

Explorando o uso de Vision Transformer na Classificação de Lesões Anais e Cervicais: uma Revisão Sistemática

Ingrid Bromerschenckel¹, Andrea Gomes Campos¹

¹ Departamento de Computação – Universidade Federal de Ouro Preto (UFOP)
Caixa Postal 35.400-000 – Ouro Preto – MG – Brazil

ingrid.bromerschenckel@aluno.ufop.edu.br, andrea@ufop.edu.br

Abstract. *Although anal canal cancer (ACC) has a low incidence among gastrointestinal tract tumors, a significant increase in the number of cases has been observed, making it an emerging public health concern. This paper presents a systematic review of the use of the Vision Transformer architecture in classifying cervical and anal cells, evaluating its performance compared with other machine learning architectures. Cervical cells were included in the study due to their histological similarity to anal cells and the widespread use of the Pap test. The objective is to identify the state of the art in the diagnosis of cervical cancer through conventional cytology and to explore the feasibility of using transfer learning for application in anal cancer.*

Resumo. *Embora o câncer de canal anal (CCA) apresente baixa incidência entre os tumores do trato digestivo, observa-se um aumento significativo no número de casos, configurando-se como uma questão emergente de saúde pública. Este artigo apresenta uma revisão sistemática sobre o uso da arquitetura Vision Transformer na classificação de células cervicais e anais, avaliando seu desempenho em comparação com outras arquiteturas de aprendizado de máquina. As células cervicais foram incluídas no estudo devido à semelhança histológica com as células anais e à ampla utilização do teste de Papanicolau. O objetivo é identificar o estado da arte no diagnóstico de câncer cervical por meio de citologia convencional e explorar a viabilidade do uso de transferência de aprendizado para aplicação no câncer anal.*

1. Introdução

De acordo com a Organização Mundial da Saúde (OMS), cerca de 14,1 milhões de novos casos de câncer são diagnosticados anualmente em todo o mundo. O câncer é a segunda principal causa de morte global, sendo responsável por aproximadamente 9,6 milhões de óbitos por ano — e a tendência é que, nos próximos 25 anos, se torne a principal. O CCA representa cerca de 4% das neoplasias malignas do trato digestivo inferior.

Apesar de sua baixa ocorrência entre os tumores do trato digestivo, o CCA demonstra um crescimento progressivo na incidência de casos, especialmente entre populações vulneráveis, como pessoas vivendo com o Virus da Imunodeficiência Humana (HIV), imunossuprimidos, receptores de transplantes, indivíduos com histórico de neoplasias ou displasias cervicais e/ou vulvares de alto grau, além daqueles que praticam sexo anal receptivo com penetração por pênis — prática associada ao maior risco

de microlesões na mucosa anal, facilitando a infecção por agentes como o Papilomavírus Humano (HPV) e o HIV. Nos Estados Unidos, a incidência de cânceres escamosos do ânus aumentou significativamente entre as décadas de 1970 e 2000, com destaque para homens, especialmente homens negros ([Stewart et al. 2018])).

A descoberta de que o HPV, particularmente os sorotipos 16 e 18, é a principal causa desse tipo de câncer, consolidou a compreensão do ACC como uma possível consequência de uma infecção sexualmente transmissível. Esse entendimento reforça seu potencial de prevenção por meio de estratégias de rastreamento baseadas em evidências, vacinação e diagnóstico precoce. Como há uma relação inversa entre o estágio da doença e a sobrevida, a detecção precoce surge como uma ferramenta essencial para reduzir a mortalidade e posiciona o câncer anal como uma preocupação relevante e tratável de saúde pública.

A maioria dos pacientes apresenta massa de crescimento lento envolvendo o canal anal ou a pele perianal. Dor e sangramento são comuns, ocorrendo em aproximadamente metade dos pacientes, embora mais de 20% dos pacientes possam ser assintomáticos ([Robb and Mutch 2006])). O diagnóstico de ACC pode ser frequentemente retardado, principalmente devido a sintomas anorretais inespecíficos, que são frequentemente atribuídos erroneamente à patologias anorretais benignas, como hemorroidas.

