函数输出各类别概率:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}}(9)$$

式中, z_i 为第 i 类的得分, $C=5$ 为类别数。模型训练采用交叉熵损失:

$$\mathcal{L} = -\frac{1}{B} \sum_{n=1}^B \sum_{i=1}^C y_i^{(n)} \log p_i^{(n)}(10)$$

式中, B 为批大小, $y_i^{(n)}$ 为第 n 个样本的真实标签 (one-hot 形式), $p_i^{(n)}$ 为模型预测的第 i 类概率。

迁移学习方面,本文在 ImageNet 预训练权重的基础上对比两种策略:特征提取式冻结卷积骨干,仅训练顶层全连接分类器;微调式不冻结骨干,允许整个网络在本数据集上联合更新。AlexNet 网络结构如图 5 所示。

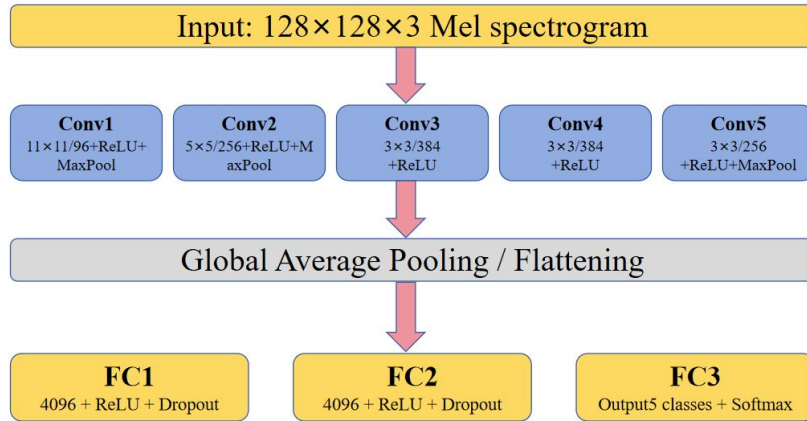


图 5: AlexNet 网络结构示意图

二、实验验证

(一) 实验环境与参数

本文实验基于 Windows 操作系统搭建,使用 PyTorch 深度学习框架实现,硬件配置为 NVIDIA RTX3050 Laptop GPU (4GB 显存),并采用 CUDA 加速。所有模型均加载 ImageNet 预训练权重初始化,在统一的实验条件下进行对比,具体环境与超参数设置如表 2 所示。

表 2 实验环境与超参数设置

类别	参数	设置
硬件	GPU	NVIDIA RTX3050 Laptop (4GB)
框架	深度学习框架	PyTorch 2.5.1 (CUDA 12.1)
输入	梅尔频谱尺寸	128×128×3
训练	训练轮次 Epoch	100
训练	批大小 BatchSize	32
训练	优化器	SGD (momentum=0.9)
训练	初始学习率	0.001
训练	权重衰减	1×10^{-4}
训练	学习率策略	余弦退火
训练	早停	验证损失连续 10 轮不下降则终止
评估	交叉验证	分层 5 折、8:1:1 划分

(二) 定量结果对比

在统一实验条件下,四种卷积网络分别在特征提取式与微调式两种迁移策略下的对比结果如表 3 所示,所有指标均为 5 折交叉验证的均值±标准差。

表 3 四种网络与两种迁移策略的性能对比

骨干网络	迁移策略	准确率 Acc(%)	Macro-F1(%)	参数量(M)	MACs(M)
MobileNet-V2	特征提取式	39.32±9.33	28.98±6.38	2.23	106.5
MobileNet-V2	微调式	54.27±10.62	39.80±10.32	2.23	106.5
EfficientNet-B0	特征提取式	47.22±7.16	34.93±6.54	4.01	135.6
EfficientNet-B0	微调式	44.59±11.19	32.15±7.39	4.01	135.6
ResNet-18	特征提取式	49.00±9.11	36.68±4.02	11.18	595.4
ResNet-18	微调式	51.41±7.06	40.47±5.51	11.18	595.4
AlexNet	特征提取式	59.54±5.20	49.10±5.83	57.02	250.8
AlexNet	微调式	52.84±9.32	41.37±6.85	57.02	250.8

