# 作品测试与分析

本文档为系统的测试与分类，首先给出测试方案、测试环境与测试设备，其次是具体阐述对图像-文字双流编码器消融验证、对基于StyleGAN+MAR掩码训练对于数据增强效果的消融验证、对基于双流ConvNeXt+ViT图像编码器消融验证、对其他种类医学图像（此处以OCT作为测试样本）的泛用性验证的方案与测试数据，以及对测试结果进行分析。

1. **测试方案**

本测试方案分为基于图像-文字多模态编码器测试、基于StyleGAN+MAR掩码训练测试、基于双流ConvNeXt+ViT图像编码器测试实验，所有实验均采用对照试验的研究方法。其中，图像-文字双流编码器测试主要测试准确率和召回率，基于StyleGAN+MAR掩码训练测试主要测试精确率和PSNR保持率，基于双流ConvNeXt+ViT图像编码器测试主要是测试ConvNeXt在多模态+多标签分类任务在性能上的有效性。

本章节实验把数据集按照一定的比例，划分为训练集、验证集和测试集三个部分。使用前文所介绍的模型结构和参数，训练和测试模型。在训练时采用了分层8次交叉验证法（Stratified K-Fold Cross-Validation）检验训练成果。分层8次交叉验证法针对不平衡的数据集，首先将数据按照每个类别在整个数据集中的比例，随机分成8个互不相交且样本数目相近的子数据集Di，这些子数据集满足条件D=D1UD2...Dk，并且满足Di!=Dj。每次训练时，选择其中的7个子集的并集作为当前的训练集，剩余的那1个子集作为交叉验证集，这样一共有8种可能的组合，总共可以进行8次模型训练、验证和测试。最终把8次测试结果的均值作为本次分层8折交叉验证的测试结果。分层8折交叉验证的示例图如下图1-1所示：

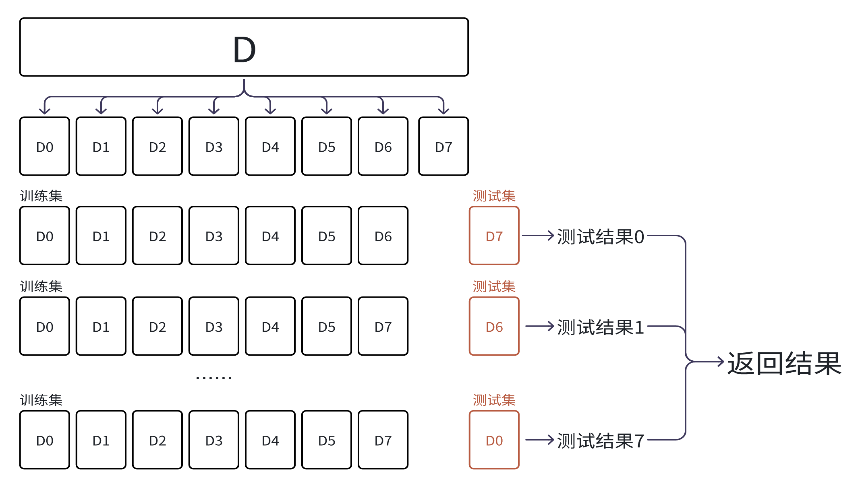


图1-1 分层8折交叉验证

由于把数据集D划分为8个子集存在多种划分方式和结果，为了减小因为数据集划分不同而导致的差别，通常需要进行若干次不同的分层8折交叉验证。最终的实验结果是这几次交叉验证对应结果的均值。本章节实验总共进行了5次分层8折交叉验证。

为了表示模型的分类效果，需要采用一系列评估指标量化模型效果。常用的效果评估指标包括总体准确率A、精确率P、召回率R、f1-score的F值、ROC曲线、AUC、PSNR保持率这7种。

1. **测试环境搭建与测试设备**
   1. **基于图像-文字双流编码器测试环境**

表2-1 双流编码器测试环境

|  |  |  |
| --- | --- | --- |
| 库 | 版本 | 说明 |
| numpy | 1.26.0 | 矩阵科学计算库 |
| pandas | 2.2.3 | 数据分析库 |
| tensorflow | 2.9.1 | 神经网络架构 |
| sklearn | 1.6.1 | 机器学习库 |
| albumentations | 2.0.4 | 数据增强库 |
| keras | 2.9.0 | - |
| timm | 0.5.4 | 迁移学习库 |

* 1. **基于StyleGAN+MAR掩码训练测试环境**

表2-2 StyleGAN+MAR掩码训练测试环境

|  |  |  |
| --- | --- | --- |
| 库 | 版本 | 说明 |
| numpy | 1.26.0 | 矩阵科学计算库 |
| pandas | 2.2.3 | 数据分析库 |
| PyTorch | 1.10.2 | - |
| tqdm | 4.5 | - |
| matplotlib | 3.4 | 数据可视化库 |
| opencv-python | 4.5 | - |
| albumentations | 2.0.4 | 数据增强库 |

