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

Ingrid Bromerschenckel[1]

[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

Abstract. *Although anal canal cancer (ACC) has a low incidence among digestive tract tumors, a significant increase in cases has been observed, establishing it as an emerging public health concern. This article presents an overview of the problem and potential solution approaches, along with a systematic review on the use of the Vision Transformer architecture for the classification of cervical and anal cells, comparing its performance with other machine learning models. Cervical cells were included in the study due to their histological similarity to anal cells and the widespread use of the Papanicolaou test. The objective is to identify the state of the art in cervical cancer diagnosis through conventional cytology and to explore the feasibility of applying transfer learning techniques to anal cancer diagnosis. Based on the findings, experimental proposals will be made using the "Cric Cervix" dataset, one of the most established and comprehensive resources in the literature for the study and classification of cervical lesions. The code will be available on GitHub, with access at: (*`https://github.com/ingridbromer/reconhecimento-padroes`*).*

Resumo. *Embora o câncer do canal anal (Anal Canal Cancer, ACC) 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 trabalho propõe uma introdução ao tema e discute possíveis abordagens para a triagem do ACC, com base em uma revisão sistemática sobre o uso da arquitetura Vision Transformer (ViT) na classificação de células cervicais e anais, comparando seu desempenho com o de outras arquiteturas de aprendizado de máquina. As células cervicais foram incluídas na análise devido à sua semelhança histológica com as células anais e à ampla aplicação do exame de Papanicolaou na prática clínica. O objetivo da revisão é identificar como o ViT vem sendo empregado no diagnóstico do câncer cervical por meio da citologia convencional, além de investigar a viabilidade do uso de transferência de aprendizado para aplicação no diagnóstico do câncer anal. Com base nos achados, serão propostos modelos utilizando a base de dados "Cric Cervix", uma das mais consolidadas na literatura para o estudo e a classificação de lesões cervicais, com o intuito de estender esses modelos para a identificação de lesões celulares associadas ao ACC. O código estará disponível no GitHub, com acesso em: (*`https://github.com/ingridbromer/reconhecimento-padroes`*).*

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 câncer de canal anal (*Anal Canal Cancer*, ACC) representa cerca de 4% das neoplasias malignas do trato digestivo inferior.

Embora o câncer anal apresente baixa incidência entre os tumores do trato digestivo, observa-se um aumento progressivo de casos, especialmente entre populações vulneráveis, como pessoas vivendo com o Vırus 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 ([Klas et al. 1999])). Dor e sangramento são comuns, ocorrendo em aproximadamente metade dos pacientes, embora mais de 20% dos pacientes possam ser assintomáticos ([Ryan et al. 2000, 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 ([Klas et al. 1999, Tanum et al. 1991])).

Não há um consenso internacional sobre a triagem de lesões anais. As recomendações variam conforme a organização e a população-alvo (Figura 1), e os testes utilizados podem incluir o teste de HPV, a citologia anal ou ambos de forma combinada. O teste de HPV possui maior sensibilidade para detectar infecção por tipos oncogênicos do vírus, enquanto a citologia anal apresenta maior especificidade para identificar alterações celulares sugestivas de lesões pré-malignas ou malignas ([Dyer et al. 2025].). A anuscopia ou proctoscopia com biópsia é essencial para estabelecer o tamanho da lesão, determinar sua localização dentro do canal anal e confirmar o diagnóstico após a triagem via citologia.

Grupos-alvo para triagem	Institutos Nacionais de Saúde dos EUA Escritório de Pesquisa em AIDS	Sociedade Brasileira de Oncologia Cirúrgica	Sociedade Portuguesa de Coloproctologia	Sociedade Nacional Francesa de Coloproctologia	Sociedade Internacional de Neoplasia Anal
HSH vivendo com HIV	X	X	X	X	X
HSM vivendo com HIV	X				X
Homens trans vivendo com HIV	X				X
Mulheres trans vivendo com HIV	X				X
Mulheres cis vivendo com HIV	X				X
Mulheres com transplante de órgão sólido > 10 anos			X	X	X
Homens com transplante de órgão sólido > 10 anos					X
Mulheres com histórico de câncer vulvar			X	X	X

HSH: Homens que fazem sexo com homens
HSM: Homens que fazem sexo com mulheres

Figura 1. Grupos-alvo para rastreamento de câncer colorretal.

