

采用多尺度交叉轴注意力融合网络的 咽喉疾病影像分类算法

胡延苏¹, 王远峥¹, 侯瑾^{2,3}

(1. 长安大学电子与控制工程学院, 710064, 西安; 2. 西安交通大学第二附属医院, 710114, 西安;
3. 陕西省耳鼻咽喉疾病精准诊疗重点实验室, 710114, 西安)

摘要: 针对咽喉疾病计算机辅助诊断中存在的病灶视角单一、分类精度不足及依赖主观经验判断等问题, 提出基于多尺度交叉轴注意力融合网络的咽喉疾病影像算法。引入多尺度交叉轴注意力机制, 通过跨空间维度的自适应特征加权, 实现细粒度病理特征的精准定位。构建双流协同学习框架, 捕获全局解剖特征并聚焦局部上下文信息, 实现多视角特征提取。设计特征耦合模块避免简单拼接导致的信息冲突与语义错位, 实现特征空间的优化、局部-全局表征的深度整合。研究结果表明: 相比于现有的 MVT-OFML 模型, 所提算法在自制单视角、低对比度咽喉影像数据集上的准确性、精确度、召回率、F1 分数、特异度可分别获取 1.76%、1.56%、3.34%、2.60%、1.10% 的性能提升。同时, 特征可视化分析揭示了融合网络的决策机制具备显著可解释性, 关键病理区域激活强度分布与临床诊断标准保持高度一致。该研究可为智能化辅助诊断系统的开发提供技术支撑。

关键词: 咽喉疾病; 多尺度交叉轴注意力; 细粒度病理特征; 双流协同学习框架; 特征耦合模块

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.7652/xjtuxb202606022 **文章编号:** 0253-987X(2026)06-0256-11

Image Classification Algorithm for Laryngeal Disease Based on a Multi-Scale Cross-Axis Attention Fusion Network

HU Yansu¹, WANG Yuanzheng¹, HOU Jin^{2,3}

(1. School of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China;

2. The Second Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710114, China;

3. Shaanxi Provincial Key Laboratory for Precision Diagnosis and Treatment of Otorhinolaryngology, Xi'an 710114, China)

Abstract: To address limitations in computer-aided diagnosis of laryngeal disease, such as single-view lesion presentation, insufficient classification accuracy, and reliance on subjective experience-based judgment, an image classification algorithm for laryngeal disease based on a multi-scale cross-axis attention fusion network was proposed. A multi-scale cross-axis attention mechanism was introduced to enable precise localization of fine-grained pathological features through adaptive feature weighting across spatial dimensions. A dual-stream collaborative learning architecture was constructed to capture global anatomical structures while focusing on local contextual information, facilitating multi-view feature extraction. A feature coupling module was designed to prevent information conflict and semantic misalignment caused by simple concatenation,

收稿日期: 2025-10-24。 作者简介: 胡延苏(1985—), 女, 副教授, 硕士生导师; 侯瑾(通信作者), 女, 副主任医生。

基金项目: 陕西省自然科学基金资助项目(2025JC-YBMS-704, 2024JC-YBQN-0889, 2021JQ-419)。

网络出版时间: 2026-01-22

网络出版地址: <https://link.cnki.net/urlid/61.1069.T.20260121.1851.002>

optimizing the feature space and deeply integrating local and global representations. Study results showed that, on an in-house single-view, low-contrast laryngeal image dataset, compared with the existing MVT-OFML model, improvements in accuracy, precision, recall, F1 score, and specificity of 1.76%, 1.56%, 3.34%, 2.60%, and 1.10%, respectively, were achieved. Also, feature visualization analysis revealed that the fusion network's decision mechanism exhibited noticeable interpretability, with the distribution of activation intensities in key pathological regions being highly consistent with clinical diagnostic criteria. This study is expected to provide technical support for the development of intelligent computer-aided diagnosis systems.

Keywords: laryngeal disease; multi-scale cross-axis attention; fine-grained pathological feature; dual-stream collaborative learning architecture; feature coupling module

近年来,喉部疾病的发病率呈上升趋势,严重影响了患者的生活质量。然而,由于人类喉部的特殊生理结构,人眼难以观察到所有的病变区域,从而导致缺乏内窥镜检查经验的医生容易做出误判^[1]。此外,不同医疗机构获得的数据具有很大的异质性^[2],严重制约了临床转化。同时,传统的咽喉类疾病诊断需要耳鼻喉科、放射科、病理科等多个学科的共同参与,综合各种检查结果和病理分析,以做出准确的诊断。因此,关注视角全面且精确的辅助诊断技术显得尤为重要。

为提升疾病筛查和辅助诊断的可靠性,研究者多模态深度学习方面进行了探索。Li等^[3]融合白光成像和窄带成像双模态特征,提高了咽喉癌的诊断准确性。Ma等^[4]展示了放射性组学模型在喉癌生存期预测中的关键作用。Yang等^[5]在喉镜视频分析中融入时间运动线索,实现了单侧声带麻痹的高精度检测。Wang等^[6]研究了条件表格式生成对抗网络在放射组学特征上的应用。Vethasri等^[7]采用稠密连接卷积系列网络(DenseNet),在早期喉癌分类任务中取得98%的准确率。然而,上述研究仍依赖大规模数据,且模型结构复杂,在跨样本差异较大或多病灶复杂结构下的可靠性仍有限。因此,如何进一步提升特征表达力,并降低模型对数据规模的敏感性成为重要的研究方向。

在此背景下,多模型集成逐渐成为近年医学图像分析的主要趋势。与单一模型相比,集成策略能够兼顾多尺度的结构信息、细粒度的病灶差异与图像噪声的鲁棒性,从而增强模型预测稳定性。Patil等^[8]在脑肿瘤磁共振成像图像分类研究中,整合浅层卷积网络与VGG16网络进行特征互补。Singh等^[9]采用了残差网络ResNet50与EfficientNet-B7的加权输出融合机制。Thakur等^[10]采用多数表决

机制整合ResNet、EfficientNet及VGG网络的预测结果,进一步证实了多模型协同决策的有效性。尽管集成策略在分类任务中表现优越,但其高计算复杂度、较大的参数量和推理成本仍限制了其在实时性或资源受限的临床场景中的应用。

在医学图像分析的拓展应用中,集成方法同样展现出广泛的适用性。Dudeja等^[11]构建了基于ResNet与EfficientNet-B4的混合架构,实现具有复杂模式识别能力的脑肿瘤分割系统。与单一算法相比,Al-Azzwi等^[12]集成的堆叠VGG19、Inception V3与深度残差网络,显著提升了脑肿瘤诊断预测任务的细粒度判别能力。针对喉癌检测的挑战,Bhattacharjee等^[13]提出了MobileNetV2、EfficientNet-B0和DenseNet121的加权投票框架,其诊断敏感度显著优于单模型。以上研究表明:集成策略在处理跨病灶类型、跨结构尺度以及高噪声医学图像任务时,展现出强大的适应性与精度优势。然而,这类方法通常面临计算复杂度高、参数规模大以及硬性集成等限制,难以在实时性要求较强或资源受限的临床应用场景中直接部署。

