改进宫颈癌筛查方法：基于视觉变换器的分类与可解释性

Khoa Tuan Nguyen[1,2], Ho-min Park[1,2], Gaeun Oh[1], Joris Vankerschaver[1,3], Wesley De Neve[1,2]

[1] Center for Biosystems and Biotech Data Science, Department of Environmental Technology, Food Technology, and Molecular Biotechnology, Ghent University Global Campus, Incheon, Korea
[2] IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium
[3] Department of Applied Mathematics, Computer Science and Statistics,Ghent University, Ghent, Belgium

arxiv:2504.21340v1 中译本

ABSTRACT

我们提出了一种使用 EVA-02 变换器模型进行宫颈癌筛查的宫颈细胞图像分类的新方法。我们开发了一个四步流程：微调 EVA-02，特征提取，通过多个机器学习模型选择重要特征，并训练一个带有可选损失权重的新人工神经网络以提高泛化能力。凭借此设计，我们的最佳模型实现了 0.85227 的 F1 分数，超过了基线 EVA-02 模型（0.84878）。我们还利用了 Kernel SHAP 分析，并确定了与细胞形态和染色特征相关的关键特性，为微调模型的决策过程提供了可解释性的见解。我们的代码可在 https://github.com/Khoa-NT/isbi2025_ps3c 获取。

Index Terms— 细胞分类，宫颈癌，可解释人工智能，视觉变换器

1. 介绍

宫颈癌仍然是一个重要的全球健康挑战，在女性中最常见的癌症中排名第四，每年有超过 60 万新病例和 30 万人死亡 [1]。通过巴氏涂片筛查进行早期检测已被证明在降低死亡率方面至关重要，因为它可以识别出癌前病变。然而，传统分析方法面临着重大挑战：它们资源密集、耗时且高度依赖细胞学家的专业知识。为了解决这些挑战，我们参加了 IEEE 生物医学成像国际研讨会 (ISBI)2025 挑战项目中的巴氏涂片细胞分类挑战 (PS3C) [2, 3]，以促进宫颈细胞图像自动化分类系统的开发。

在这个挑战中，我们被要求开发深度学习模型来将宫颈涂片细胞图像分类为三类：没有异常的健康细胞、可能表明病理变化的不健康细胞以及由于伪影或质量不佳而不合适的图像。该挑战提供了一个包含四个类别（包括额外的 “Both cells” 类别）的综合训练数据集，而测试数据集则集中在三个主要类别上 [2, 3]。有效性使用 F1 分数进行评估，针对每个类别计算并取所有类别的平均值以考虑数据不平衡。

2. 方法

我们采用了一个四步方法来实现高 F1 分数。首先，我们微调了基于预训练的变换器图像分类模型 EVA-02 [4]。然后，如图所示 2，我们（A）从模型中提取特征，（B）选择重要特征，并且（C）使用选定的特征带或不带损失加权来训练一个新的人工神经网络 (ANN) 模型以应对不平衡标签的问题。最后，（D）我们运用了内核夏普利添加解释 (Kernel Shapley Additive Explanations, SHAP) 来分析微调后的模型的决策过程，并提供可解释的见解，说明模型如何进行分类 [5]。

2.1. 微调 EVA-02

我们使用基于变压器架构的 EVA-02 [4] 作为基准分类模型。选择 EVA-02 是因为它在 PyTorch 图像

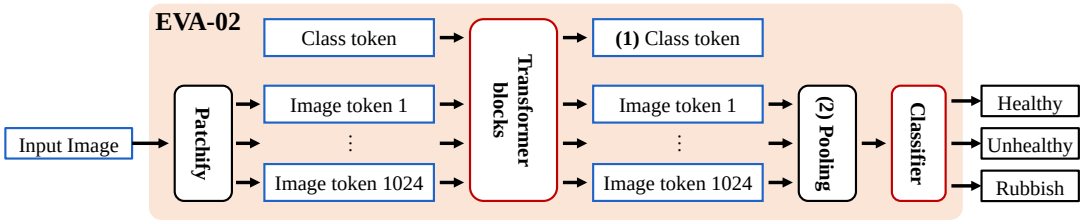


Fig. 1: EVA-02 模型的一个简单示例工作流程。输入的细胞图像被划分为小块，创建图像标记。一系列基于变压器的模块处理可训练的类别标记以及这些图像标记。输出的图像标记在传递给分类器进行健康、不健康和垃圾类别的预测之前先进行平均池化。类别标记在 (1) 处提取，而图像标记在 (2) 处提取。