O diagnóstico citológico clínico correto do CCA como primeira instância diagnóstica pode permitir uma abordagem terapêutica mais rápida e direcionada e retardar o avanço da doença. O objetivo dessa revisão sistemática é estudar o modelo de Vision Transformer (ViT) na classificação de células cervicais e investigar sua aplicação para células anais via transferência de aprendizado, em vista que a base de dados de células anais ainda está em desenvolvimento. Quando comparado à CNN, o ViT aprende representações mais uniformes, é capaz de absorver mais informações globais e estabelece conexões entre diferentes regiões das imagens, o que é crucial para classificação precisa do câncer, segmentação e outras tarefas.

2. Metodologia

O modelo PICOC foi utilizado para estruturar as perguntas de pesquisa, considerando os seguintes elementos: População (pacientes com lesões e câncer do colo do útero e anal), Intervenção (aplicação ou extensão do ViT em modelos de classificação de imagens de citologia convencional), Comparação (modelos tradicionais ou de transferência de aprendizado entre tipos de câncer), Resultado (desempenho em termos de precisão diagnóstica e generalização), e Contexto (aplicações de Deep Learning para apoio a diagnósticos citopatológicos). Com base nessa estrutura, foram formuladas duas perguntas de pesquisa principais: 1) Como o ViT tem sido empregado na classificação de lesões e câncer de colo do útero e anal em lâminas de citologia convencional? e (2) É possível estender o modelo de classificação de lesões cervicais para lesões anais?

A busca foi realizada em quatro acervos: (1) IEEE, (2) PubMed, (3) Scopus, (4) Web of Science (Figura 1). O critério de aceite incluiu artigos que estivessem dentro do escopo da proposta, ou seja, que abordassem sobre as células-alvo (cervicais e anais) e empregassem a citologia convencional como método de triagem. Os critérios de exclusão foram: (1) Documentos duplicados; (2) Literatura cinza; (3) Artigos cuja menção a ViT em células-alvo foi superficial; (4) Artigos com ViT executado sobre imagens de outros

órgãos (que não cólon, reto ou útero); (5) Artigos com ViT executado sobre imagens obtidas por outros métodos (que não citologia convencional); (6) Artigos com ViT empregado para outra finalidade que não a classificação de células.

Ao todo, foram importados 277 artigos, sendo eles: IEEE – 62, Pubmed – 36, Scopus – 108, Web of Science – 71. Dos artigos identificados, 16 foram selecionados para leitura com base nos critérios de inclusão e exclusão estabelecidos. Os estudos selecionados foram publicados entre os anos de 2022 e 2024.

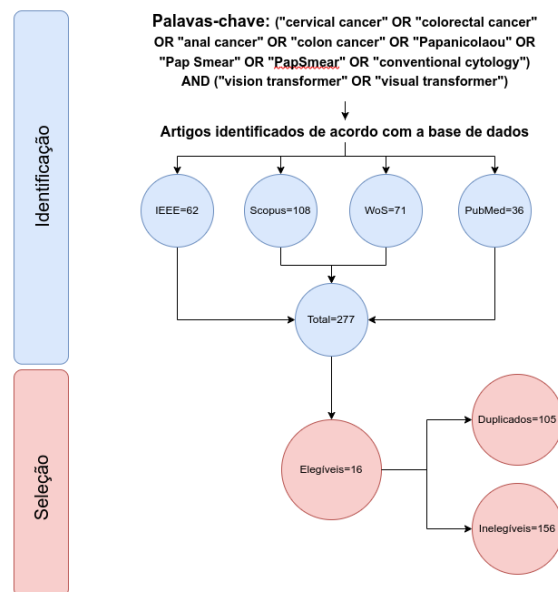


Figura 1. Descrição da Revisão Bibliográfica

3. Resultados

Estudos recentes exploram diferentes variações do Vision Transformer (ViT) e sua integração com redes convolucionais (CNNs) para aprimorar a classificação de imagens cervicais e a detecção de câncer. As abordagens revisadas exploram a combinação das características locais capturadas pelas CNNs com a capacidade dos ViTs para modelar dependências globais, visando aprimorar a precisão e a eficiência computacional.