综上所述,尽管现有方法取得了显著进展,但在咽喉疾病影像诊断领域仍面临4大关键挑战:①单一卷积架构的感受野受限,无法同时覆盖不同尺度的病灶特征,导致多视角特征缺失;②局部-全局特征协同机制不足,无法捕获病灶的细粒度纹理与解剖上下文信息;③特征融合策略缺乏自适应性,存在信息冗余和特征冲突,导致融合不稳定;④模型决策的可解释性不足,缺乏实验结果的可视化分析。因此,为弥补轻量级网络在高阶表征能力上的不足,本文选取ResNet50与EfficientNet-B0作为基础网络,并在此基础上提出了一种基于多尺度交叉轴注意力融合网络(MCAF-Net)的咽喉疾病影像算

法,从多尺度特征视角、多轴向空间视角和多网络结构视角等多个特征,表达维度捕获喉部病灶信息的复合特征,旨在解决咽喉病灶影像中存在的多视角特征缺失与局部-全局特征协同不足的双重问题。

1 多尺度交叉轴注意力协同优化

本文提出一种基于多尺度交叉轴注意力(MCA)与 EfficientNet-B0 的双流融合架构,以实现咽喉疾病影像的精准分类。该网络包含 2 个分支网络:MCA 分支通过深层网络提取的语义特征,能够有效捕获跨区域的大范围结构差异,为疾病类型判别提供结构-语义层面的全局深层特征;EfficientNet-B0 分支负责捕获多尺度上下文信息,即喉部疾病具有的多尺度病灶特征,如细粒度病灶细节、病灶与临近组织的关联上下文、影像干扰因素的识别等。通过网络的协同优化,该双流融合架构能够模拟临床的多层次观察过程,同时捕获整体的形态变化(肿胀、闭合不全、对称性破坏),细粒度病灶纹理(白斑、突起),病灶与周围组织关系(边缘清晰度、纹理过渡特征)等特征。最终通过特征耦合模块对双分支提取的病灶特征进行协同优化,在提升特征多样性的同时增强模型泛化性能。

1.1 MCA 分支网络

MCA 分支网络采用经典的编码器-解码器结构,通过层级堆叠的残差块^[14]与多尺度交叉轴注意力模块协同工作。该网络通过逐级降采样实现多尺度特征的分层捕获,即:浅层保留高分辨率空间细节,深层提取低分辨率语义特征。

(1)编码器。轻量级的 ResNet50(LW-ResNet50)是基于 ResNet50 架构优化的轻量化卷积神经网络,其采用分阶段参数冻结策略实现分支压缩与泛化性能的提升。对原始架构的初始阶段 Stage1~2 进行参数冻结,在保证特征提取鲁棒性的同时将总参数量降低。由于浅层网络提取的低级纹理特征具有跨数据集的通用性,通过约束浅层卷积核的更新自由度,可以抑制小样本训练场景下的过拟合现象。后 3 个阶段(Stage3~5)输出的特征图分别标记为 E_2 、 E_3 和 E_4 ,联合原始输入的特征图 E_1 ,作为 MCA 分支网络的编码器。

(2)解码器。采用 SegNeXt^[15]中多模态自注意力与协作注意力网络(MSCAN)提取多尺度特征的思想,获取来自编码器的后 3 个阶段的特征图进行上采样组合,然后连接作为解码器的输入。设原始

图片的高度和宽度为 H 和 W ,则 E_1 、 E_2 、 E_3 、 E_4 的分辨率分别为 $H \times W$ 、 $H/8 \times W/8$ 、 $H/16 \times W/16$ 和 $H/32 \times W/32$ 。

本文采用的多尺度交叉轴注意力通过以下流程实现特征融合:首先,对 E_2 、 E_3 、 E_4 进行线性采样,将其尺寸调整为与 E_1 一致,得到 E'_2 、 E'_3 、 E'_4 ;随后,将这 3 个上采样的特征进行通道拼接,并通过 1×1 卷积实现通道降维,使输出通道数与 E_1 保持一致;最后,将降维后的特征输入多尺度交叉轴注意力模块。该过程可形式化表示为

$$\mathbf{F} = \text{MCA}(\text{Conv}_{1 \times 1}([\mathbf{E}'_2, \mathbf{E}'_3, \mathbf{E}'_4])) \quad (1)$$

式中:MCA($\cdot$)为多尺度交叉轴注意力操作; $\mathbf{F}$ 为输出特征图; $\text{Conv}_{1 \times 1}(\cdot)$ 为 1×1 卷积; $[\cdot]$ 为连接操作。

多尺度交叉轴注意力模块^[16]采用双分支并行架构,分别建模水平与垂直轴向特征关联,即: x 轴分支执行水平向一维卷积, y 轴分支实施垂直向一维卷积。各轴向处理单元包含 3 组核尺寸(1×3 、 1×5 、 1×7)的卷积层,能够确保多尺度的上下文捕获。以 x 轴处理流程为例,层归一化后,输入特征由一组并行一维卷积层处理,提取具有多级感受野的特征,然后输出多尺度轴向特征张量,其跨通道信息通过可学习权重矩阵实现动态融合,其表达式为

$$\mathbf{F}_x = \text{Conv}_{1 \times 1} \left(\sum_{i=0}^3 \text{Conv1D}_i^x(\text{Norm}(\mathbf{F})) \right) \quad (2)$$

类似地,在 y 轴方向表达式为

$$\mathbf{F}_y = \text{Conv}_{1 \times 1} \left(\sum_{i=0}^3 \text{Conv1D}_i^y(\text{Norm}(\mathbf{F})) \right) \quad (3)$$

式中: $\mathbf{F}_x$ 、 $\mathbf{F}_y$ 分别为 x 、 y 轴分支输出的多尺度轴向特征张量; $\text{Conv1D}_i^x(\cdot)$ 、 $\text{Conv1D}_i^y(\cdot)$ 分别为第 i 个卷积层沿 x 、 y 轴维度的一维卷积; $\text{Norm}(\cdot)$ 为层归一化。

为充分融合 2 个空间方向的多尺度卷积特征,在 $\mathbf{F}_x$ 与 $\mathbf{F}_y$ 之间建立交叉注意力机制,令 $\mathbf{F}_x$ 作为键矩阵 $\mathbf{K}$ 和值矩阵 $\mathbf{V}$, $\mathbf{F}_y$ 作为查询矩阵 $\mathbf{Q}$,则该机制的数学表达式为

$$\mathbf{F}_T = \text{MHCA}_y(\mathbf{F}_y, \mathbf{F}_x, \mathbf{F}_x) \quad (4)$$