Fonte: Adaptado de [Albuquerque and Fontes 2025].

O emprego do rastreamento de rotina de lesões intraepiteliais anais nas populações de risco baseia-se no sucesso obtido na redução do câncer cervical, por meio da citologia convencional em exames de Papanicolaou, em vista que o colo uterino e o canal anal possuem histologias semelhantes. A interpretação dos resultados segue os critérios atuais para a citologia cervicovaginal empregando o Sistema Bethesda, reconhecido internacionalmente para avaliação de lesões celulares ([Hopp et al. 2023].).

O Sistema Bethesda foi criado em 1988 para proporcionar uniformidade nos laudos citopatológicos, sobretudo os de citologia cervical, permitindo uma comunicação mais clara entre citopatologistas e clínicos. Esse sistema foi revisando em 1991, 2001 e 2014, refletindo novas evidências e práticas médicas ([Nayar and Wilbur 2015])). Nele, as células são classificada em: (1) Negativo para Lesão Intraepitelial ou Malignidade (NILM); (2) Células Escamosas Atípicas de Significado Indeterminado, possivelmente não neoplásicas (ASC-US); (3) Células Escamosas Atípicas, não podem excluir uma lesão de alto grau (ASC-H); (4) Lesão Intraepitelial Escamosa de Baixo Grau (LSIL); (5) Lesão Intraepitelial Escamosa de Alto Grau (HSIL); (6) Carcinoma Espinocelular (CEC).

O câncer cervical é o quarto câncer mais comum entre mulheres no mundo, segundo dados da OMS ([Bray et al. 2018])). A periodicidade do exame de Papanicolaou é um fator crucial na prevenção do câncer do colo do útero. A realização regular desse exame permite a detecção precoce de lesões precursoras, possibilitando intervenções eficazes antes que evoluam para formas invasivas da doença ([Bernal et al. 2023].).

Há duas maneiras de realizar o exame de Papanicolaou, que diferem principalmente na forma de preparo e preservação da amostra coletada. No primeiro método, conhecido como citologia convencional, a coleta inicia-se com a inserção de um instrumento chamado espéculo na vagina, permitindo a visualização do colo do útero. Utilizando uma

espátula de madeira para a ectocérvice e uma escova de nylon para a endocérvice, o profissional coleta as amostras celulares e as espalha diretamente em uma lâmina de vidro. Essa lâmina é fixada e enviada para análise em laboratórios especializados em citopatologia. Entretanto, esse método pode sofrer interferência de muco, sangue ou sobreposição de células, o que pode dificultar a leitura e reduzir a sensibilidade do exame.

O segundo método, conhecido como citologia em meio líquido, começa de forma semelhante, utilizando geralmente apenas a escova endocervical. A diferença está no preparo da amostra: as células coletadas são depositadas em um frasco com líquido conservante, em vez de serem aplicadas diretamente na lâmina. Esse meio permite que, no laboratório, o material passe por um processo de separação que remove impurezas como muco, bactérias e hemácias, resultando em uma lâmina mais limpa, com células distribuídas de forma uniforme e com menor sobreposição. Esse método também permite a realização de exames adicionais, como a detecção do HPV, a partir da mesma amostra. Apesar das vantagens em termos de qualidade e sensibilidade diagnóstica, a citologia em meio líquido apresenta um custo mais elevado, o que ainda limita sua ampla adoção no sistema público de saúde brasileiro ([Souza et al. 2021, Chandwani et al. 2021, Manickadevi et al. 2022])).

Durante a análise da lâmina com as amostras celulares coletadas, o patologista enfrenta o desafio de examinar uma grande quantidade de imagens — em torno de 15.000, em média —, o que implica em um volume de dados considerável. Esse processo de análise manual exige não apenas uma atenção extrema, mas também uma grande carga cognitiva e física, o que pode levar à fadiga tanto física quanto mental dos profissionais envolvidos.

Além disso, o exame de Papanicolaou é altamente dependente da experiência do citopatologista. Essa característica torna o processo interpretativo e aumenta a possibilidade de variações nos resultados, uma vez que a precisão da análise pode ser influenciada pela capacitação e pelo nível de experiência do profissional ([Claro et al. 2021].). Essa dependência da interpretação humana também contribui para o elevado custo associado ao exame, visto que é necessário contar com profissionais altamente qualificados para realizar a leitura das lâminas, o que encarece significativamente a mão de obra ([Campos and Carvalho 2024].).