由表 3 可见,AlexNet 在特征提取式策略下取得最优整体性能,准确率达 59.54%、Macro-F1 为 49.10%。值得注意的是,迁移策略的优劣并非绝对:轻量级的 MobileNet-V2 在微调式下准确率提升了约 15 个百分点,而结构较复杂的 AlexNet 与 EfficientNet-B0 在小样本下采用特征提取式(冻结骨干)反而更稳定、波动更小

为进一步分析最佳模型在各类别上的表现,将 AlexNet 特征提取式方案的逐类结果列于表 4

表 4 最佳模型(AlexNet 特征提取式)逐类识别结果

类别	精确率 P (%)	召回率 R (%)	F1 (%)
求救呼叫	93.3±5.5	95.0±6.1	93.9±4.0
愉悦音	70.0±26.7	63.3±36.0	58.2±27.6
短促啾啾	53.4±21.6	35.9±9.4	39.9±9.3
颤音	59.4±5.4	53.8±19.3	53.4±11.8
其他发声	0.0±0.0	0.0±0.0	0.0±0.0

由表 4 可知,"求救呼叫"类识别效果最好(F1 达 0.94),这与其特征明显、样本较清晰有关;"其他发声"类因仅 5 个样本、且特征混杂,模型未能有效学习。

(三) 可视化分析

本文以最佳实验模型 AlexNet(特征提取式)为例进行可视化分析。图 6 为模型在测试集上的混淆矩阵,"求救呼叫"类几乎全部被正确识别、对角元素最高,而"愉悦音""颤音"等类别之间存在少量相互混淆,"其他发声"类由于样本量过少、特征混杂,模型未能有效区分。图 7 为模型高层特征的 t-SNE 二维可视化结果,"求救呼叫"类在特征空间中聚簇明显、与其他类别分离较好,其余类别之间存在一定程度的重叠,这与混淆矩阵和定量结果反映的情况一致,也说明当前模型在类别可分性较好的发声上表现稳定,而在样本稀缺、特征相近的类别上仍有提升空间。

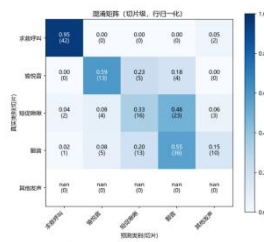


图 6: AlexNet 特征提取式模型测试集混淆矩阵

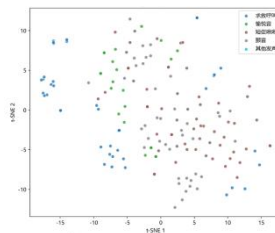


图 7: AlexNet 特征提取式模型测试集特征 t-SNE 可视化

三、结论

本文针对肉鸡叫声识别中传统人工听辨费时费力、主观一致性差,且实际养殖场景下标注样本稀缺、骨干网络与迁移学习策略缺乏系统比较的问题,以自建 5 类、54 条肉鸡叫声数据集为研究对象,将音频转换为 128×128 维梅尔频谱图作为输入,系统对比了四种典型卷积神经网络在特征提取式与微调式两种迁移学习策略下的识别性能,并结合数据增强与分层 5 折交叉验证进行多维度评估,得出以下主要结论:

1. AlexNet 特征提取式方案在本任务中取得最优综合性能。实验结果表明, AlexNet 在冻结骨干、仅训练分类器的策略下取得总体准确率 59.54%、宏平均 F1 值 49.10%,为八组实验中的最优结果。这一结果说明,在小样本条件下,预训练 AlexNet 所提取的通用时频特征已能较好地表征肉鸡叫声,全网络微调反而容易在有限数据上发生过拟合;同时也表明模型性能并非随网络规模增大而单调提升,选择与数据规模相匹配的骨干网络比盲目使用更深模型更为重要。

2. 迁移学习策略应根据骨干网络复杂度协同选择,数据增强与规范划分有效抑制了过拟合。对比可见,轻量级的 MobileNet-V2 在微调式策略下准确率由 39.32% 提升至 54.27%,提升显著;而结构较复杂的 AlexNet 与 EfficientNet-B0 在微调时性能不升反降,说明小样本下复杂网络微调容易破坏预训练通用特征。此外,仅在训练集上施加 SpecAugment 增强、按原始音频进行分层 5 折划分并使用训练集统计量做归一化,有效避免了数据泄露,训练曲线与 t-SNE 可视化均显示模型在类别可分性较好的样本上收敛稳定。