式中: $\text{MHCA}_y(\cdot)$ 为沿 x 轴的多头交叉关注; $\mathbf{F}_T$ 为 x 轴向的输出。对于底部分支,使用类似的方法,在 y 轴方向上处理编码上下文,其表达式为

$$\mathbf{F}_B = \text{MHCA}_x(\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_y) \quad (5)$$

式中: $\text{MHCA}_x(\cdot)$ 为沿 y 轴的多头交叉关注; $\mathbf{F}_B$ 为 y 轴向的输出。

在给定 $\mathbf{F}_T$ 和 $\mathbf{F}_B$ 的基础上,多尺度交叉轴注意

力的输出 F_{out} 可以表示为

$$F_{out} = \text{Conv}_{1 \times 1}(F_T) + \text{Conv}_{1 \times 1}(F_B) + F \quad (6)$$

1.2 EfficientNet 分支网络

EfficientNet-B0^[17] 是基于深度可分离卷积架构的预训练模型,其主干网络采用 MobileNetV2 的核心组件,即倒置残差模块。该模块通过深度卷积提取空间特征后,采用逐点卷积进行通道信息融合,实现局部与全局特征的协同建模。模块内嵌的通道扩展与投影机制,支持跨尺度特征的空间-通道交互;压缩与激励(SE)模块通过注意力机制,增强关键特征表征,强化多尺度建模能力。同时,模型采用自适应缩放策略,能够动态调节网络深度、宽度及输入分辨率以平衡计算量与性能。倍率因子通过统一调控各层级增长率,实现网络维度的同步优化。该架构通过上述组件的协同作用,在保证计算效率的同时提升了模型性能。

1.3 特征耦合模块

MCA 分支网络和 EfficientNetB0 分支网络提取的特征表征在语义层次、空间尺度和通道分布上存在显著差异,直接拼接会导致:① 语义不一致性:不同网络对同一病灶的特征编码方式不同;② 通道冗余:简单拼接会使特征维度翻倍,引入大量冗余信息;③ 权重失衡:同分支对不同病灶类型的判别能力存在差异,需动态调整贡献权重。因此,本文设计了一种基于双层感知机的特征拼接模块,与传统的 1×1 卷积采用静态线性变换不同,该模块利用内容感知的策略对特征降维之前的权重动态重校准。

采用全局平均池化将空间信息聚合为通道级描述,其表达式为

$$z_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W f_c(i, j, c) \quad (7)$$

式中: z_c 为该通道在整个特征图上的平均响应强度; $f_c(i, j, c)$ 表示第 c 个通道在空间位置 (i, j) 处的特征值。 $\mathbf{z} = [z_1, z_2, \dots, z_C]^T \in \mathbf{R}^C$ 反映了该通道的全局激活水平, C 为 2 个分支网络通道数的总和。该操作将原始的三维特征图 $(H \times W \times C)$ 转换为一维通道向量 $\mathbf{z}$ 。该降维操作一方面可大幅减少参数量,降低计算开销;另一方面,通过限制中间层的维度,强制网络学习通道间最本质的依赖关系,增强了模型的泛化能力并防止过拟合。

在获得一维通道向量 $\mathbf{z}$ 后,通过双层感知机网络构成的学习通道间的非线性依赖关系,生成自适应的通道注意力权重。该过程可表示为

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \cdot \mathbf{z})) \quad (8)$$

式中: $\mathbf{W}_1$ 为第一层感知机网络的权重矩阵,实现通道信息的表征; $\delta(\cdot)$ 为 ReLU 激活函数,引入非线性以捕捉通道间的复杂交互模式; $\mathbf{W}_2$ 为第二层感知机网络的权重矩阵,完成通道依赖关系的建模; $\sigma(\cdot)$ 为 Sigmoid 激活函数,将输出值约束在 $(0, 1)$ 之间,并生成归一化的通道关注权重; $\mathbf{s}$ 为最终的重要性评分向量,具有样本特异性,即对于不同的输入图像,网络会动态生成不同的权重分布,从而实现内容自适应的权重选择。MCAF-Net 的整体结构如图 1 所示。

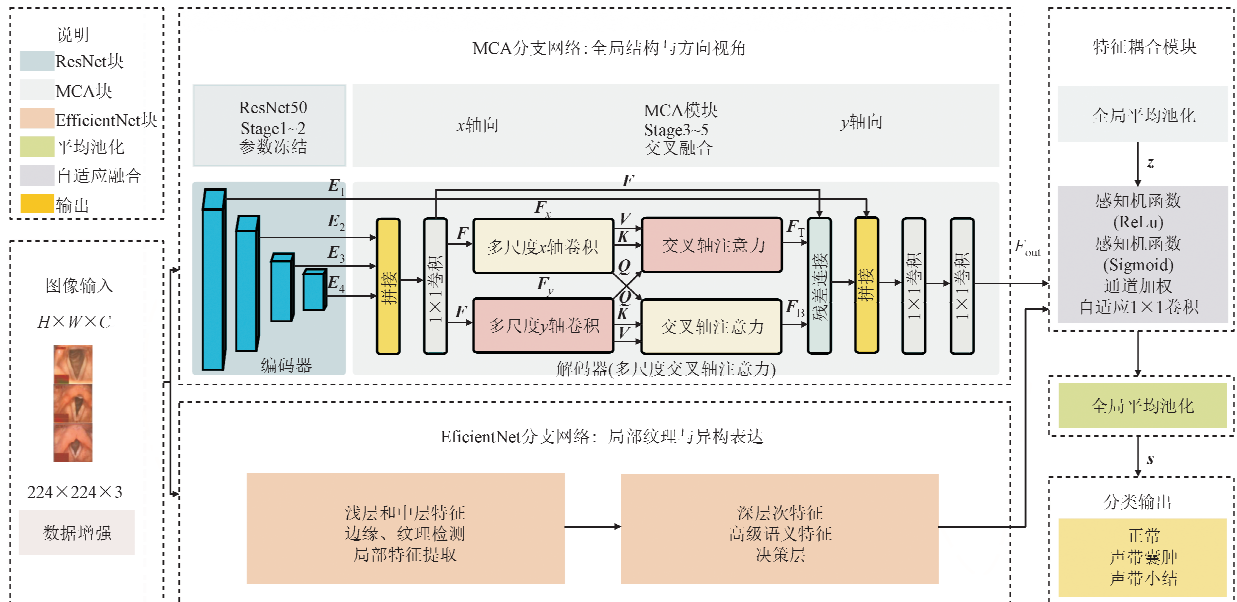


图1 MCAF-Net 整体结构

Fig. 1 MCAF-Net overall structure

2 实验数据集

2.1 自制数据集

本文实验数据集来自西安交通大学第二附属医院病人临床检测视频,选取了近 3 年 118 例咽喉疾病患者,并且所有选取的病例都经过了病理检验确认。在截取的 118 个病例(共 13 880 幅图像)中,包含 52 个声带囊肿病例(共 6 086 幅图像),50 个声带小结病例(共 5 848 幅图像),16 个正常类型患者样本(共 1 946 幅图像)。该数据集已得到批准可以研究,并放弃知情同意。