A análise manual das lâminas, portanto, está sujeita a erros, incluindo falsos positivos e falsos negativos, que podem afetar tanto o diagnóstico quanto o tratamento dos pacientes. Falsos positivos ocorrem quando lesões são diagnosticadas erroneamente, resultando em tratamentos desnecessários e aumento de ansiedade para o paciente, enquanto falsos negativos podem levar à falta de diagnóstico e ao atraso no tratamento de lesões que poderiam ser tratadas precocemente. Esses erros podem, por sua vez, impactar a saúde psicológica do paciente e prejudicar o desempenho físico e psicológico dos profissionais de saúde envolvidos no processo.

Diante dessas dificuldades, surge a necessidade de buscar metodologias que possam auxiliar na análise das lâminas, com o objetivo de melhorar a precisão do diagnóstico e reduzir a carga de trabalho dos citopatologistas. Tecnologias emergentes, como sistemas de inteligência artificial e automação, têm mostrado grande potencial para melhorar a qualidade dos resultados, reduzindo erros humanos e proporcionando diagnósticos mais

rápidos e precisos ([Hou et al. 2022].). Dessa forma, essas metodologias contribuem para a otimização de exames de citopatológicos, oferecendo não apenas uma melhoria no. diagnóstico, mas também uma redução no custo e no tempo necessário para realizar as análises.

1.1. Redes Neurais Convolucionais

Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*) são uma classe especializada de redes neurais artificiais voltada para o processamento de dados com estrutura em grade, como imagens. Inspiradas na organização do córtex visual de mamíferos, as CNNs têm se destacado em tarefas de visão computacional, como classificação, segmentação e detecção de objetos em imagens ([LeCun et al. 2015])).

As CNNs fazem parte do campo do Aprendizado Profundo, que é um subconjunto do Aprendizado de Máquina baseado em redes neurais artificiais com múltiplas camadas. No caso das CNNs, essas camadas incluem operações como convolução, *pooling* ou subamostragem, e camadas totalmente conectadas, que permitem extrair representações hierárquicas dos dados de entrada — desde bordas simples até padrões morfológicos complexos ([Krizhevsky et al. 2012])). Essa estrutura torna as CNNs particularmente eficazes em reconhecer padrões visuais em imagens médicas, incluindo lâminas citológicas.

A Figura 2 apresenta um exemplo de representação esquemática de uma CNN básica. Essa rede é composta por cinco camadas distintas: entrada, convolução, subamostragem, uma camada totalmente conectada e a última camada de saída. Essas etapas estão organizadas em duas partes principais: extração de características e classificação. A natureza sequencial dessas camadas possibilita que o modelo aprenda representações de características cada vez mais complexas e detecte padrões não lineares de maneira sistemática ([Liu et al. 2019])).

A etapa de extração de características inclui as camadas de entrada, convolução e subamostragem. Já a etapa de classificação compreende a camada totalmente conectada e a de saída. A camada de entrada define um tamanho fixo para as imagens de entrada, que são redimensionadas quando necessário. Em seguida, a imagem passa pela camada de convolução, onde é processada por múltiplos filtros aprendidos, aplicados com pesos compartilhados. Depois, a camada de subamostragem reduz o tamanho da imagem, tentando preservar ao máximo as informações relevantes. As saídas dessa etapa são conhecidas como mapas de características (em inglês, *feature maps*).

Na fase de classificação, as características extraídas são combinadas nas camadas totalmente conectadas. Por fim, a camada de saída possui um neurônio para cada categoria de objeto, e o resultado da classificação é obtido a partir da ativação desses neurônios ([Phung and Rhee 2019])).