As bases de dados utilizadas nos artigos analisados foram: SIPaKMeD [Plissiti et al. 2020], Herlev [Jantzen et al. 2005] e CRIC [Group 2021], sendo as duas primeiras de imagens com células unicelulares. As métricas de desempenho dos artigos com destaque foram divididas em duas tabelas, de acordo com as bases de dados empregadas (Tabelas 1 e 2).

Alguns modelos se destacam em termos de acurácia e métricas complementares (precisão, revocação e F1-score), indicando potencial para aplicação em diagnóstico de citologia cervical. Em relação a base de dados SipakMed, cabe destacar alguns artigos:

MaxCerVixT – Alcança acurácia de 99,02%, precisão de 99,03%, revocação de 99,04% e F1-score de 99,02% [Pacal 2024]. Esse modelo adapta a arquitetura do ViT para melhorar a capacidade de generalização em conjuntos de dados reduzidos, equilibrando

Tabela 1. Resultados das Bases de Dados SIPaKMeD

Referência	Método	Base de Dados	Métricas de Desempenho
[Khowaja et al. 2023]	Cervix Visionator ELM	SIPaKMeD	Acur.: 98,89%, Prec.: 98,42%, Rec.: 97,87%, F1: 98,76%
[AlMohimeed et al. 2024]	ViT-PSO-SVM	SIPaKMeD	Acur.: 97,25%, Prec.: 97,25%, Rec.: 97,25%, F1: 97,23%
[Fang et al. 2024]	DIFF (CNN + ViT)	SIPaKMeD	Acur.: 96,02%, Prec.: 96,09%, Rec.: 96,04%, F1: 96,04%
[Liu et al. 2022]	CVM-Cervix	CRIC + SIPaKMeD	Acur.: 91,72%, Prec.: 91,80%, Rec.: 91,60%, F1: 91,70%
[Khowaja et al. 2024]	IATS	SIPaKMeD	Acur.: 95,88%, Prec.: 95,37%, Rec.: 96,25%, F1: 95,79%
[Hemalatha et al. 2023]	CervixFuzzyFusion (segm.)	SIPaKMeD	Acur.: 98,13%, Prec.: 97,4%, Rec.: 97,6%, F1: 97,6%
[Hemalatha et al. 2023]	CervixFuzzyFusion (não segm.)	SIPaKMeD	Acur.: 96,13%, Prec.: 90,2%, Rec.: 90,4%, F1: 90,2%
[Maurya et al. 2023]	VisionCervix CNN-LSTM	SIPaKMeD	Acur.: 95,80%, Prec.: 100%, Rec.: 95,80%, F1: 97,5%
[Maurya et al. 2023]	VisionCervix ViT-CNN	SIPaKMeD	Acur.: 97,65%, Prec.: 99,54%, Rec.: 97,65%, F1: 98,58%
[Pacal 2024]	MaxCerViXT	SIPaKMeD	Acur.: 99,02%, Prec.: 99,03%, Rec.: 99,04%, F1: 99,02%
[Emara et al. 2024]	HViT-Cerv	SIPaKMeD	Acur.: 98%, Prec.: 98%, Rec.: 98%, F1: 98%
[Sholik et al. 2024]	CNN+ViT PCA e LDA	SIPaKMeD	Acur.: 98,52%, Prec.: 98,81%, Rec.: 97,80%, F1: 98,28%

profundidade e eficiência computacional. Sua arquitetura otimizada e alta precisão sugerem que é uma alternativa promissora para a detecção de câncer cervical em pequenos conjuntos de dados.

Cervix Visionator ELM – Com acurácia de 98,89% e precisão de 99,42%, o modelo atinge F1- score de 98,76% [Khowaja et al. 2023]. A combinação de EfficientNet e ViT com uma Máquina de Aprendizado Extremo (ELM) explora as vantagens complementares dessas arquiteturas, resultando em um desempenho robusto e eficiente para classificação de células cervicais.