3. "求救呼叫"类叫声识别效果突出,具备实际养殖预警价值。在最佳模型下,"求救呼叫"类的 F1 值达到 0.94,精确率与召回率均超过 90%,说明该类叫声时频特征显著、易于识别,可用于鸡群应激与危险状态的早期自动预警;而"其他发声"类因仅 5 个样本且特征混杂,模型未能有效学习,符合小样本学习的客观规律,也提示后续应补充该类样本。总体而言,本文以较小代价建立了一套规范的肉鸡叫声识别对比流程,为智慧养殖中禽声自动识别系统的模型选型提供了参考。

展望未来,在样本规模扩充、细粒度发声类别划分以及面向养殖现场的边缘设备轻量化部署等方面仍有提升空间,后续将围绕扩大标注数据规模、引入注意力机制与多模态融合、优化模型推理效率三个方向展开研究,推动肉鸡叫声识别技术在实际养殖场景中的落地应用。

参考文献:

- [1] LI G B, LIU W X. Overcoming data scarcity: few-shot pig vocalization recognition via domain expansion, knowledge transfer, and feature alignment[J]. *Animals*, 2026, 16(13): 2074.
- [2] LI G B, SUN J Y, GUAN M Y, et al. A new method for non-destructive identification and tracking of multi-object behaviors in beef cattle based on deep learning[J]. *Animals*, 2024, 14(17): 2464.
- [3] Automated detection of broiler vocalizations: a machine learning approach for broiler chicken vocalization monitoring[J/OL]. 2024.
- [4] 基于改进 MobileNetV3 的笼养蛋鸡声音分类识别方法[J]. *农业机械学报*, 2025, 56(4).
- [5] 基于轻量化卷积神经网络的笼养鸡异常发声识别方法[J]. *农业机械学报*, 2026, 57(2).

Implementation and Teaching Research on Few-shot Recognition of Broiler Vocalizations Based on Mel Spectrogram and Convolutional Neural Networks

ZHAOYongcai¹ LIMan¹ GAOJingru¹ GUANManyu¹ LIGuangbo^{1*}

1 College of Electronic and Information Engineering, Huaibei Institute of Technology, Huaibei 235000, China

Abstract: Broiler vocalizations can intuitively reflect their physiological and behavioral states such as feeding, stress and distress. Accurate recognition of broiler vocalizations is of great significance for animal welfare monitoring and abnormal early warning in large-scale farming. However, labeled samples of broiler vocalizations are difficult to obtain. Most existing studies focus on a single network model, lacking a systematic comparison of different backbone networks and transfer learning strategies under few-shot conditions, resulting in insufficient basis for model selection.

To address this issue, this paper constructs a broiler vocalization dataset containing 54 original audio clips covering 5 categories: distress call, pleasure sound, short chirp, trill and other vocalizations. All audio samples are uniformly converted into 128×128 Mel spectrograms as model input. Four typical convolutional neural networks are adopted, and comparative experiments are carried out under two transfer learning strategies: feature extraction and fine-tuning. Stratified 5-fold cross-validation is applied to evaluate the model performance.

The experimental results show that the AlexNet feature extraction scheme achieves the optimal recognition performance, with an overall accuracy of 59.54% and a macro-F1 score of 49.10%, among which the F1 score of distress call reaches 0.94. The lightweight network MobileNet-V2 obtains significant performance improvement under the fine-tuning strategy, while for networks with more complex structures, feature extraction training with frozen backbone is more stable under few-shot conditions.

The research indicates that in few-shot broiler vocalization recognition tasks, the backbone network and transfer learning strategy should be selected collaboratively according to model complexity. This conclusion can provide methodological reference for the selection of automatic poultry sound recognition systems in farming scenarios.

Keywords: broiler vocalization recognition; Mel spectrogram; convolutional neural network; transfer learning; few-shot learning