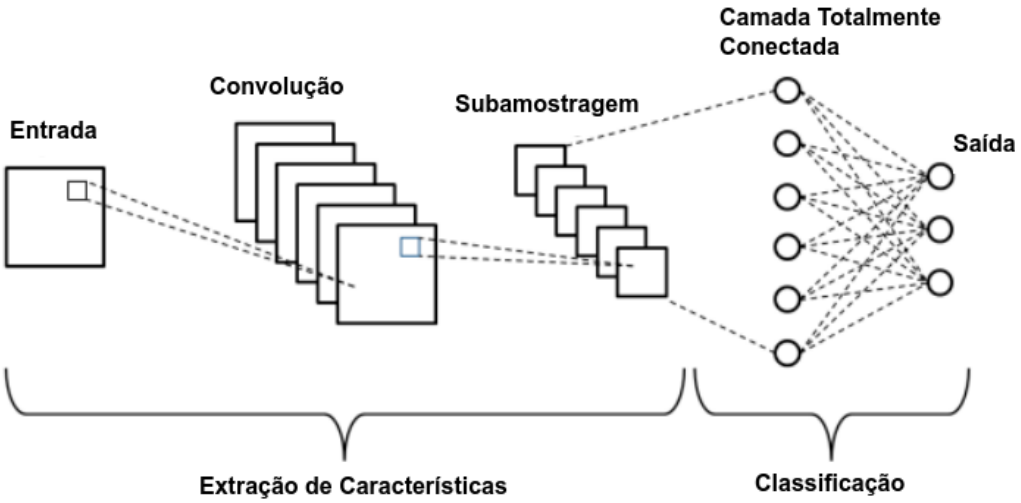

Figura 2. Diagrama da arquitetura básica de uma CNN.

Fonte: Adaptado de [Phung and Rhee 2019].

Vale destacar que, embora imagens sejam percebidas visualmente por seres humanos, elas são representadas numericamente por computadores. Uma imagem digital em tons de cinza, por exemplo, é modelada como uma matriz bidimensional, na qual cada elemento corresponde à intensidade de brilho de um *pixel*, comumente variando de 0 (preto) a 255 (branco). Já imagens coloridas são compostas por três matrizes — uma para cada canal de cor (vermelho, verde e azul). Esse formato matricial é essencial para que operações matemáticas, como a convolução, possam ser aplicadas. Assim, durante o processamento em CNNs, tanto as entradas (imagens) quanto os filtros (*kernels*) são tratados como estruturas numéricas, permitindo a multiplicação e soma de seus valores ao longo do deslocamento da janela convolucional. Essa conversão da imagem para números é o que possibilita à rede extrair padrões, bordas, formas e texturas relevantes para tarefas como reconhecimento de objetos e classificação de imagens.

A operação central em uma CNN é a convolução, que combina duas funções matemáticas para produzir uma terceira. Ela é obtida ao somar os produtos das funções sobre uma região delimitada onde ambas se sobrepõem, levando-se em conta o deslocamento relativo entre elas. A convolução discreta é formalmente definida da seguinte forma:

$$(f * g)(n) = h(n) = \sum_{j=0}^{n} f(j) \cdot g(n - j) \tag{1}$$

onde:

- f é a função de entrada (por exemplo, uma imagem),
- g é o filtro ou *kernel*,
- $h(n)$ é o resultado da convolução na posição n.

Ambas as funções f e g são sequências finitas de tamanho k, e o resultado $h(n)$ também será uma sequência.

Operação com Janela Deslizante

Em termos práticos, a operação de convolução em uma CNN pode ser compreendida como o deslocamento S de uma janela $W = \{w_1, w_2, ..., w_m\}$ sobre os valores de uma entrada $I = \{i_0, i_1, ..., i_n\}$, onde $S \in N^+$ e $S \leq n$. Aqui, m e n representam os tamanhos da janela (filtro) e da entrada, respectivamente, com $m \leq n$.

Durante esse deslocamento, ocorre a multiplicação elemento por elemento entre os valores da janela e os valores correspondentes na entrada, seguida da soma dos produtos resultantes. Essa operação é essencial para a extração de características nas camadas convolucionais das redes neurais.

Exemplo Numérico da Convolução em 1D

Considere a seguinte entrada (n=5) e o filtro (m=3):

$$Entrada(\textbf{\textit{Input}}) : [1, 0, 2, 3] \qquad Filtro(\textbf{\textit{Kernel}}) : [1, 0, 1]$$

Aplicando a convolução com *stride* S=1 (deslocamento de um elemento por vez), temos os seguintes cálculos:

$$(1 \cdot 1) + (0 \cdot 0) + (2 \cdot 1) = 1 + 0 + 2 = 3$$
$$(0 \cdot 1) + (2 \cdot 0) + (3 \cdot 1) = 0 + 0 + 3 = 3$$

Logo, a saída da convolução (output) será:

$$Saída(Output) : [3, 3]$$

Esse processo é repetido para cada posição válida da janela sobre a entrada, formando o mapa de características extraído pela camada convolucional. A convolução assume que características relevantes tendem a estar agrupadas localmente, e por isso a operação é eficaz principalmente para capturar padrões de proximidade entre elementos vizinhos.

