Modeling and Simulation 建模与仿真, 2025, 14(4), 50-60
Published Online April 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.144264

Hans汉斯

SC-SENet：一种基于阴道镜图像的宫颈上皮内瘤变分类诊断模型

徐志扬[1]，谷雪莲[1*]，管　睿[2]，邹任玲[1]，楚胜轩[1]，方庆斌[1]

[1]上海理工大学健康科学与工程学院，上海
[2]海军军医大学第一附属医院长海医院妇产科，上海

收稿日期：2025年3月2日；录用日期：2025年3月25日；发布日期：2025年4月7日

摘　要

宫颈上皮内瘤变(Cervical Intraepithelial Neoplasia, CIN)是宫颈浸润癌的癌前病变阶段。为了提高诊断的效率与准确性，本研究使用深度学习技术对阴道镜图像进行三分类识别，搭建了一个结合了Swin Transformer与卷积神经网络(CNN)的分类诊断模型Swin-Conv Squeeze-and-Excitation Network (SC-SENet)。该模型包括Swin Transformer分支与结合了通道注意力机制的卷积分支，通过并联的方式结合分别用于图像全局特征与局部特征的提取，结合两种特征实现病灶区域信息的特征识别以提高诊断准确度。本研究使用临床采集的阴道镜图像作为原始数据集，SC-SENet模型在CIN三分类诊断任务的总体准确率为90.81%。实验结果表明，本研究提出SC-SENet模型能有效诊断宫颈上皮内瘤变的病变等级。

关键词

深度学习，宫颈上皮内瘤变，阴道镜图像，卷积神经网络，Swin Transformer

SC-SENet: A Classification Diagnosis Model for Cervical Intraepithelial Neoplasia Based on Colposcopic Images

Zhiyang Xu[1], Xuelian Gu[1*], Rui Guan[2], Renling Zou[1], Shengxuan Chu[1], Qingbin Fang[1]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai
[2]Department of Gynecology and Obstetrics, Changhai Hospital, The First Affiliated Hospital of Naval Medical University, Shanghai

Received: Mar. 2nd, 2025; accepted: Mar. 25th, 2025; published: Apr. 7th, 2025

*通讯作者。

文章引用: 徐志扬, 谷雪莲, 管睿, 邹任玲, 楚胜轩, 方庆斌. SC-SENet: 一种基于阴道镜图像的宫颈上皮内瘤变分类诊断模型[J]. 建模与仿真, 2025, 14(4): 50-60. DOI: 10.12677/mos.2025.144264

Abstract

Cervical Intraepithelial Neoplasia (CIN) is a precancerous stage of invasive cervical cancer. In order to improve the efficiency and accuracy of diagnosis, this study used deep learning technology to classify colposcopy images into three categories. In this paper, a classification diagnosis model Swin-Conv Squeeze-and-Excitation Network (SC-SENet) combining Swin Transformer and Convolutional neural Network (CNN) is built. The model includes the Swin Transformer branch and the convolution branch combined with the channel attention mechanism, which are combined in parallel to extract the global features and local features of the image respectively. The two features are combined to realize the feature recognition of the lesion area information to improve the diagnostic accuracy. In this study, clinically collected colposcopy images were used as the original data set, and the overall accuracy of SC-SENet model in CIN three-classification diagnosis task was 90.81%. The experimental results show that the SC-SENet model proposed in this study can effectively diagnose the lesion grade of cervical intraepithelial neoplasia.

Keywords

Deep Learning, Cervical Intraepithelial Neoplasia, Colposcopic Images, CNN, Swin Transformer

1. 引言

宫颈上皮内瘤变(Cervical Intraepithelial Neoplasia, CIN)是宫颈病变发展的重要病理阶段，病变的不同阶段反映了细胞增生、核异型性及其向浸润性宫颈癌演变的潜力[1]，及时诊断并治疗 CIN 有够有效预防宫颈癌的发生[2]。根据其病变程度，由低到高可以划分为 CIN1、CIN2 及 CIN3 三种等级，CIN1 也称为低级别鳞状上皮内病变(LSIL)，CIN2/3 称为高级别鳞状上皮内病变(HSIL) [3]。阴道镜检查是临床广泛应用于 CIN 诊断手段之一，医师可依据阴道镜图像进行病变评估[2]。然而，因为阴道镜图像的数据量通常较小，且易受到设备、操作者的拍摄习惯、照明条件以及患者生理情况等因素的影响，诊断结果受限于医师的专业经验与技能水平，故存在较大的主观性和不确定性[4]。因此，基于医学图像任务构建深度学习分类模型进行 CIN 辅助诊断研究，将有助于提高医生的工作效率，增强诊断的准确性，为临床应用提供理论依据[5]-[8]。