Em relação as bases de dados Herlev e CRIC, cabe destaque aos artigos:

3D CNN-ViT com classificador KELM - Esse modelo combina a capacidade de extração de características espaciais e temporais da 3D CNN com a habilidade do ViT de capturar representações visuais complexas. Sua arquitetura inclui camadas de convolução 3D para extrair características das imagens cervicais e reduzir informações redundantes, seguida por módulos do ViT para extrair representações em diferentes níveis de abstração. Essas características são então concatenadas por uma Rede de Pirâmide de Características 3D (3D FPN), que é uma CNN, aprimoradas por um bloco de exaltação 3D (3D SE), também uma CNN e, finalmente, classificadas por uma Máquina de Aprendizado Extremo com Kernel (KELM). Os resultados demonstraram uma alta precisão de 98,6% [Abinaya and Sivakumar 2024].

Tabela 2. Resultados das Bases de Dados Herlev e CRIC

Referência	Método	Base de Dados	Métricas de Desempenho
[Abinaya and Sivakumar 2024]	3D CNN-ViT + KELM	Herlev	Acc. 98,6%, Prec. 97,5%, Rec. 98,2%, F1 98,4%
[AlMohimeed et al. 2024]	ViT-PSO-SVM	Herlev (Binário)	Acc. 97,78%, Prec. 97,94%, Rec. 97,77%, F1 97,80%
[Fang et al. 2024]	DIFD (CNN + ViT)	Herlev	Acc. 94,55%, Prec. 94,13%, Rec. 91,66%, F1 92,78%
[Li et al. 2024]	VTCNet	Herlev	Acc. 95,95%, Prec. 96,18%, Rec. 95,95%, F1 95,17%
[Sholik et al. 2024]	CNN+ViT (PCA/LDA)	Herlev	Acc. 97,83%, Prec. 98,75%, Rec. 92,86%, F1 95,52%
[Liu et al. 2022]	CVM-Cervix	CRIC + SIPaKMeD	Acc. 91,72%, Prec. 91,80%, Rec. 91,60%, F1 91,70%

ViT-PSO-SVM – O modelo alcançou acurácia de 97,78% e é beneficiado pela otimização com o algoritmo *Particle Swarm Optimization* (PSO) e a utilização de *Support Vector Machine* (SVM) para classificação, resultando em uma integração eficiente para cenários de detecção de câncer cervical [AlMohimeed et al. 2024].

Os modelos demonstram alto potencial para o diagnóstico de citologia cervical devido à sua elevada precisão e estratégias inovadoras de fusão e otimização de características. O MaxCerVixT, o Cervix Visionator ELM em SIPaKMeD, e o 3D CNN-ViT com classificador KELM na base de dados Herlev, apresentaram desempenho ligeiramente superior em métricas gerais, destacando-se como potenciais candidatos para diagnósticos de alta confiabilidade.

4. Conclusões

A integração entre CNN e ViT tem demonstrado, de maneira geral, um melhor desempenho na classificação de imagens médicas, especialmente em contextos com dados de baixa resolução ou estruturas complexas. Dos 16 artigos analisados, 9 utilizaram a combinação entre CNN e ViT. Nessa abordagem, a CNN é responsável por extrair características locais detalhadas, como bordas e texturas, enquanto o ViT capta relações globais, contribuindo para uma detecção mais precisa, conforme demonstrado em diversos estudos. Seis artigos utilizaram ViT isolado, ou com outras abordagens que não envolvem CNN. Embora o ViT isolado seja eficaz na captura de relações globais, ele tende a apresentar desempenho inferior em tarefas que exigem alta resolução de detalhes locais, um aspecto crítico na citologia cervical. Além disso, técnicas de transferência de aprendizado em células sanguíneas e o uso de modelos pré-treinados foram aplicados para melhorar a precisão e reduzir o tempo de treinamento.