整个数据集按照病例数量 8 : 2 的比例划分训练集与测试集,训练集病例的 20%作为验证集。其中,训练集由 95 个病例(11 555 幅图像)组成,包括:42 个声带囊肿病例(5 108 幅图像)、40 个声带小结病例(4 745 幅图像)、13 个正常类型患者样本(1 702 幅图像)。测试集由 23 个病例(2 325 幅图像)组成,包括:10 个声带囊肿病例(978 幅图像)、10 个声带小结病例(1 103 幅图像)、3 个正常类型患者样本(244 幅图像)。

2.2 数据预处理

首先,将图像尺寸统一调整为 224×224 以实现大小归一化。随后,对咽喉病变图像进行标签分类:声带囊肿标注为 1,声带小结标注为 2,正常样本标注为 0。最后,将图像数据与对应标签按行合并,并随机打乱数据集以消除切片顺序对病灶特征的潜在影响。

针对数据量不足及类别不均衡的问题,采用数据增强与交叉验证双重优化策略以增强模型泛化能力。图像增强结果如图 2 所示。

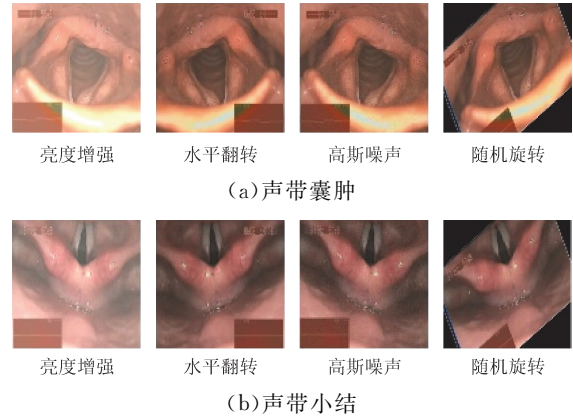


图 2 咽喉类疾病病理学图像增强

Fig.2 Pathological images enhancement of throat diseases

由图可知,对声带囊肿及声带小结样本进行几

何增强(旋转、翻转)和光度变换(亮度、高斯噪声)^[6],以实现样本空间的有效扩展。同时,采用五折交叉验证^[18],在提高数据利用率的基础上,降低评估过程的随机偏差,并增强验证结果的说服力,有效抑制模型过拟合风险。上述方法改善了模型的抗干扰性,也为小样本条件下的医学影像分类任务提供了借鉴。

3 实验与分析

3.1 实验设置

实验环境为 Windows 操作系统,硬件配置为 NVIDIA GTX 1660 Ti 显卡,采用 TensorFlow 深度学习框架进行构建。其中,输入图片大小为 224×224,训练周期设置为 50 轮次,批次训练样本数量为 8,余弦退火学习率设为 1×10^{-3} ,丢弃率为 0.4,训练冻结 ResNet50 前 2 阶段。

3.2 评价指标

从得到的混淆矩阵结果计算精确度 P 、准确率 A 、召回率 R 、F1 分数 F_1 、特异度 S 。令 T_P 为真正例的数量, F_N 为假反例的数量, F_P 为假正例的数量; T_N 为真反例的数量,各计算式如下

$$A = (T_P + T_N)/(T_P + T_N + F_P + F_N) \quad (9)$$

$$P = T_P/(T_P + F_P) \quad (10)$$

$$R = T_P/(T_P + F_N) \quad (11)$$

$$F_1 = 2PR/(P + R) \quad (12)$$

$$S = T_N/(T_N + F_P) \quad (13)$$

3.3 对比实验

ResNet 系列网络独特的模块化设计,能够在医学图像中提取更高层次且更具判别力的特征。本文将该系列网络选作 MCA 分支的主干网络,并在自建的咽喉疾病数据集上开展实验,其结果如表 1 所示。

表 1 主干网络对比实验结果

Table 1 Results of the comparative experiment on backbone network

实验策略	准确率	精确度	召回率	F1 分数	特异度
ResNet18	83.57	82.55	79.91	80.84	90.74
ResNet18	87.61	82.95	83.94	83.39	93.44
融合 MCA	85.94	84.16	81.40	82.61	91.98
ResNet34	88.56	85.06	86.29	85.61	93.83
融合 MCA	88.22	84.20	87.97	85.79	93.72
ResNet50	88.60	87.54	87.37	87.45	93.51
融合 MCA	88.17	87.56	84.44	85.84	93.17

通过对比实验结果可以看出,ResNet50在精确度、召回率和F1分数方面处于较高水平,这表明ResNet50在提取咽喉疾病特征方面具有较强的判别能力。引入MCA模块后,模型在保持较高准确率的基础上,进一步提高了精确度与召回率,使得整体性能更优。另一方面,ResNet101尽管在精确度指标上略优于ResNet50和MCA的融合,但其训练时长约为这个融合组合的3倍。这主要是因为ResNet101层数更多,导致计算复杂度提高。在硬件资源有限的情况下,训练效率明显降低,且性能提升有限。考虑到模型复杂度与训练时间之间的折中,最终选择ResNet50作为MCA分支的主干网络。

在此基础上,进一步将MCAF-Net算法与参考文献中的DenseNet^[7]、SCNN-VGG^[8]、Res-Eff^[9]和Res-Eff-VGG^[10]算法,以及当前先进的LG-CAFN^[19]、Mobile-VIT^[20]和MVT-OFML^[21]这3种分类算法在自制咽喉疾病数据集上进行了对比实验,其结果见表2。

从准确率分析,本文算法(MCAF-Net)在复杂病理影像分类中具有突出优势。值得注意的是,在数据分布不均衡的条件下,多数算法呈现出精确度高于准确率的现象,这反映出模型对正类样本的识别能力较强,且预测为正类的样本中假阳性比例较低。而MCAF-Net的召回率相比LG-CAFN、

Mobile-VIT和MVT-OFML分别提升了4.16%、3.46%和3.34%,这也验证了MCAF-Net在病理特征捕获方面的全面性。另外,该模型的F1分数达到了95.11%,在综合平衡精确度与召回率的评价指标下仍保持最优性能。特异度指标达到了97.71%,对比表现较好的MVT-OFML,该指标提升了1.10%,从而证实了其在非病理样本的鉴别中具有更强的判别能力。