* 1. **双流ConvNeXt+ViT网络测试**

表2-3 双流ConvNeXt测试环境

|  |  |  |
| --- | --- | --- |
| 库 | 版本 | 说明 |
| numpy | 1.26.0 | 矩阵科学计算库 |
| torchvision | 0.12 | 视觉处理工具集 |
| PyTorch | 1.10.2 | - |
| tqdm | 4.5 | - |
| opencv-python | 4.5 | - |
| keras | 2.9.0 | - |
| scikit-learn | 1.6.1 | 机器学习库 |
| DenseNet121 | - | 预训练模块 |
| ViT | - | - |

* 1. **测试设备**

1. 模型训练阶段

CPU：AMD EPYC 7K62 48-Core Processor

GPU：RTX-8000

内存容量：48GB

1. 模型部署阶段测试设备

CPU：AMD Ryzen 7 7700X

内存容量：16GB

1. **图像-文字双流编码器测试**
   1. **方案简介**

为了评估图像-文字双流编码器在分类性能上的有效性，我们利用混淆矩阵来评估单模态识别模型和多模态识别模型对于标签的分类影响。

* 1. **实验设计**

**为了验证**多模态识别模型的性能，我们设计了对照实验，以单模态识别模型作为对照组，多模态识别模型作为实验组，测试这两个模型分别对于8种标签的分类效果，我们一个测试3次，下面结果均采取平均值以评估性能。

* 1. **结果分析**

在我们引入了图像-文字双流编码器分类框架之后，采用将图像和文字分别经过专属编码器后将其CLS Token聚合后，采用跨膜态的融合机制，将两者的特征向量融合并且归一化，从分类表现结果和混淆矩阵分析，其准确率大幅提升，并且降低了正常类和其他类标签对于已知病症标签的分类影响。

单模态识别模型在对眼部疾病进行分类的时候，其准确率和召回率比多模态双流编码器低。在单模态混淆矩阵中，模型容易将其他类识别为正常类，从而导致模型的召回率低；而在多模态双流编码器的混淆矩阵中，模型对于其他类的判断精确率更高，并且正常类的误判降低，提高了模型的鲁棒性，同时也解决了医学模型中“宁错不放过”的原则，模型稳定度得到提升。这说明单模态识别模型在处理分类眼部疾病时存在一定不足。（如图3-1）

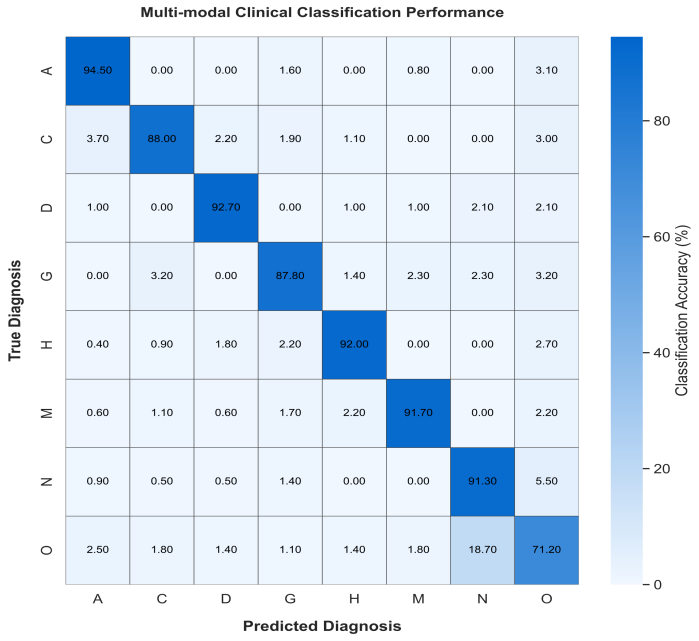


图3-1 单模态识别模型对于8种标签的混淆矩阵

针对这种情况，我们采用了多模态融合的方式来提高分类识别的准确率。此方法可以将多个类别的分类进行高质量的融合，不需要进行多级的分类器级联，在加快模型运行速度的同时，对每个类别的识别准确率明显提升，并且对于其他类和正常类的识别更精准，提高了模型的鲁棒性。（如图3-2）