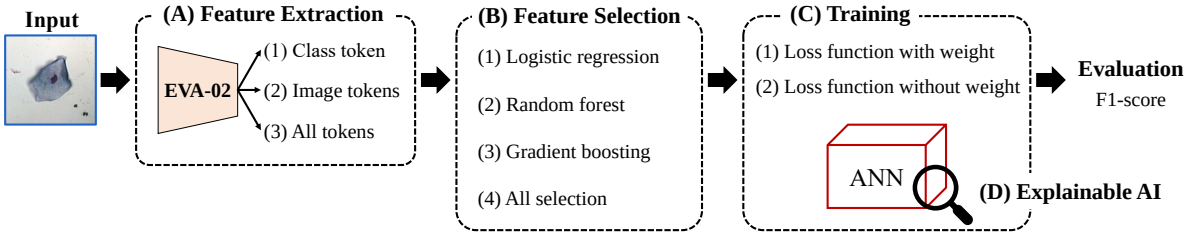


Fig. 2: 实验概述。

模型 [1] [6] 中可用的基于变压器的图像分类模型中表现出色。我们没有从头开始训练，而是微调预训练的 ImageNet 模型 [2] 在整个训练数据集上进行健康、不健康和垃圾输出类别的分类，如图 1 所示。此外，在挑战组织者 [3] 的鼓励下，我们发现使用三个类别（将“两个细胞”类别合并到“不健康”类别中）训练比使用四个类别训练产生了略微更好的结果。因此，我们将这种合并作为默认设置。与预训练的 ImageNet 模型相同的配置用于图像预处理和增强。我们采用 AdamW 优化器（weight decay = 0.05），学习率为 1e−5，使用 ReduceLROnPlateau 调度器（patience = 2，factor = 0.5）以及 PyTorch 中的 CrossEntropyLoss [7]。我们将模型微调 20 个周期，并选择损失值最低的检查点。

2.2. （A）特征提取

训练完成后，我们使用了三种选项提取图像特征：类别标记、图像标记和所有标记。类别标记如图 1 中所示的 (1)，是一个额外设计的标记，旨在捕捉数据集 [8] 中图像的全面信息。图像标记如图 1 所示的 (2)，通过将输入图像分割成小块生成，每个标记代表来自其对应小块的位置和视觉信息。“所有标记”选项包括类别标记和图像标记，结合全局和局部信息提供最全面的特征表示。

2.3. （B）特征选择

EVA-02 的特征，即类别标记、图像标记和所有标记，共享相同的格式形状 (N, L, E)，其中 N 是训练数据集中的图像数量，L 表示标记长度，E 表示嵌入大小。通过沿标记长度轴取平均值（用 $L_{\text{Class token}} = 1$、$L_{\text{Image tokens}} = 1024$ 和 $L_{\text{All tokens}} = 1025$ 表示），从每个图像中提取的特征获得了一个通用维度 $E = 768$。鉴于相对于图像大小可用的训练数据量有限，我们使用三个机器学习模型（逻辑回归、随机森林、梯度提升）及其相应的标签（垃圾、健康、不健康）来训练这些特征，以减少过拟合的风险并消除可能引入噪声的不太重要的特征 [9]。

我们从每个训练好的模型中提取了各个特征的重要性。例如，在逻辑回归中，系数的绝对值作为相应特征的重要性的度量。重要性值在三个类别之间进行了平均以生成排名，并通过为每个模型应用不同的阈值来进行特征选择。

- 逻辑回归：通过手动检查确定，应用了手动阈值

[1] 根据基准测试结果：https://github.com/huggingface/pytorch-image-models/blob/main/results/results-imagenet.csv

[2] 模型：‘timm/eva02_base_patch14_448.mim_in22k_ft_in22k_in1k’

1e−16（过滤了 31.12%）
- 随机森林：应用了一个阈值 3e−6，该阈值通过手动识别数据点数量及其数值在分布中急剧下降的点来确定（过滤了 1.95%的数据）
- 梯度提升：排除所有重要性值为 0 的数据点（过滤了 40.10%）

这一特征选择过程使我们能够在减少过拟合风险和促进更高效训练的同时保持模型的有效性。为了进行比较，我们也包括了一个“全选”选项，在这个选项中所有特征维度都被使用而没有任何选择过程。

2.4. (C) 训练一个新的 ANN 模型

为了将提取的特征整合到分类模型中，我们构建了一个具有三个隐藏层（分别为 1024、512 和 256 个神经元）的 ANN，这些值是根据其实用性能经验选择的。与仅包含一个线性层（输入至输出：$768 \to 3$）的 EVA-02 基准分类器相比，这种更深的架构允许进行更丰富的特征学习。