目前，已有学者使用阴道镜图像来训练深度学习网络模型，以实现 CIN 的计算机辅助诊断[9]-[15]，但针对 CIN 的研究集中于 LSIL 和 HSIL 的二分类诊断。Tao Zhang 等基于预训练的密集连接卷积网络 DenseNet，采用迁移学习技术对来自 ImageNet 和 Kaggle 的两个预训练模型进行了微调，以是否含有 CIN2 及 CIN2+为分类标准，达到了 73.08%的准确率[9]。Miyagi 等利用 310 名患者的阴道镜图像数据构建了一个 11 层的卷积神经网络模型，该模型对 LSIL 和 HSIL+的分类准确率为 82.3% [10]。Saini S K 等使用 NCI 公共数据集进行模型训练，搭建了一种高效的模型 ColpoNet，将宫颈炎症与 CIN1 归为一类，CIN2+归为一类，模型的准确率达到 81.35% [11]。Lu Liu 等将残差神经网络(ResNet)与临床特征结合构建一个综合模型，专注于两个二分类任务：一般炎症与低级别鳞状上皮内病变或更严重(LSIL+)之间的分类，以及高级别鳞状上皮内病变(HSIL)与 HSIL+之间的分类，准确率分别为 88.6%和 80.7% [12]。主要原因是阴道镜

图像数据普遍存在噪声干扰严重、数据类别不均衡等问题[16] [17]，使用常见的神经网络难以高效地提取CIN 各个等级的病灶特征并识别。

为了解决诊断准确率低、分类精度低等问题，本研究基于阴道镜图像构建了一种适用于 CIN 三分类诊断的网络模型 Swin-Conv Squeeze-and-Excitation Network (SC-SENet)。模型采用了多尺度特征融合的方法，将 Swin Transformer 分支与带有通道注意力机制的卷积网络分支并联，充分结合全局与局部特征，并且使用 Focal loss 损失函数等策略以解决类别不平衡和模型过拟合等问题。实验结果表明，SC-SENet 模型在阴道镜图像的 CIN 三分类任务中取得了良好的表现。

2. SC-SENet 模型的构建

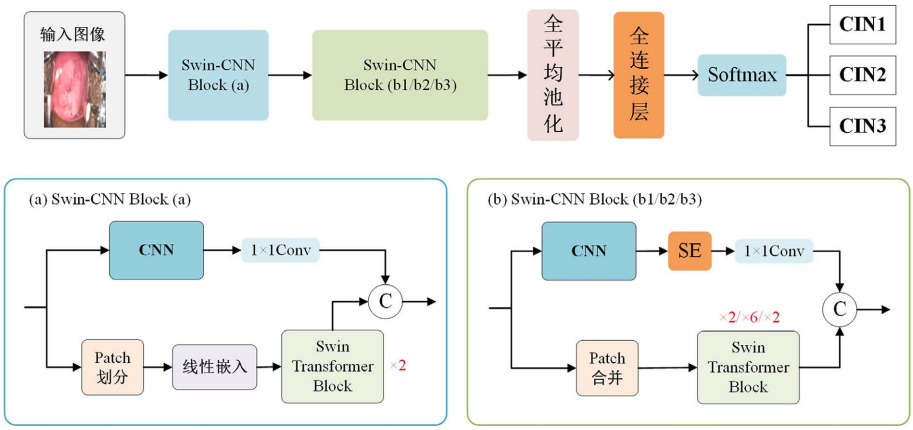

Figure 1. SC-SENet model architecture and module details
图 1. SC-SENet 模型具体架构及模块细节

SC-SENet 模型的主体架构主要由 Swin-CNN Blcok(a)、三个结构相似的 Swin-CNN Blcok(b)、全平均池化层、全连接层以及 Softmax 层构成。其中 Swin-CNN Block 为模型的核心部分，负责输入图像的多尺度特征提取及融合工作，不同 Swin-CNN Block 的结构有一定的细节差异，根据结构可将其划分为 Block(a)与 Block(b)。该模块采用了分支并联的方式，选择 Swin Transformer 作为图像全局特征提取的分支网络，另一个分支为结合了通道注意力机制的卷积神经网络用于图像局部特征提取，将两个分支的特征以逐元素相加的方式进行融合，既能够更好的对宫颈图像的复杂特征进行处理提取，增加分类准确度，还能防止模型计算参数的增加，保证运算效率。在通过多个 Swin-CNN Block 进行特征提取之后，图像的高维特征图会被送入全平均池化层。该层对每个特征图进行全局平均池化操作，将每个特征图的空间维度压缩为单一的平均值生成一个固定长度的特征向量，有效地减少了模型的参数量和计算复杂度。全连接层将该向量映射到不同的类别空间，模型能够根据提取到的特征作出分类决策，并学习各类别的权重关系，进一步优化模型的分类能力。最后，经过全连接层输出的结果会输入至 Softmax 层，Softmax 函数会将其转化为概率分布，表示每个类别的预测概率。SC-SENet 模型的具体架构及 Swin-CNN Block 的细节信息如图 1 所示。