表2 分类算法对比实验结果

Table 2 Results of the comparative experiment on classification algorithms

算法	准确率	精确度	召回率	F1分数	特异度
DenseNet	88.30	85.91	83.21	84.37	93.42
SCNN-VGG	91.33	91.17	85.93	88.00	95.06
Res-Eff	91.95	90.61	87.27	88.68	95.54
Res-Eff-VGG	90.25	90.39	83.51	85.97	94.40
LG-CAFN	93.12	93.75	89.97	91.64	95.96
Mobile-VIT	93.76	94.01	90.67	92.17	96.29
MVT-OFML	94.28	94.66	90.79	92.51	96.61
MCAF-Net	96.04	96.22	94.13	95.11	97.71

同时,本文对编码器和解码器的选择进行了对比试验,以评估不同编码器和解码器对MCAF-Net的影响,其结果见表3。

表3 编码器解码器实验结果

Table 3 Encoder decoder experimental results

编码器	解码器	准确率	精确度	召回率	F1分数	特异度
DenseNet121	MCA	88.90	86.25	89.07	87.44	94.00
MobileNet		92.86	93.30	90.40	91.64	95.91
ResNet50	SE	95.91	94.87	94.72	94.79	97.71
	CAA	95.35	93.19	94.71	93.91	97.53
ResNet50	MCA	96.04	96.22	94.13	95.11	97.71

表3的前2行是通过使用2种不同的编码器,包括DenseNet121^[22]和MobileNet^[23],验证不同编码器对MCAF-Net的影响。此外,也测试了不同解码器对MCAF-Net的影响,如比较常见的SE^[24]和上下文锚点注意力机制(CAA)^[25]。虽然其他主流的解码器也能带来良好的实验性能,但MCA是轻量级的,且效果表现

最佳。这一系列实验表明:优化编码器与解码器设计,是提升医学图像分类模型性能的重要手段之一。

同样,为验证所提算法的有效性和在公开数据集上的泛化能力,选取了具有代表性的肺部和脑部疾病图像数据集,并对比了几种算法模型的性能,其结果见表4。

表4 公开数据集对比实验结果表

Table 4 Results of the comparative experiment of publicly available datasets

算法	肺部CT图像			脑部CT图像		
	准确率	F1分数	特异度	准确率	F1分数	特异度
LG-CAFN	92.52	92.65	96.17	98.01	98.01	99.01
Mobile-VIT	92.95	93.10	96.40	98.12	98.12	99.06
MVT-OFML	91.94	92.12	95.87	98.34	98.34	99.17
MCAF-Net	93.38	93.51	96.61	98.45	98.46	99.23

综合公开数据集的验证结果,MCAF-Net 不仅在自制数据集上表现优异,还展现了良好的泛化能力。结合在喉部数据集的测试结果(如表 2 所示)可以看出,自适应特征耦合能够学习到网络分支的实际贡献比例,因此相比于硬性权重组合和投票机制,在准确率、F1 分数、特异度等评估指标上取得稳定且显著的提升,这表明了灵活的特征融合策略能够获得更稳定且更具判别力的特征表示。

图 3 的可视化分析进一步验证了 MCAF-Net 的优越性能。通过观察,MCAF-Net 在肺部和脑部

图像中的热点区域更集中,表现出对关键病灶的精准捕捉能力。同时,MCAF-Net 的热力图边界更精细,说明其特征提取和空间注意机制更有效地激活了重要区域,有利于提高分类的准确性。肺部 CT 图像中,MCAF-Net 对病变区域的关注度明显高于其他模型,减少了误区的风险。在脑部 CT 图像中,其热力图突出显示了脑部异常区域,增强了疾病定位的可靠性。这一可视化结果有效验证了前面定量指标的优势,增加了模型在实际诊断中的解释性和可信度。

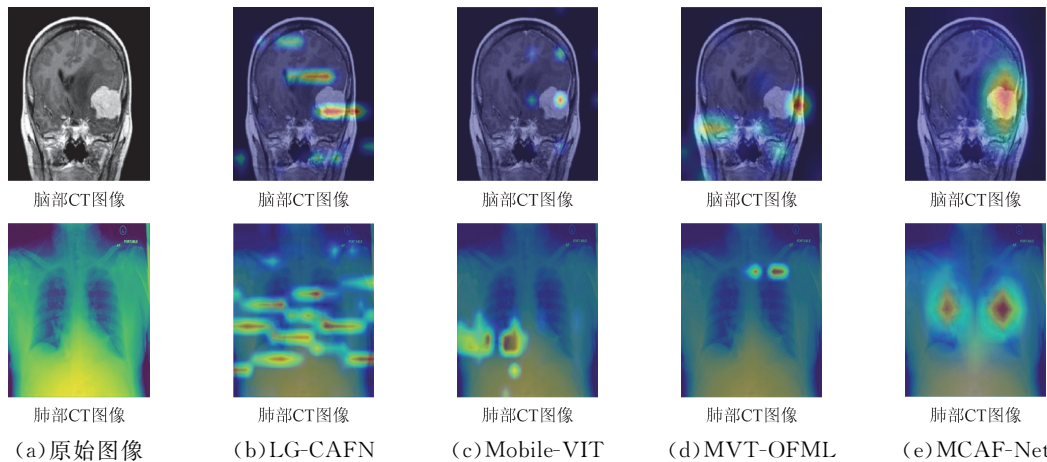


图 3 公开数据集可视化分析
Fig. 3 Visualization analysis of public datasets

3.4 消融实验

为全面评估各组件在 MCAF-Net 中的作用,通

过喉部病变分类任务的自制数据集开展了消融实验,其结果见表 5。

表 5 MCAF-Net 的消融实验结果
Table 5 Results of ablation experiment of MCAF-Net

RES	是否加入模型			准确率	精确度	召回率	F1 分数	特异度
	EF	MCA	COUPL					
否	是	否	否	84.22	78.58	84.02	80.21	91.95
是	否	否	否	84.56	83.94	81.28	82.44	90.98
是	是	否	否	87.10	87.86	82.85	84.53	92.69
是	否	是	否	89.16	86.85	86.40	86.51	94.06
是	是	是	否	94.92	93.11	91.85	92.44	97.16
是	是	是	是	96.04	96.22	94.13	95.11	97.71

注:RES 代表 ResNet50 网络,EF 表示 EfficientNet-B0 网络,MCA 表示多尺度交叉轴注意力应用到 ResNet 网络中,COUPL 表示特征耦合块。