人工神经网络使用交叉熵损失训练了 100 个周期，我们选择了验证损失最低的检查点进行评估。

损失加权。与其在交叉熵损失中平等地对待所有类别，我们建议根据每个类别的数据点数量向损失函数添加权重，如表 2 所示。我们还对基准模型进行了带有损失加权的微调以进行比较。

加权交叉熵损失函数定义如下：

$$\mathcal{L} = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log(p_c),$$

其中：

- C 是类别的数量。
- y_c 是二进制指示符（0 或 1），表示类别标签 c 是否为当前实例的正确分类。
- p_c 是类别 c 的预测概率。
- w_c 是类别 c 的权重，计算方式为：

$$w_c = \frac{\max(\text{num\_data\_points})}{\text{num\_data\_points}_c}$$

其中：

- $\max(\text{num\_data\_points})$ 是所有类别中数据点的最大数量。
- num_data_points_c 是类 c 中的数据点数量。

2.5. (D) 基于 SHAP 的特征分析

尽管实现高效果很重要，理解模型决策同样关键，特别是在医学应用中，可解释性直接关系到患者的信任和权益 [10] 。我们采用了 Kernel SHAP [5] 来分析最佳模型的决策过程。该分析识别出输入特征中最具影响力的特性，并考察这一特性的值的变化如何与实际图像特征相关联。我们将这个重要特征表现出极低值和极高值的图像进行了比较，提供了关于这一特征所代表的视觉模式或特性的定性见解。

3. 结果与讨论

3.1. （A）特征提取的影响

我们的分析表明，特征提取方法的选择对模型的有效性有显著影响。如表 1 所示，在三种提取方法（类标记、图像标记和所有标记）中，总体上“所有标记”方法获得了最高的 F1 分数，使用“全选”在未加权损失条件下达到了最高有效性 0.85227。这表明全局信息（类标记）与局部信息（图像标记）的结合提供了最全面的分类表示。

3.2. （B）特征选择的有效性

我们的特征选择方法，如表 1 所示的结果，基于模型特定的阈值（梯度提升为 40.10%，随机森林为 1.95%，逻辑回归为 31.12%）过滤掉不太重要的特征，证明了在减少计算复杂性的同时保持模型有效性的有效性。具有筛选特征的模型的竞争效果表明，我们的选择标准成功地识别出分类中最相关的特征。

3.3. (C) 损失加权的影响

损失加权对模型效果的影响在表 1 中显示出显著的模式。尽管加权和未加权损失场景之间的差异相对较小（差异通常小于 0.002），但在不同的提取（A）和选择（B）方法组合中存在一致的模式。例如，使用类标记提取时，加权损失通常导致在不同机器学习模型上的 F1 分数更稳定（范围：0.85007-0.85039），相比之下未加权损失（范围：0.84921-0.85112）。

Table 1: F1 分数↑(越高越好) 在不同特征提取方法和机器学习模型组合下的加权和非加权损失条件下的表现。显著的分数显示在**粗体**。

(A) Extraction method	(B) Selection	(C-1) Weighted loss	(C-2) Unweighted loss
分类标记	Gradient boosting	0.85007	0.85033
	Random forest	0.85020	0.85112
	Logistic regression	0.85031	0.84921
	All selection	0.85039	0.85015
图像标记	Gradient boosting	0.85069	0.85037
	Random forest	0.85070	0.85166
	Logistic regression	0.85078	0.85141
	All selection	0.85154	0.85108
所有令牌	Gradient boosting	0.85056	0.85044
	Random forest	0.85225	0.85008
	Logistic regression	0.85004	0.85072
	All selection	0.85169	0.85227
EVA-02 基线		0.84170	0.84878

Class	# of datapoints	Weights
Rubbish	50,371	1.000
Healthy	28,895	1.743
Unhealthy + Both cells	5,814	8.664

Table 2: 每个类别的权重。

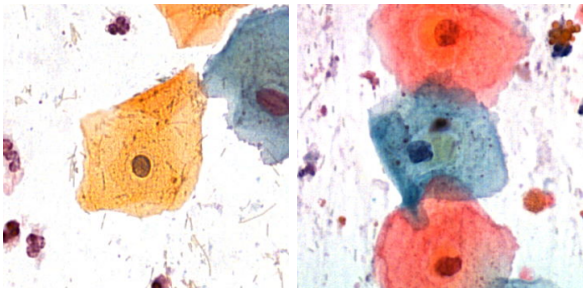
3.4. 与基线 EVA-02 的比较

表 1 中的结果表明，我们提出的特征提取和选择管道始终优于基线 EVA-02 模型 (F1 分数: 0.84878)。我们的最佳配置使用所有标记、全部选择和未加权损失，实现了 0.85227 的 F1 分数，比基线提高了 0.35%。