Nos últimos anos, a aplicação de CNNs para análise automatizada de lâminas citológicas tem ganhado destaque. Estudos demonstram que modelos de Aprendizado Profundo podem alcançar desempenho comparável ao de especialistas humanos em tarefas como detecção de anormalidades celulares e classificação de tipos celulares ([Zhang et al. 2017, Xu et al. 2020])).

Por exemplo, CNNs têm sido usadas para automatizar a triagem de lâminas de Papanicolau digitalizadas, com o potencial de aumentar a acurácia diagnóstica e reduzir o tempo de análise ([Bora et al. 2017].). Essas abordagens, no entanto, ainda enfrentam desafios como variabilidade na qualidade das imagens, necessidade de grandes bases de dados rotuladas e validação clínica rigorosa.

1.2. Transformers

O modelo *Transformer* representa uma arquitetura de rede neural introduzida por [Vaswani et al. 2017]), que marcou uma revolução no campo do aprendizado profundo e revolucionou o de processamento de linguagem natural, especialmente em tarefas com

dados sequenciais. O principal diferencial dessa arquitetura é o uso da autoatenção, que permite ao modelo analisar e relacionar diferentes partes da entrada ao mesmo tempo, sem a necessidade de processá-las em ordem, capturando contextos amplos com eficiência. Isso torna os *Transformers* altamente paralelizáveis e eficazes na identificação de dependências de longo alcance.

O mecanismo de autoatenção se baseia em três conjuntos de vetores:

- **Consultas (*queries*)**: representam o elemento da sequência que busca informações relevantes nos demais.
- **Chaves (*keys*)**: representam os elementos da sequência que serão comparados com a consulta para medir relevância.
- **Valores (*values*)**: são os vetores efetivamente combinados, com base na relevância determinada pela comparação entre consultas e chaves.

No caso do *Transformer*, esses três vetores são obtidos a partir da mesma entrada (por isso o termo "autoatenção"), o que permite que cada elemento da sequência compare-se com todos os demais, inclusive consigo mesmo, para extrair relações contextuais úteis.

Matematicamente, a operação de autoatenção no *Transformer* pode ser descrita da seguinte forma: dado um conjunto de vetores de consulta Q, chaves K e valores V, a atenção escalonada é calculada por:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

onde:

- $Q \in R^{m \times d_k}$ representa o conjunto de vetores de consulta (*queries*), sendo m o número de consultas e d_k a dimensão dos vetores de consulta.
- $K \in R^{n \times d_k}$ é o conjunto de vetores de chaves (*keys*), onde n é o número de chaves e d_k é a dimensão dos vetores de chave.
- $V \in R^{n \times d_v}$ são os vetores de valores (*values*), com n sendo o número de valores e d_v a dimensão dos vetores de valor.
- O termo QK^T refere-se ao produto escalar entre as consultas e as chaves, que mede a similaridade entre os vetores de consulta Q e os vetores de chave K. Isso resulta em uma matriz de similaridade de dimensões $m \times n$, que representa a relevância de cada chave para uma consulta específica.

O fator de escalonamento $\frac{1}{\sqrt{d_k}}$ é aplicado para evitar que o produto escalar QK^T gere valores excessivamente grandes à medida que a dimensão d_k aumenta. Sem esse escalonamento, a magnitude dos valores do produto escalar poderia crescer desproporcionalmente, resultando em uma função softmax que se tornaria muito sensível a pequenas variações.

A função *softmax* é aplicada ao resultado de $\frac{QK^T}{\sqrt{d_k}}$, normalizando os pesos de atenção para que somem 1. Isso garante que cada valor associado ao vetor V seja ponderado de acordo com sua importância relativa em relação à consulta.

Após a aplicação da *softmax*, esses pesos de atenção são usados para ponderar os vetores de valor V, resultando em uma nova representação. Essa nova representação

captura as informações mais relevantes para cada posição da sequência de entrada, o que é essencial nas tarefas de atenção subsequentes do *Transformer* ([Vaswani et al. 2017])).