当仅使用 EF 网络时,准确率、精确度和召回率均较低,表明单一基础网络能够完成初步分类任务,但受限于网络特征提取能力,难以充分捕捉疾病特异度特征。而 RES 将 EF 网络精确度提高至 83.94%,F1 分数达到 82.44%,说明能够提取更具代表性的特征。然而,略低的特异度,表明其对非病灶区域的辨别能力仍存在不足。当仅采用 2 个主干

网络并联时,相较于单一主干网络,在 F1 分数、特异度等关键性能指标上能够带来性能提升,说明双流架构可利用两网络特征之间的互补性,但仍低于加入 MCA 与 COUPL 后的模型性能。MAC 机制的融合,能够有效聚焦病变区域,强化关键信息的提取,进一步提高了其特征表达能力。当三者都使用时,实现了全面的性能提升,展现了模块协同作用对

疾病分类任务的优化效果。最后,引入特征耦合块进一步优化不同模块间的特征关联信息,使得准确率达到 96.04%,精确度和特异度分别为 96.22%和 97.71%。这一结果说明特征耦合块能够最大化模型对疾病分类的细粒度能力。

为研究双流融合架构中,每个组件的工作原理,针对声带囊肿、声带小结,各选择 3 种病例图像(分别为病例图像 1~3、病例图像 A~C),通过 MCAF-Net 生成的特征图进行了可视化,其结果如图 4 和图 5 所示。由图可知,单一视角模型关注区域存在明显偏移,更多关注的是非病理区域,而多视角融合模型能够定位病理观察区。具体来说,单独的 EfficientNetB0 更关注 y 轴向特征;ResNet50 则侧重病灶周围纹理;MCA 显著改善了关注区域但全

局表征不足;而本文所提出的双流融合架构可很好的实现全局的特征表征与病变区域的聚焦。这也说明局部视角、x 轴向视角、y 轴向视角以及多网络结构视角在多视角特征中每一部分都不可替代。

此外,将临床关注的区域在图 4、图 5 的原始图像中用绿色高亮方框标注。从热力图的空间分布可以看出,该改进算法的特性关注点与临床诊断中医师所聚焦的区域高度吻合,具有较高的医学解释性和可靠性。值得注意的是,由于能够广泛覆盖病灶区域的特征分布,使得本文算法能更敏感地捕捉与病理相关的细微特征。这一优势不仅提升了模型的分类性能,还显著增强了模型输出结果的医学可解释性,为后续的临床应用和辅助诊断提供了有力支持。

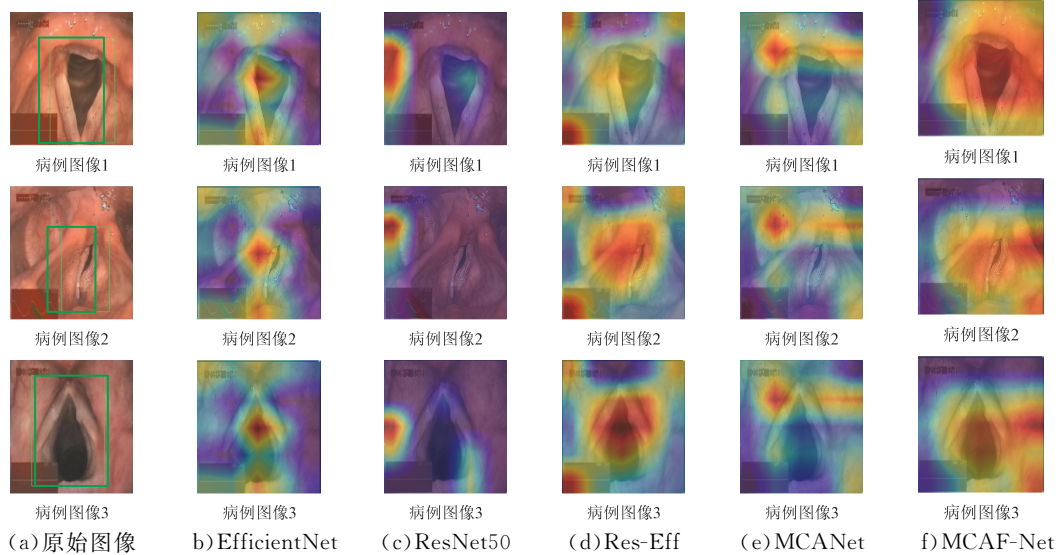


图 4 基于 MCAF-Net 消融研究的可视化结果(声带囊肿)

Fig. 4 Visualization results based on MCAF-Net ablation research (vocal fold cysts)

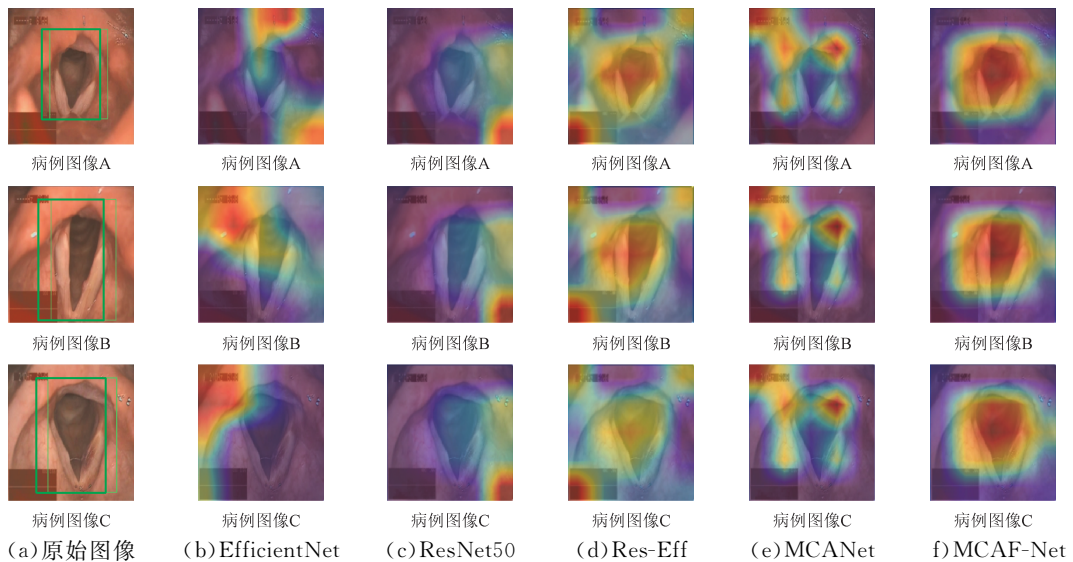


图 5 基于 MCAF-Net 消融研究的可视化结果(声带小结)

Fig. 5 Visualization results based on MCAF-Net ablation research (vocal nodules)

此外,进一步探讨编码器中特征图(即 E_1 、 E_2 、 E_3 和 E_4)的选取策略,对 ResNet50 主干网络进行消融实验,在此过程中保证 E_1 特征图的持续存在,以使其包含必要的边界线索,实验结果见表 6。结果显