2.1. Swin-CNN Block(a)

在 Swin-CNN Block(a)中，Swin Transformer 分支由一个 Patch 划分层、一个线性嵌入层以及两个 Swin Transformer Block 构成[18]。Patch 划分层选取不同的 Patch 大小将图像分割成固定大小的块，通过线性嵌入层将这些块在通道维度展平，映射为向量。Swin Transformer Block 则进行图像的全局特征提取，采

用了移动窗口机制计算自注意力，提高了模型的计算效率。在每一对 Swin Transformer Block 中，除了归一化处理层 LN 与非线性特征变换层 MLP 之外，分别由窗口多头自注意力机制(Windows Multi-Head Self-Attention, W-MSA)与位移窗口多头自注意力机制(Shifted Windows Multi-Head Self-Attention, SW-MSA)组成。两种不同的自注意力模块成对使用，所以 Swin Transformer Block 的数量也都是偶数，结构如图 2 所示。W-MSA 是一种添加了窗口机制的多头自注意力模块，能有效减少计算量，该部分在进行自注意力计算时，将已经分割好的不同 patch 进一步划分为多个窗口，在每个窗口内部进行自注意力计算，减少了模型计算过程中的参数数量，普通 MSA 与 W-MSA 的参数量计算公式分别为公式(1)与公式(2)。其中 h 和 w 代表每个 patch 的高度与宽度，C 代表深度，M 代表 W-MSA 中的窗口大小，δ 代表参数数量。

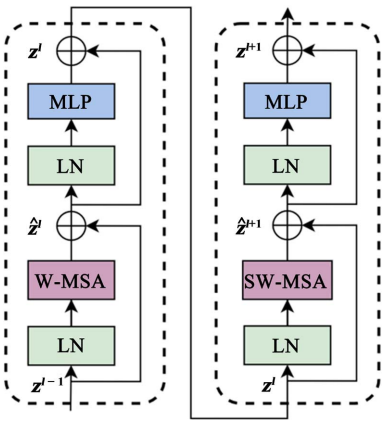

Figure 2. Diagram of two consecutive Swin Transformer Block structures
图 2. 两个连续的 Swin Transformer Block 结构图

$$\delta(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2 C \tag{1}$$

$$\delta(W-\mathrm{MSA}) = 4hwC^2 + 2(M)^2 hwC \tag{2}$$

采用 W-MSA 模块时，只会在每个窗口内进行自注意力计算，窗口与窗口之间是无法进行信息传递的，在 W-MSA 后添加 SW-MSA 则可以解决这个问题。SW-MSA 模块的窗口在原窗口基础上进行偏移，并且新的部分窗口进行位置融合，可以实现在与 W-MSA 不同的窗口内部的信息交互。

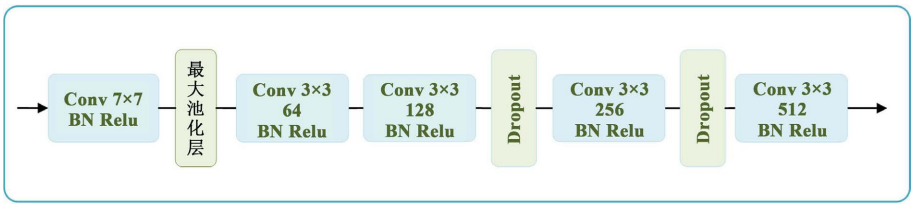

Figure 3. The specific network structure of CNN
图 3. CNN 的具体网络结构

卷积分支由一个卷积神经网络(CNN)以及一个用于通道维度转换的 1 × 1 的卷积构成，CNN 主要承担局部特征提取任务，通过卷积操作增强模型对 CIN 图像中细节特征的感知能力。CNN 包含多个关键组件，包括 7 × 7 大核卷积层、批归一化层(Batch Normalization, BN)、ReLU 激活函数、最大池化层(Max