Adicionalmente, não foram identificados estudos que explorassem a transferência de aprendizado para adaptação do modelo a outros tipos celulares. A única abordagem observada envolveu a aplicação de um modelo treinado em células cervicais para a classificação de células sanguíneas [Fang et al. 2024] [Liu et al. 2022].

Referências

Abinaya, K. and Sivakumar, B. (2024). A deep learning-based approach for cervical cancer classification using 3d cnn and vision transformer. *Journal of Imaging Informatics*

in Medicine, 37:280–296.

- AlMohimeed, A., Shehata, M., and et. al. (2024). Vit-pso-svm: Cervical cancer prediction based on integrating vision transformer with particle swarm optimization and support vector machine. *Bioengineering*, 11(7):729.
- Emara, H. M., El-Shafai, W., and et. al. (2024). Cervical cancer detection: A comprehensive evaluation of cnn models, vision transformer approaches, and fusion strategies. *IEEE Access*.
- Fang, M., Fu, M., Liao, B., and et. al. (2024). Deep integrated fusion of local and global features for cervical cell classification. *Computers in Biology and Medicine*, 171:108153.
- Group, C. R. (2021). Cric searchable image database. *Federal University of Minas Gerais*. Available at: <http://www.cric.com.br/>.
- Hemalatha, K., Vetrivel, V., Dhandapani, M., and Gladys, A. A. (2023). Cervixfuzzy-fusion for cervical cancer cell image classification. *Biomedical Signal Processing and Control*, 85:104920.
- Jantzen, J., Dounias, G., and Norup, J. (2005). The herlev pap smear dataset. *University of Denmark*. Available at: <http://mde-lab.aegean.gr/index.php/downloads>.
- Khowaja, A., Zou, B., and Kui, X. (2024). Enhancing cervical cancer diagnosis: Integrated attention-transformer system with weakly supervised learning. *Image and Vision Computing*, 149:105193.
- Khowaja, A., Zou, B., and Xiaoyan, K. (2023). Cervix visionator elm: A novel approach to early detection of cervical cancer.
- Li, M., Que, N., Zhang, J., Du, P., and Dai, Y. (2024). Vtcnet: A feature fusion dl model based on cnn and vit for the classification of cervical cells. *International Journal of Imaging Systems and Technology*, 34(5).
- Liu, W., Li, C., Xu, N., and et. al. (2022). Cvm-cervix: A hybrid cervical pap-smear image classification framework using cnn, visual transformer, and multilayer perceptron. *Pattern Recognition*, 130:108829.
- Maurya, R., Pandey, N. N., and Dutta, M. K. (2023). Visioncervix: Papanicolaou cervical smears classification using novel cnn-vision ensemble approach. *Biomedical Signal Processing and Control*, 79(Part 2):104156.
- Pacal, I. (2024). Maxcervix: A novel lightweight vision transformer-based approach for precise cervical cancer detection. *Knowledge-Based Systems*, 289:111482.
- Plissiti, M. E., Nikou, C., and Charchanti, A. (2020). Sipakmed: A dataset for classification and detection of cervical cancer cells. *IEEE Journal of Biomedical and Health Informatics*, 24(4):1175–1185.
- Robb, B. W. and Mutch, M. G. (2006). Epidermoid carcinoma of the anal canal. *Clinics in Colon and Rectal Surgery*, 19(2):108–115.
- Sholik, M., Fatichah, C., and Amaliah, B. (2024). Deep feature extraction of pap smear images based on convolutional neural network and vision transformer for cervical can-

cer classification. In *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 290–296, Bali, Indonesia.

Stewart, D. B., Gaertner, W. B., Glasgow, S. C., Herzig, D. O., Feingold, D., and Steele, S. R. (2018). The american society of colon and rectal surgeons clinical practice guidelines for anal squamous cell cancers (revised 2018). *Diseases of the Colon & Rectum*, 61(7):755–774. Prepared on behalf of the Clinical Practice Guidelines Committee of the American Society of Colon and Rectal Surgeons.