示,去除主干网络初始的 2 个阶段是可行的,这可能是因为前 2 个阶段的卷积层较浅,而低级特征不适合捕捉高级语义,只能类似于 E_1 特征图,从中获取辅助病变区域的边界细节。

表 6 MCANet 的消融实验结果

Table 6 Results of ablation experiment of MCANet

是否加入阶段或模型			准确率	精确度	召回率	F1 分数	特异度
Stage1	Stage2	MCA					
是	是	否	88.22	84.20	87.97	85.79	93.72
是	是	是	88.60	87.54	87.37	87.45	93.51
否	是	否	86.32	81.13	84.43	82.51	92.85
否	是	是	87.66	85.14	80.18	82.07	93.07
否	否	否	84.56	83.94	81.28	82.44	90.98
否	否	是	89.16	86.85	86.40	86.51	94.06

3.5 可视化分析

为进一步提升模型的可解释性并直观观察提取特征的分布特性,本文采用 t -分布随机邻域嵌入算法(t -SNE),将咽喉疾病测试集的高维特征降维至二维空间进行可视化分析,即:将声带囊肿、声带小结及正常样本的特征投影至低维空间,以揭示模型对不同类别特征的辨别能力及相关特性分布规律。

如图 6 所示,不同疾病类别的特征聚类边界更清晰,仅在少数局部区域出现重叠现象,这说明多视角融合能够有效提升类间可分性。此外,特征分布的适度离散性进一步反映了医学图像数据的固有异质性,为优化数据采集和标注提供了有力的理论依据。然而,类别之间局部混淆区域的存在,特别是声带囊肿与声带小结样本的部分交叠,暗示了 2 类疾病在形态特征上具有一定的相似性。这类边界样本的形态模糊性可能会对模型的分类能力造成挑战。因此,将边界样本进行更加精确的人工标注和特征强化,将成为模型进一步优化的关键方向。

图 7 展示了基于分水岭算法的局部可解释解释器(LIME)可视化结果,其中颜色深浅反映了模型对各区域的关注程度。分水岭算法先对图像进行超像素分割,将其划分为若干结构连贯的区域,进而实现局部区域的扰动分析,量化每个区域对模型输出的贡献度。由图可知,模型对影像边缘的凹陷区域和医疗器械背景给予了较低置信度,对病灶中央区域关注度显著提高,这证明了多视角特征增强了模型对细微纹理和边缘的响应能力。分水岭算法提供的超像素划分,有效隔离了关键组织结构与周围环境,减少了背景噪声的干扰,提高了解释的准确性和可靠性,体现了其在医学图像分析中的应用潜力。

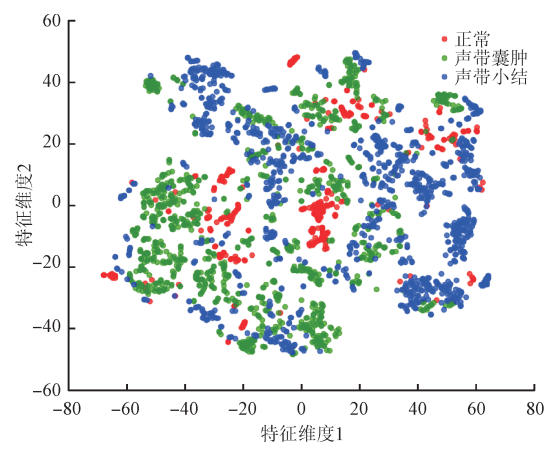


图 6 t -SNE 可视化结果

Fig. 6 t -SNE visualization results

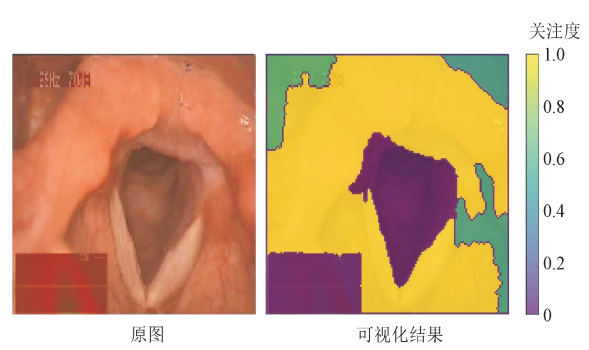


图 7 基于分水岭算法的 LIME 可视化结果

Fig. 7 LIME visualization results based on watershed algorithm

图 8 为基于 SHAP 方法的可视化结果。热力图中,红色区域表示对模型预测具有正向贡献,蓝色区域代表负向贡献。颜色深浅反映了 SHAP 值 ϕ 的大小,越接近红色说明该区域对该类别预测的重要性越高,蓝色则表示对预测结果有抑制作用。值得注意的是,SHAP 方法分析更多关注的是病灶上

部区域,这可能是该部分在博弈时会带来更高的收益。比如在声带小结分类中,图像中央部分出现了明显的红色高亮,表明该区域特征对模型识别结节具有显著推动作用。相反,部分外围区域以蓝色显示,说明这些区域特征对模型决策影响较小甚至产生反向影响。而在声带囊肿和正常类别中,模型关注区域呈现部分重叠且分布特征差异化的特点。

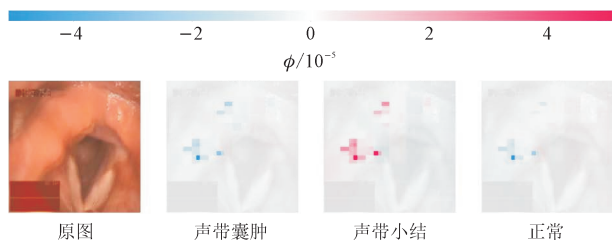


图8 SHAP可视化结果(声带小结)

Fig. 8 SHAP visualization results (vocal nodules)

4 结 论

为实现 MCA 特征学习与 EfficientNet 全局表征能力的深度协同,本文构建了 MCAF-Net 融合网络以提升咽喉疾病影像的良恶性鉴别精度。通过数据增强、模型结构冻结策略及训练集重排方法,有效缓解样本分布不均衡引发的过拟合风险。实验结果表明:该网络在咽喉病理图像分类任务中取得 96.04% 的准确率,特征可视化分析揭示了融合网络的决策机制具备显著可解释性,关键病理区域激活强度的分布与临床诊断标准保持高度一致。然而,数据量的限制与类别分布的不均衡是本研究的重要局限性之一。采用按病例划分的方式构建训练集与测试集,导致测试集中正常类别仅包含 3 个案例,尽管这一分布反映了真实临床数据的自然不均衡,但确实限制了该类别性能评估的统计可靠性。因此,正常类别的评价结果需谨慎解读,其性能并不能完全代表模型在包含更大规模正常类别数据集上的泛化能力。本文建立的深度学习框架为咽喉疾病计算机辅助诊断提供了有效解决方案,具有显著的临床实用价值。