Pooling)、深层 3 × 3 卷积层和 Dropout 层等，具体网络结构如图 3 所示。在特征提取过程中，输入特征首先通过 7 × 7 大核卷积层进行初步特征提取，并通过 BN 层和 ReLU 激活函数进行数据标准化和非线性变换。随后，特征经过最大池化层对空间尺寸进行下采样，以压缩特征图大小并保留主要信息。接下来使用多个 3 × 3 卷积层、BN 层与 ReLU 激活函数用，于提取更深层次的局部特征，进一步增强模型对 CIN 影像的病变区域细节的捕捉能力。并且在部分卷积层后，引入 Dropout 层随机失效部分神经元以防止过拟合，提高模型的泛化能力[19] [20]。特征提取之后会通过 1 × 1 卷积进行维度转换，使得卷积分支的输出通道数与其他分支保持一致，输出特征能够有效融合。最终，经过逐像素加和后，CNN 分支的局部特征和 Transformer 分支的全局特征得以融合，增强了模型的特征表达能力。

2.2. Swin-CNN Block(b)

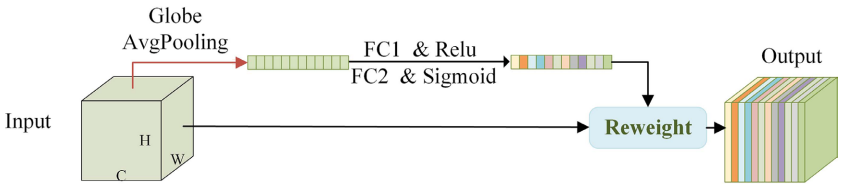

Figure 4. The structure of Squeeze-and-Excitation
图 4. 通道注意力机制模块结构

在 Swin-CNN Block(b)的主干部分，Swin Transformer Block 没有差异，根据其数量的不同，可将 Swin-CNN Block 分为 Block(b1)、(b2)和(b3)。在 Swin Transformer Block 特征提取之前，使用了 Patch 合并模块，类似于卷积网络中的池化层，该模块会对输入的特征图像进行下采样操作，之后特征图的高和宽会减半，深度会翻倍。在卷积分支部分，使用与 Swin-CNN Block(a)中相同的卷积模块进行特征提取，该卷积分支还添加了通道注意力机制 Squeeze-and-Excitation(SE) [21]。SE 主要由压缩和激励两部分组成，具体结构如图 4 所示。第一步对大小为 $H \times W \times C$ 的输入特征图进行空间特征压缩，实现全平均池化，得到一个 1 × 1 × C 的向量。使用公式(3)将每个通道的 $H \times W$ 个像素值压缩为一个实数，式中 H、W、C 分别为图像的长宽及通道数，Z_c 是通道的全局特征，表示了通道 c 上的平均特征值。

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j) \tag{3}$$

第二步进行激励操作，通过两个全连接层 FC_1 及 FC_2 对压缩后的向量 Z 进行通道特征学习，得到通道权重值 S。不同的数值表示不同通道的权重信息，赋予通道不同的权重，得到具有通道注意力的特征向量，维度仍然保持 1 × 1 × C。使用公式(4)实现上述操作，式中的 δ 与 σ 分别代表两个全连接层使用的激活函数 ReLU 与 Sigmoid。

$$S = \sigma\left(FC_2 \delta\left(FC_1 Z\right)\right) \tag{4}$$

将得到的权重系数与输入特征图对应通道上的每个像素相乘，得到最终具有通道注意力的特征图像，其大小仍然为输入时的 $H \times W \times C$。

2.3. 训练过程优化

在网络训练过程中，使用了焦点损失函数(Focal Loss) [22]，用来解决数据集存在的类别不平衡问题。焦点损失函数在交叉熵损失函数(Cross-entropy Loss) [23]的基础上，增加了调节因子以降低易分类样本的权重，使模型更多地关注那些困难样本，从而提高模型在类别不平衡情况下的性能。Focal Loss 的基本公

式和带有 α 权重的公式见公式(5)和公式(6)。

$$FL(p_t) = -(1-p_t)^{\gamma} \cdot \log(p_t) \tag{5}$$

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} \cdot \log(p_t) \tag{6}$$

式中 p_t 是模型预测目标类别的概率；γ 是焦点因子，用于调节易分类样本的权重，当 $\gamma = 0$ 时，焦点损失函数退化为交叉熵损失函数；α_t 是一个用于平衡正负样本的权重因子，通常用于处理类别不平衡的情况。在该模型中，选取的权重因子 $\alpha_t = 0.25$，$\gamma = 2.0$。

学习率是模型中的一个重要超参数[24]。为了更好地训练网络模型，使用了一种阶梯式学习率下降策略，每经过一定的 epoch，学习率参数会进行一次固定倍率的衰减，可以用公式(7)表示。

$$Lr = Lr \times \beta^{\left\lfloor \frac{\text{epoch}}{\text{step-size}} \right\rfloor} \tag{7}$$