A operação realizada na Equação 2 ocorre várias vezes em paralelo por meio do mecanismo de *multi-head attention*, que permite ao modelo aprender diversos tipos de relacionamentos entre os elementos da sequência de entrada, em diferentes níveis de granularidade ([Vaswani et al. 2017])). Esse mecanismo é central para a robustez do *Transformer* em tarefas de sequência, pois o modelo pode focar em múltiplas partes relevantes da entrada simultaneamente.

Embora inicialmente projetados para tarefas textuais, os *Transformers* têm sido adaptados com sucesso para o domínio da visão computacional, especialmente com a introdução do *Vision Transformer* (ViT) ([Dosovitskiy et al. 2020])). Com o ViT, a imagem é dividida em pequenos blocos fixos chamados em inglês de *patches*, que são transformados em vetores de características, de forma análoga às palavras em uma frase. Esses vetores passam por *embeddings* posicionais, que preservam a informação espacial da imagem, e são então processados pelo mecanismo de autoatenção, permitindo que o modelo aprenda relações globais entre diferentes regiões da imagem. Essa abordagem altamente paralelizável tem demonstrado desempenho competitivo — e por vezes superior — em tarefas clássicas como classificação, detecção e segmentação de imagens, especialmente em contextos nos quais as relações espaciais de longo alcance são importantes.

Conforme ilustrado na Figura 3, o processamento da imagem pelo ViT inicia-se com a divisão da imagem de entrada em pequenos blocos fixos, denominados *patches*. Cada *patch* é achatado (*flattened*) e transformado em um vetor de características (*embedding*). Em seguida, esses vetores são projetados em um espaço vetorial de dimensão fixa por meio de uma camada linear. Para preservar a informação da ordem espacial dos *patches*, vetores de codificação posicional absolutos são adicionados aos *embeddings*.

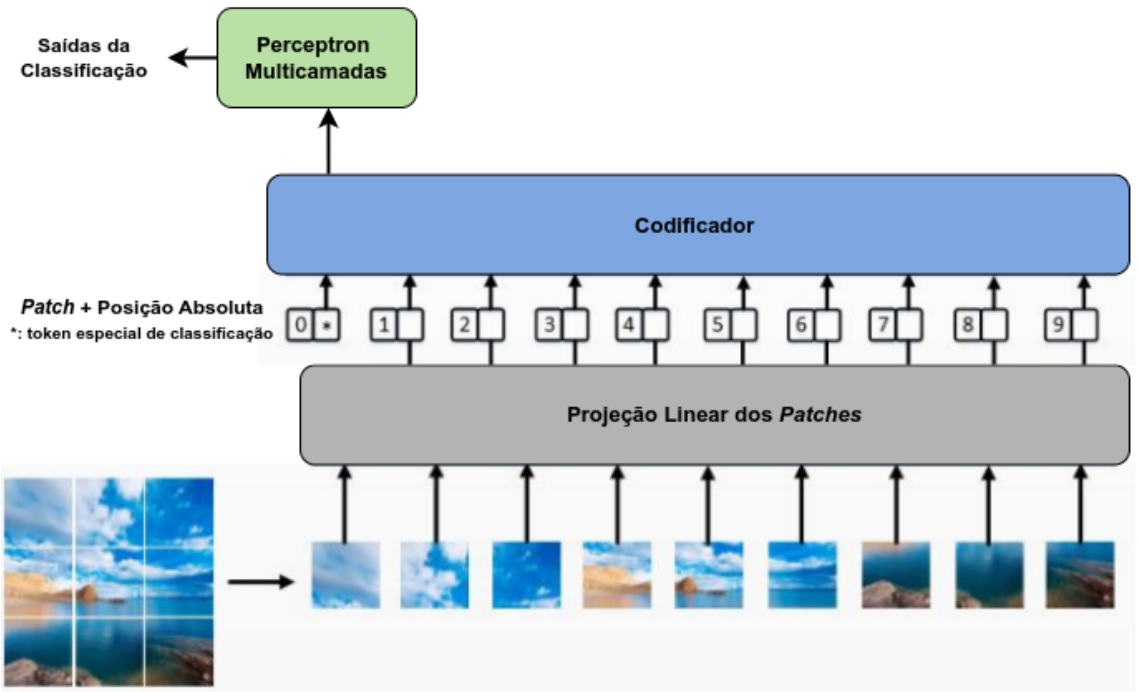

Figura 3. Diagrama do Modelo de ViT.