参考文献:

- [1] 付嘉,李丽娟,闫燕,等. 深度学习辅助电子喉镜诊断喉白斑的应用研究[J]. 临床耳鼻咽喉头颈外科杂志, 2021, 35(5): 464-467.
FU Jia, LI Lijuan, YAN Yan, et al. Application of deep learning assisted electronic laryngoscope in diagnosis of laryngeal leukoplakia [J]. Journal of Clinical
- Otorhinolaryngology Head and Neck Surgery, 2021, 35(5): 464-467.
- [2] 马梓博,米悦,张波,等. 面向异质性医学图像处理的深度学习算法综述[J]. 软件学报, 2023, 34(10): 4870-4915.
MA Zibo, MI Yue, ZHANG Bo, et al. Review on deep learning algorithms for heterogeneous medical image processing [J]. Journal of Software, 2023, 34(10): 4870-4915.
- [3] LI Yun, GU Wenxin, YUE Huijun, et al. Real-time detection of laryngopharyngeal cancer using an artificial intelligence-assisted system with multimodal data [J]. Journal of Translational Medicine, 2023, 21: 698.
- [4] MA Haomiao, WEI Wenfeng, ZHANG Junhong, et al. Multi-channel deep learning radiomics model based on contrast-enhanced CT for predicting postoperative prognosis in laryngeal carcinoma [J]. BMC Cancer, 2025, 25(1): 1597.
- [5] YANG K O, KIM S Y, KANG C W, et al. Diagnosis of unilateral vocal fold paralysis using auto-diagnostic deep learning model [J]. Scientific Reports, 2025, 15(1): 27635.
- [6] WANG Haiyang, MAINARDI L. CTGAN in augmentation of radiomics features classification from narrow band imaging for laryngeal cancer [C]//2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA). Piscataway, NJ, USA: IEEE, 2024: 1-5.
- [7] VETHASRI V, DEVI AKALYA B, SHRIRAM P M, et al. Deep learning-based classification of laryngeal abnormalities using DenseNet architectures [C]//2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). Piscataway, NJ, USA: IEEE, 2025: 1235-1239.
- [8] PATIL S, KIRANGE D. Ensemble of deep learning models for brain tumor detection [J]. Procedia Computer Science, 2023, 218: 2468-2479.
- [9] SINGH R, GUPTA S, BHARANY S, et al. Ensemble deep learning models for enhanced brain tumor classification by leveraging ResNet50 and EfficientNet-B7 on high-resolution MRI images [J]. IEEE Access, 2024, 12: 178623-178641.
- [10] THAKUR S, SHARMA S. Ensemble fusion: skin cancer detection using ResNet, EfficientNet, and VGG architectures [C]//2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). Piscataway, NJ, USA: IEEE,

- 2024; 1-8.
- [11] DUDEJA T, DUBEY S K, BHATT A K. Multinomial classification of CT-MRI image retrieval by optimizing EffResNet-4 architecture in deep neural models [J]. International Journal of System Assurance Engineering and Management, 2024, 15(8): 3971-3987.
- [12] AL-AZZWI Z H N, NAZAROV A N. Brain tumor classification based on improved stacked ensemble deep learning methods [J]. Asian Pacific Journal of Cancer Prevention; APJCP, 2023, 24(6): 2141-2148.
- [13] BHATTACHARJEE R, SUGANYA DEVI K, VIJAYKANTH S. Detecting laryngeal cancer lesions from endoscopy images using deep ensemble model [C]//2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IconSCEPT). Piscataway, NJ, USA; IEEE, 2023; 1-6.
- [14] 张龙, 胡燕青, 赵丽娟, 等. 采用递归图编码技术与残差网络的滚动轴承故障诊断 [J]. 西安交通大学学报, 2023, 57(2): 110-120.
- ZHANG Long, HU Yanqing, ZHAO Lijuan, et al. Fault diagnosis of rolling bearings using recurrence plot coding technique and residual network [J]. Journal of Xi'an Jiaotong University, 2023, 57(2): 110-120.
- [15] GUO M H, LU C Z, HOU Q, et al. Segnext: rethinking convolutional attention design for semantic segmentation [J]. Advances in Neural Information Processing Systems, 2022, 35: 1140-1156.
- [16] SHAO Hao, ZENG Quansheng, HOU Qibin, et al. MCANet: medical image segmentation with multi-scale cross-axis attention [J]. Machine Intelligence Research, 2025, 22(3): 437-451.
- [17] TAN Mingxing, LE W. EfficientNet: rethinking model scaling for convolutional neural networks [C]//Proceedings of the 36th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2019; 6105-6114.
- [18] HAO Jie, LIU Yangrui, MO Zhiqiang, et al. Integrated multi-omics and whole slide images for survival prediction in glioblastoma using multiple instance learning and co-attention [C]//2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Piscataway, NJ, USA; IEEE, 2024; 1-4.
- [19] 赵小明, 廖越辉, 张石清, 等. 融合局部与全局特征的 DCE-MRI 乳腺肿瘤良恶分类 [J]. 中国图象图形学报, 2024, 29(1): 256-267.
- ZHAO Xiaoming, LIAO Yuehui, ZHANG Shiqing, et al. Method of classifying benign and malignant breast tumors by DCE-MRI incorporating local and global features [J]. Journal of Image and Graphics, 2024, 29(1): 256-267.
- [20] MEHTA S, RASTEGARI M. Mobil eViT: lightweight, general-purpose, and mobile-friendly vision transformer [EB/OL]. (2022-03-04) [2025-09-04]. <https://arxiv.org/abs/2110.02178>.
- [21] 李广丽, 叶艺源, 吴光庭, 等. 联合多视角 Transformer 编码与在线融合互学习的乳腺癌病理图像分类模型 [J]. 电子学报, 2024, 52(7): 2369-2381.
- LI Guangli, YE Yiyuan, WU Guangting, et al. Breast cancer pathological image classification model via combining multi-view transformer coding and online fusion mutual learning [J]. Acta Electronica Sinica, 2024, 52(7): 2369-2381.
- [22] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA; IEEE, 2017; 2261-2269.
- [23] HOWARD A G, ZHU Menglong, CHEN Bo, et al. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17) [2025-08-24]. <https://arxiv.org/abs/1704.04861>.
- [24] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA; IEEE, 2018; 7132-7141.
- [25] CAI Xinhao, LAI Qiuxia, WANG Yuwei, et al. Poly kernel inception network for remote sensing detection [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA; IEEE, 2024; 27706-27716.

(编辑 刘懿莹)