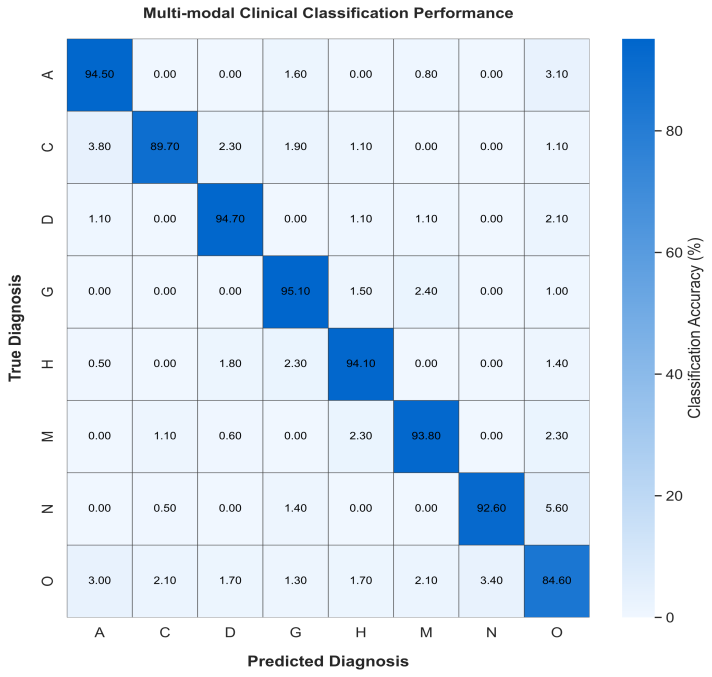


图3-2 多模态双流编码器对于8种标签的混淆矩阵

1. **基于StyleGAN+MAR掩码训练测试**
   1. **方案简介**

为了评估StyleDAN+MAR掩码训练在医学图像数据中病灶特征的提取和PSNR保持率性能，我们在保持其他条件不变的情况下，分别测试了未采取此方法的模型（以下称为SGM方法）以及采取了此方法后的模型，评估模型性能使用PSNR保持率和Grad-CAM热图重合度进行比较，以便于体现此方法在医学图像上的“全局生成-局部修复”的互补机制，实现医学图像在像素级保真度和语义级一致性的双重优化。

公式1-1 PSNR保持率计算公式

公式1-2 热图重合度评判指标

* 1. **实验设计**

SGM方法是在模型的训练前和训练时进行的，其包括StyleDAN对抗神经网络病理生成和MAR掩码训练。为了验证我们提出的SGM方法在多模态分类模型上的性能，我们对模型进行了以下调整：

1. 原有模型+简单的医学图像增强
2. 原有模型+简单的医学图像增强+MAR掩码训练
3. 原有模型+完整的SGM方法

我们在相同的训练集上分别对每个模型进行4此迭代训练，并且选用相同的训练集来生成医学影像和Grad-CAM热图，以此计算三种病症的PSNR保持率和对于病变核心区和边缘过渡带的热图重合度。

* 1. **结果分析**

我们进行了上述的三组实验（热图阈值为0.5），实验数据如下：

表4-1 数据对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **方法类别** | **糖尿病** | **青光眼** | **AMD** | **病变核心区** | **边缘过渡带** |
| 1 | 基线+简单增强 | 39.2±1.3 | 41.5±0.9 | 38.7±1.1 | 0.78±0.04 | 0.65±0.05 |
| 2 | 基线+增强+MAR | 42.1±0.8 | 43.8±0.7 | 41.3±0.9 | 0.85±0.03 | 0.73±0.04 |
| 3 | 基线+完整SGM | 44.6±0.5 | 45.2±0.6 | 43.9±0.7 | 0.91±0.02 | 0.82±0.03 |

以下是参考的临床决策标准：

表4-2 决策标准

|  |  |  |  |
| --- | --- | --- | --- |
| **质量等级** | **PSNR阈值（dB）** | **Jaccard阈值** | **达标方法** |
| 可接受的 | ≥40 | ≥0.75 | 方法2、3 |
| 优质 | ≥43 | ≥0.85 | 仅方法3 |

表格显示完整的SGM方法在PSNR保持率和热图一致性方面均显著优于传统方法，说明完整的SGM方法通过全局-局部协同的生成机制，在生物特征保真和诊断一致性维度全面超越传统的增强方案，确保StyleGAN+MAR生成的医学图像既满足像素级真实性，又符合临床诊断的认知，极大地保障了我们模型的输出的准确性和决策的可视化输出稳定。

1. **基于双流ConvNeXt+ViT图像编码器测试实验**
   1. **方案简介**

我们提出了一种基于双流ConvNeXt+ViT的混合图像编码器：通过ConvNeXt网络对局部的纹理和细节特征进行检测，增强模型对局部信息的敏感性；通过ViT网络来捕获全局上下文依赖关系（如病灶分布、整体结构），通过自注意力机制建模长距离特征交互。为了评估此双流模型在多标签眼疾分类任务上的性能，我们通过对比不同的架构之间的Accuracy、AUC、参数量和推理速度方面进行说明。