式中 Lr 为当前学习率，β 表示衰减因子，即每个阶段的学习率衰减的比例；epoch 表示当前的训练轮数；step-size 表示学习率调整的周期，即经过多少个 epoch 后进行一次学习率的调整；$\lfloor\cdot\rfloor$表示向下取整操作符，表示除法后向下取整。该模型中，各参数设置分别为 $\beta = 0.8$，step-size = 2。此外还设置了最小学习率为 0.00001，即学习率小于该值时不再进行衰减。

3. 实验准备

3.1. 数据集与实验设置

实验使用的阴道镜宫颈图像数据取自上海市某医院妇产科门诊 2019.11~2023.10 共计约 4500 名患者，根据图像的模糊程度、病灶区域可见度等因素，筛选出研究可用图像 2325 例，均为为 6000 × 4000 像素的三通道图像。其中，CIN1、CIN2、CIN3 三种类型数量分别为 1305、449、571 例，按照约 8:1:1 的固定比例将其划分为训练集、验证集和测试集，具体数据信息见表 1。由于数据量较大且单张图像尺寸较高，直接用于训练可能会增加计算负担并降低模型的训练效率。因此，在图像预处理阶段，首先将所有图像统一缩放至 512 × 512 像素，以减少计算复杂度并提高训练效率。此外，数据集中存在多个来自同一患者的图像，其中某些图像的特征高度相似，可能会导致模型过拟合。为缓解这一问题，在训练过程中采用在线数据增强策略，包括随机平移(图像尺寸的 10%)、随机旋转(±15˚)、随机缩放(范围 90%~110%)、水平翻转等方法，以提升模型的泛化能力。通过数据增强策略，可以在保证图像信息不失真的前提下，提升模型的鲁棒性，减少过拟合风险，并提高其在不同数据分布下的适应能力。

整个网络模型代码基于深度学习 PyTorch 框架实现，并在 NVIDIA GeForce RTX 2080Ti GPU (12G) 上执行训练过程。此外，模型超参数如迭代次数、初始学习率、批次大小和梯度优化器分别设置为 150、0.001、4 和 AdamW 优化器，patch 大小为 4，每个 Swin-CNN Block 模块中的窗口大小均为 7，从前到后四个 Swin-CNN Block 中多头自注意力模块 head 的维度大小依次为 3、6、12、24。

Table 1. Details of the dataset used for the experiments
表 1. 实验使用的数据集详细信息

类别	训练集	验证集	测试集	总计
CIN1	1045	130	130	1305
CIN2	359	45	45	449
CIN3	457	57	57	571
总计	1861	232	232	2325

3.2. 评价指标

为了评估所提出的诊断模型的性能，选取了分类模型常用的评估指标，如精确率(Precision)、召回率(Recall)、F1 Score、准确率(Accuracy)和混淆矩阵等[25]。精确率表示被分类器判定为正类别的样本中，实际上是正类别的比例；召回率表示实际为正类别的样本中，被分类器正确预测为正类别的比例；F1 Score是精确率和召回率的调和平均，它综合考虑了分类器的准确性和召回能力；准确率表示被分类器正确分类的样本数量占总样本数量的比例；混淆矩阵则能够表示模型预测的各类具体结果，并且能用混淆矩阵中的数据计算出上述指标。这些指标的定义如下所示：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$F1\text{Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

各式中 TP (True Positives)表示正确预测为正类别的样本数量；FP (False Positives)表示错误预测为正类别的样本数量；TN (True Negatives)表示正确预测为负类别的样本数量；FN (False Negatives)表示错误预测为负类别的样本数量。

4. 实验结果与分析

SC-SENet 模型进行了 150 个轮次的训练，损失值最终趋于收敛，且在验证集达到了较高的准确率。使用训练后得到的最优权重对测试集进行预测，得到分类结果的混淆矩阵。模型评估使用的测试集为预先划分好的样本数据，没有参与模型的训练与验证过程。