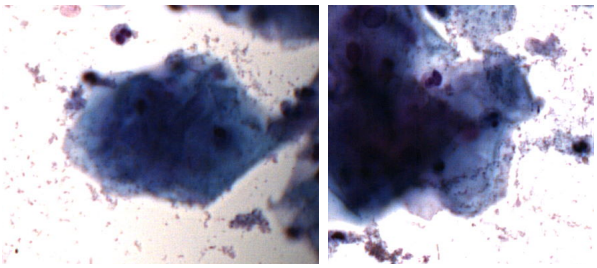
3.5. (D) 基于 SHAP 的特征分析

我们使用最佳模型（所有标记 + 全部选择 + 非加权损失）分析了 768 维输入特征中每个特征的相对重要性。该分析表明，特征#10 在模型决策过程中的影响力最大。

图 3 提供了特征#10 的可视化见解。图 3(a) 显示具有高特征#10 值的图像，而图 3(b) 则显示低值的图像。视觉特征表现出显著差异：具有高特征#10 值的图像是由富含细胞质的大细胞和强烈的红/蓝染色以及清晰的核边界构成的，而低特征#10 值的图像则显示出暗淡区域和不明显的细胞。



(a) 具有高特征#10 值的样本图像，其特点是大细胞、丰富的胞质和强烈的红/蓝染色以及清晰的核边界。



(b) 特征值#10 较低的样本图像，显示了一个较大的暗不确定性区域。

Fig. 3: 基于特征值#10 的细胞图像比较，展示了不同的形态和染色特性。

4. 结论

在这项研究中，我们提出了一种创新的方法来对宫颈细胞图像进行分类，以帮助早期发现宫颈癌。我们的流程结合了基于 EVA-02 变换器模型的特征提取、特征选择和更深的 ANN 分类器，在 F1 分数上比 EVA-

02 基线更高。一个值得注意的发现是，All tokens 方法利用了全局信息（Class token）和局部信息（Image tokens），表现出了最高的有效性。此外，我们确认通过特征选择过程可以在降低模型复杂性的同时保持其有效性。进一步地，通过 Kernel SHAP，我们确定特征#10 对模型决策的影响最大。特别地，通过对图像的分析，我们能够确认细胞大小、胞浆特性以及染色强度在分类中具有显著影响。

我们的研究发现的主要贡献在于提出了一种方法，该方法在医学环境中满足了自动宫颈癌筛查所需的高度准确性和可解释性。然而，这项研究的局限性在于最小有效性和最大有效性之间的差异为 1.06%（最小值：0.8417，最大值：0.85227），并且最终结果尚未得到确认，因为评估仅在公共排行榜上进行。

5. REFERENCES

[1] World Health Organization, "Cervical Cancer: Key Facts," 2024, Accessed: 2025-02-03.

[2] David Kupas, Andras Hajdu, Ilona Kovacs, Zoltan Hargitai, Zita Szombathy, and Balazs Harangi, "Annotated Pap cell images and smear slices for cell classification," Scientific Data, vol. 11, no. 1, pp. 743, 2024.

[3] Balázs Harangi, András Hajdu, Dávid Kupás, Ilona Kovács, Péter Kovács, and Nicolai Spicher, "Pap Smear Cell Classification Challenge (PS3C)," https://kaggle.com/competitions/pap-smear-cell-classification-challenge, 2024.

[4] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao, "Eva-02: A visual representation for neon genesis," Image and Vision Computing, p. 105171, 2024.

[5] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[6] Ross Wightman, "PyTorch Image Models," https://github.com/huggingface/pytorch-image-models, 2019.

[7] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala, "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). Apr. 2024, ACM.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.

[9] Jundong Li, Ke Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu, "Feature selection: A data perspective," ACM Computing Surveys (CSUR), vol. 50, no. 6, pp. 94:1–94:45, 2017.

[10] Bryce Goodman and Seth Flaxman, "Euro-

pean Union regulations on algorithmic decision-making and a right to explanation," AI Magazine, vol. 38, no. 3, pp. 50–57, 2017.