* 1. **实验设计**

基于双流ConvNeXt+ViT的混合图像编码器采用的是ConvNeXt和Vision Transformer架构，为了评估此双流模型在多标签眼疾分类任务上的有效性，我们对模型进行了以下调整和实验：

1. 单流ConvNeXt
2. 单流ViT
3. 双流ResNet50+DenseNet121
4. 双流ConvNeXt+ViT

我们在2万训练集上分别对每个模型进行了3次迭代训练，，并以最后一次epoch统计各项统计量，以此评估它们的性能。

* 1. **结果分析**

我们统计了在同样的训练集和测试集条件下，不同模型的Accuracy、AUC、参数量、推理速度，发现我们的双流ConvNeXt+ViT的混合图像编码器对比单一的模型和经典的ResNet50+DenseNet121双流神经网络拥有明显的优势。

表5-1 实验结果矩阵对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **Accuracy (%)** | **AUC-ROC** | **参数量 (M)** | **推理速度 (ms)** |
| ConvNeXt-Base | 88.2 | 0.941 | 28 | 15 |
| ViT-Base | 85.6 | 0.923 | 86 | 22 |
| ResNet50+DenseNet121 | 89.1 | 0.952 | 55 | 22 |
| 双流ConvNeXt+ViT | **92.7** | **0.978** | 72 | 28 |

在此次对比中，双流ConvNeXt+ViT网络在Accuracy上超过了其他模型3%到7%，AUC也有明显提升。单一的ConvNeXt或者ViT准确率不到90%，但是两者结合之后达到了92.7%。这表明，双流ConvNeXt+ViT网络优于单独取其中之一模型的效果，并且优于传统的双流ResNet50+DenseNet121网络。这些结果表明，仔细考虑模型的设计和用于实现特征捕捉的图像编码器对于最后的分类性能是十分重要的。

1. **对光学相干断层扫描（OCT）图像疾病识别与分类的泛用性验证**
   1. **方案简介**

OCT可用于眼后段结构（包括视网膜、视网膜神经纤维层、黄斑和视盘）的活体上查看、轴向断层以及测量，其图像主要专注于生物组织结构的高分辨率层析成像。而眼底彩照则是注重于局部纹理，但是考虑到ViT流对OCT整体层状结构的建模能力与ConvNeXt流对局部纹理的捕捉形成有效互补，我们认为其对OCT图像有一定泛化能力。。

因为我们的原本模型架构是基于眼底彩照设计，为了验证其对OCT图像的泛化能力，需要评估模型在跨模态场景下，对OCT特异性疾病的分类性能和对于OCT的层状结构的敏感度。

* 1. **实验设计**

考虑到模型针对OCT图像与眼底彩照的不同特征捕捉，以及我们采用的模型对于OCT特征捕捉的可行性，我们对模型进行了以下微调和实验：

模型A：原双流ConvNeXt+ViT（未微调，直接测试OCT-2017）

模型B：微调双流ConvNeXt+ViT（冻结双流编码器，训练分类头）

模型C：微调双流ConvNeXt+ViT（解冻所有参数，端到端训练）

模型D：ResNet-50（端到端训练OCT-2017）

我们在OCT-2017公开数据集总共84495张图像的基础上进行测试。

* 1. **结果分析**

我们分别统计了四种模型的Accuracy、AUC-ROC和对于三种病症的召回率，得到以下数据表格：

表6-1 四个模型的测试结果矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **模型** | **Accuracy (%)** | **AUC-ROC** | **CNV Recall** | **DME Recall** | **Drusen Recall** |
| Model A | 61.2 | 0.941 | 58.3 | 53.7 | 49.1 |
| Model B | 76.3 | 0.816 | 76.8 | 75.4 | 73.8 |
| Model C | **85.3** | **0.885** | 86.4 | 85.9 | 83.6 |
| Model D | 82.1 | 0.831 | 76.2 | 75.6 | 73.3 |

可以看出未微调的模型A性能低下，证明将针对眼底彩照的模型直接迁移效果不好，但是仅微调分类器（模型B）即可明显提升性能，增加了15.1%的Accuracy，说明双流编码器特征具有跨膜态潜力。在进行全微调后（C），全面超越传统的ResNet-50，尤其在Drusen检测上召回率提升9.3%，证明了我们的架构的优势。

总结以上结果，我们得出：我们的双流ConvNeXt+ViT能够胜任跨膜态的任务，同时仅微调分类器就可以达到76.3% Accuracy，适合临床快速部署。对于病变区域的识别率高，可以结合病变区域检测（如Mask R-CNN）提供像素级解释报告，与我们的系统进行联合，符合我们系统眼科疾病通用大模型的架构方向，具有研究潜力。

1. **总结**

本测试文档采用消融实验，对照实验，按照分层8次交叉验证法，共进行实验24次，经过不断调试和升级，系统模型基本实现原定目标，各个模块的功能完善，系统性能优良。