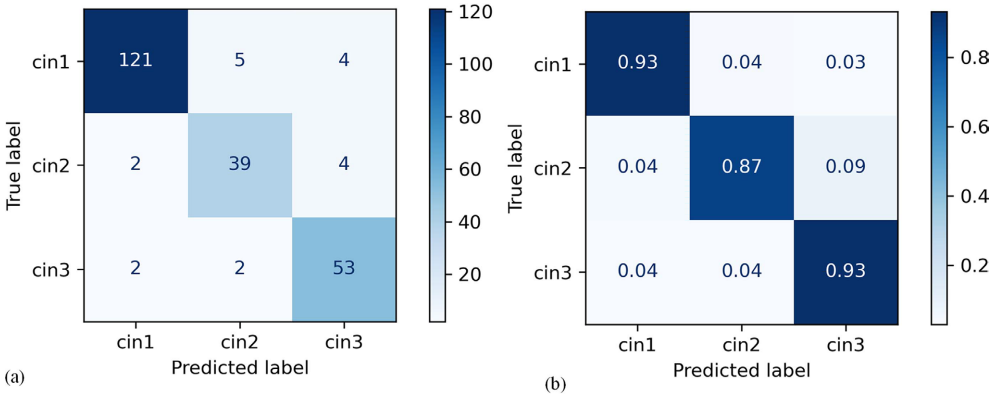

Figure 5. Confusion matrix of model predictions
图 5. 模型预测结果的混淆矩阵

图 5 为 SC-SENet 模型预测结果的混淆矩阵，展示了 CIN1、CIN2 和 CIN3 三种类别在测试集上的具体预测结果，上述提到的不同模型评估指标所需的数值均可通过混淆矩阵计算得到，模型的整体精度为 90.81%。表 2 是三种类别在模型中预测效果的具体参数结果，对比了模型训练使用 Focal Loss 损失函数

与普通交叉熵损失函数的各类别表现差异，Focal Loss 的权重因子设置为$\alpha_t = 0.25$，$\gamma = 2.0$。数据集中三种类别比例不均衡，CIN2 与 CIN3 两类数量明显少于 CIN1，并且 CIN2 类别处于病变中间时期，病灶包含的特征信息与另外两类尤其是 CIN3 相似，属于较难预测的类别。根据表 2 中的数据可以看出，使用 Focal Loss 损失函数，CIN1 预测结果差异不大，而 CIN2 与 CIN3 各参数均有一定提高，在一定程度上改善了这两种类别的预测效果，CIN2 的准确率由 76.57%提高至 85.44%，CIN3 的准确率由 86.69%提高至 91.26%，表明 Focal Loss 对类别不均衡的数据集及具有一定的优化效果。

Table 2. Parameter performance of different loss functions across the three categories
表 2. 不同损失函数在三种类别上的参数表现

损失函数	类别	精确率	召回率	F1	准确率
Focal loss	CIN1	0.9680	0.8897	0.9272	0.9307
	CIN2	0.8578	0.7647	0.8086	0.8544
	CIN3	0.8689	0.8448	0.8518	0.9126
Cross-entropy loss	CIN1	0.9385	0.8677	0.8970	0.9048
	CIN2	0.8104	0.6604	0.7368	0.7657
	CIN3	0.8143	0.7692	0.8034	0.8669

此外，进行了消融实验来验证了该研究提出的创新方法的有效性，SC-SENet 主要优化方法除损失函数外为双特征提取网络并联以及学习率下降策略。因此在使用 Focal loss 损失函数的标准下，实验部分对比了使用了一般学习率的 Swin Transformer 网络、使用了学习率下降策略的 Swin Transformer 网络、使用了一般学习率的并联网络和使用了学习率下降策略的并联网络这四种不同处理方式的结果。

Table 3. Results of model parameters using different optimization methods
表 3. 使用不同优化方法的模型参数结果

方法		精确率	召回率	F1	准确率
Swin Transformer	一般学习率	0.8376	0.8484	0.8389	0.8528
	学习率下降策略	0.8513	0.8575	0.8536	0.8644
并联网络	一般学习率	0.8793	0.8804	0.8746	0.8875
	学习率下降策略	0.8949	0.9091	0.8915	0.9081

表 3 可以看到，使用一般的学习率策略对 Swin Transformer 模型进行训练，最终准确率为 85.28%，使用学习率下降策略对 Swin Transformer 模型进行训练准确率提高至 86.44%，证明了学习率下降策略在该任务中具有一定优化效果。使用一般的学习率策略对并联网络进行训练，最终准确率为 88.75%，相对于 Swin Transformer 有较大的提升，说明将结合了通道注意力机制的卷积网络与 Swin Transformer 并联，实现多尺度特征提取，能够对基于阴道镜图像的 CIN 三分类诊断任务有较好的优化效果，提高分类准确性。使用学习率下降策略对并联网络进行训练后，进一步提高了模型效果，最终达到 90.81%准确率。

为了更加直观地体现出本研究提出方法的优化效果，将 SC-SENet 模型与其他常用的分类模型(VGG, ResNet, ResNeXt, DenseNet, Vision Transformer, Swin Transformer)进行了比较[18] [26]-[30]。均使用相同图像数据集进行模型训练，选择模型最优权重进行对测试集进行预测。

Table 4. Parameter performance of different models in the CIN tri-classification task
表 4. 不同模型在 CIN 三分类任务中的参数表现

模型	精确率	召回率	F1 score	准确率
VGG	0.7374	0.7404	0.7367	0.7629
ResNet	0.7783	0.7805	0.7788	0.8060
ResNeXt	0.7653	0.7856	0.7866	0.7940
DenseNet	0.7454	0.7520	0.7458	0.7716
Vision Transformer	0.8099	0.8234	0.8163	0.8319
Swin Transformer	0.8269	0.8311	0.8289	0.8448
SC-SENet	**0.8949**	**0.9091**	**0.8915**	**0.9081**

表 4 详细展示了不同模型的各个参数结果。可以看出，在阴道镜图像的宫颈上皮内瘤变三分类诊断任务中，Swin Transformer 模型相对于其他模型具有更好的分类效果，本研究在此模型基础上进行进一步优化，将其与添加了带有通道注意力机制的卷积分支并联实现多尺度特征提取，提出的 SC-SENet 模型在该数据集任务中取得了良好的结果，准确率达到 90.81%，明显高于其他模型。

5. 总结与讨论

综上所述，本研究提出并搭建了一个创新的分类网络模型 SC-SENet，旨在提升阴道镜图像中宫颈上皮内瘤变(CIN)三分类诊断的准确性与效率。模型采用了多分支并联的方式，通过将 Swin Transformer 提取的全局特征与传统卷积神经网络(CNN)提取的局部特征相结合，充分利用了两者在不同层次上捕捉图像特征的优势。Swin Transformer 通过其自注意力机制，能够有效地建模图像中的长距离依赖关系，提取全局信息；而 CNN 则在局部特征提取方面表现优越，能够精确地捕捉病灶区域的细微变化。通过这两种特征的互补和融合，SC-SENet 在病灶区域的识别能力上得到了显著提升，从而增强了模型在不同类别(CIN1, CIN2, CIN3)诊断任务中的准确性。

在实验验证过程中，本研究使用了来自临床的阴道镜图像数据集，评估了模型在实际应用中的性能。经过训练与测试，SC-SENet 在 CIN 三分类任务中的准确率达到了 90.81%，这一结果表明该模型具有较强的分类能力，能够准确地区分不同病变级别。与当前领域中其他诊断模型相比，SC-SENet 不仅在最终准确率上有所提升，还能在较为复杂的病变图像中实现更高精度的病变等级诊断。

SC-SENet 的高准确率和较好的病变等级识别能力使其具有广泛的临床应用前景，特别是在大规模筛查和早期检测方面，能够为宫颈病变的早期诊断和干预提供有力支持。通过自动化的诊断系统将 SC-SENet 应用于临床诊断中，能够为医务人员提供更加详细和精确的病变信息，从而帮助医生在诊断过程中做出更为准确的判断。此外，本研究中的网络架构为其他医学图像分类任务提供了一定的参考，未来可以考虑将这一框架推广到其他疾病的早期检测和诊断中，进一步扩充人工智能在医学影像分析领域的应用。

伦理声明

本研究使用了完全匿名的过往阴道镜图像，不包含任何个人身份信息且对患者无风险，这些数据由[海军军医大学第一附属医院长海医院]提供。根据[海军军医大学第一附属医院长海医院]伦理委员会的意见，本研究无需伦理审批，符合伦理原则。

参考文献

[1] Mattiuzzi, C. and Lippi, G. (2019) Cancer Statistics: A Comparison between World Health Organization (WHO) and Global Burden of Disease (GBD). *European Journal of Public Health*, **30**, 1026-1027. https://doi.org/10.1093/eurpub/ckz216

[2] Rowland, A.G. and Brady, J. (2020) Colposcopy and Cervical Intraepithelial Neoplasia. *Obstetrics, Gynaecology & Reproductive Medicine*, **30**, 133-138. https://doi.org/10.1016/j.ogrm.2020.02.008

[3] Waghe, T. and Acharya, N. (2024) Advancements in the Management of Cervical Intraepithelial Neoplasia: A Comprehensive Review. *Cureus*, **16**, e58645. https://doi.org/10.7759/cureus.58645

[4] Xue, P., Ng, M.T.A. and Qiao, Y. (2020) The Challenges of Colposcopy for Cervical Cancer Screening in LMICs and Solutions by Artificial Intelligence. *BMC Medicine*, **18**, Article No. 169. https://doi.org/10.1186/s12916-020-01613-x

[5] Dong, S., Wang, P. and Abbas, K. (2021) A Survey on Deep Learning and Its Applications. *Computer Science Review*, **40**, Article 100379. https://doi.org/10.1016/j.cosrev.2021.100379

[6] Kather, J.N., Weis, C., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., *et al*. (2016) Multi-Class Texture Analysis in Colorectal Cancer Histology. *Scientific Reports*, **6**, Article No. 27988. https://doi.org/10.1038/srep27988

[7] Lundervold, A.S. and Lundervold, A. (2019) An Overview of Deep Learning in Medical Imaging Focusing on MRI. *Zeitschrift für Medizinische Physik*, **29**, 102-127. https://doi.org/10.1016/j.zemedi.2018.11.002

[8] Nogales, A., García-Tejedor, Á.J., Monge, D., Vara, J.S. and Antón, C. (2021) A Survey of Deep Learning Models in Medical Therapeutic Areas. *Artificial Intelligence in Medicine*, **112**, Article 102020. https://doi.org/10.1016/j.artmed.2021.102020

[9] Zhang, T., Luo, Y., Li, P., Liu, P., Du, Y., Sun, P., *et al*. (2020) Cervical Precancerous Lesions Classification Using Pre-Trained Densely Connected Convolutional Networks with Colposcopy Images. *Biomedical Signal Processing and Control*, **55**, Article 101566. https://doi.org/10.1016/j.bspc.2019.101566

[10] Miyagi, Y., Takehara, K., Nagayasu, Y. and Miyake, T. (2019) Application of Deep Learning to the Classification of Uterine Cervical Squamous Epithelial Lesion from Colposcopy Images Combined with HPV Types. *Oncology Letters*, 75, Article 103589. https://doi.org/10.3892/ol.2019.11214

[11] Saini, S.K., Bansal, V., Kaur, R. and Juneja, M. (2020) Colponet for Automated Cervical Cancer Screening Using Colposcopy Images. *Machine Vision and Applications*, **31**, Article No. 15. https://doi.org/10.1007/s00138-020-01063-8

[12] Liu, L., Wang, Y., Liu, X., Han, S., Jia, L., Meng, L., *et al*. (2021) Computer-aided Diagnostic System Based on Deep Learning for Classifying Colposcopy Images. *Annals of Translational Medicine*, **9**, 1045-1045. https://doi.org/10.21037/atm-21-885

[13] Chen, J., Li, P., Xu, T., Xue, H., Wang, X., Li, Y., *et al*. (2022) Detection of Cervical Lesions in Colposcopic Images Based on the Retinanet Method. *Biomedical Signal Processing and Control*, **75**, Article 103589. https://doi.org/10.1016/j.bspc.2022.103589

[14] Zhang, X. and Zhao, S. (2018) Cervical Image Classification Based on Image Segmentation Preprocessing and a Capsnet Network Model. *International Journal of Imaging Systems and Technology*, **29**, 19-28. https://doi.org/10.1002/ima.22291

[15] Xu, T., Li, P. and Wang, X. (2021) Cervical Lesions Classification Based on Pre-Trained Mobilenet Model. 2021 *IEEE* 15*th International Conference on Anti-counterfeiting, Security, and Identification* (*ASID*), Xiamen, 29-31 October 2021, 93-96. https://doi.org/10.1109/asid52932.2021.9651726

[16] Moghtaderi, S., Yaghoobian, O., Wahid, K.A. and Lukong, K.E. (2024) Endoscopic Image Enhancement: Wavelet Transform and Guided Filter Decomposition-Based Fusion Approach. *Journal of Imaging*, **10**, Article 28. https://doi.org/10.3390/jimaging10010028

[17] Al-Stouhi, S. and Reddy, C.K. (2015) Transfer Learning for Class Imbalance Problems with Inadequate Data. *Knowledge and Information Systems*, **48**, 201-228. https://doi.org/10.1007/s10115-015-0870-3

[18] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[19] Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.R. (2012) Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. https://doi.org/10.48550/arXiv.1207.0580

[20] Srivastava, N. (2013) Improving Neural Networks with Dropout. https://api.semanticscholar.org/CorpusID:17084851

[21] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/cvpr.2018.00745

[22] Lin, T., Goyal, P., Girshick, R., He, K. and Dollar, P. (2020) Focal Loss for Dense Object Detection. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, **42**, 318-327. https://doi.org/10.1109/tpami.2018.2858826

[23] Mao, A., Mohri, M. and Zhong, Y. (2023) Cross-Entropy Loss Functions: Theoretical Analysis and Applications. https://doi.org/10.48550/arXiv.2304.07288

[24] Smith, L.N. (2017) Cyclical Learning Rates for Training Neural Networks. 2017 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Santa Rosa, 24-31 March 2017, 464-472. https://doi.org/10.1109/wacv.2017.58

[25] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[26] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[27] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. http://arxiv.org/abs/1409.1556

[28] Xie, S., Girshick, R., Dollar, P., Tu, Z. and He, K. (2017) Aggregated Residual Transformations for Deep Neural Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5987-5995. https://doi.org/10.1109/cvpr.2017.634

[29] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2261-2269. https://doi.org/10.1109/cvpr.2017.243

[30] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al*. (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929