Fonte: Adaptado de [Liang et al. 2021].

Em seguida, os vetores dos *patches*, juntamente com um *token* especial de classificação inserido na primeira posição da sequência, são fornecidos ao codificador. Esse *token* é um vetor que tem a função de agregar informações globais da imagem ao longo das camadas do codificador, permitindo que o modelo, ao final do processamento, utilize essa representação como um resumo da imagem.

O codificador do modelo (representado pela área azul na Figura 3) é composto por múltiplas camadas que contêm o mecanismo de autoatenção do tipo *multi-head* — o núcleo da arquitetura *Transformer*. A autoatenção permite que o modelo avalie e relacione cada bloco com todos os outros da imagem, aprendendo padrões e dependências espaciais globais. Em termos práticos, essa etapa realiza internamente os cálculos descritos na Equação 2, possibilitando a modelagem eficiente das relações entre regiões distantes da imagem. Após o processamento pelo codificador, a representação correspondente ao *token* de classificação é extraída e passada por um perceptron multicamadas (bloco verde), que atua como classificador e gera a saída final de predição da classe da imagem.

Na citologia convencional, os *Transformers* vêm sendo explorados para superar limitações dos métodos baseados em CNNs, como o campo de visão local limitado e a dificuldade em modelar relações espaciais de longo alcance entre estruturas celulares. Modelos híbridos (como CNN-ViT) e arquiteturas puramente baseadas em atenção têm sido aplicados na classificação de imagens de esfregaços citológicos, com resultados promissores ([Li et al. 2022, Chen et al. 2022]).)

Estudos recentes indicam que *Transformers* podem oferecer vantagens importantes na triagem automatizada de exames citológicos, como os de Papanicolau, fornecendo maior sensibilidade na detecção de anomalias e facilitando a integração de múltiplas resoluções e contextos espaciais ([Li et al. 2022]).) No entanto, como nas CNNs, essas abordagens enfrentam desafios, como a escassez de dados rotulados e a necessidade de validação clínica rigorosa antes da adoção em ambientes hospitalares.

2. Métricas de Avaliação

As métricas de avaliação como precisão, revocação, *F1-Score* e acurácia são fundamentais para medir a qualidade de modelos de classificação. Segundo Manning e Schütze ([Manning and Schütze 1999])), em bases de dados desbalanceadas, a acurácia pode ser uma métrica enganosa. Nesses casos, métricas como precisão, revocação e *F1-Score* são mais adequadas para refletir o desempenho real do modelo.

2.1. Precisão

A precisão mede quantas das previsões positivas feitas pelo modelo realmente eram positivas:

$$Precisão = \frac{VerdadeirosPositivos}{VerdadeirosPositivos + FalsosPositivos} \tag{3}$$

2.2. Revocação (*Recall*)

A revocação mede quantos dos exemplos positivos reais foram capturados pelo modelo:

$$Revocação = \frac{VerdadeirosPositivos}{VerdadeirosPositivos + FalsosNegativos} \tag{4}$$

2.3. *F1-Score*

O *F1-Score* é a média harmônica entre precisão e revocação, buscando equilibrar esses dois aspectos:

$$F1 = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{5}$$

2.4. Acurácia

A acurácia mede a proporção de todas as previsões corretas feitas pelo modelo:

$$Acurácia = \frac{VerdadeirosPositivos + VerdadeirosNegativos}{TotaldeAmostras} \tag{6}$$

3. Justificativa

Citopatologistas identificam lesões celulares com base em alterações morfológicas específicas, como a desproporção na relação núcleo/citoplasma, modificações na distribuição da cromatina, hipercromasia e irregularidades na membrana nuclear. Essas lesões costumam apresentar aumento do núcleo, variações em sua forma e textura, além de condensação cromatínica anormal ou distribuição irregular da cromatina ([Plissiti et al. 2011])).

"CRIC Cervix" é uma coleção de imagens obtidas a partir de Papanicoloau convencional formulada pelo grupo de pesquisa. Esta possui 400 imagens de lâminas e 11.534 células segmentadas, rotuladas e classificadas, como o exemplo da Figura 4. Seu diferencial das duas outras bases abertas existentes para imagens de Papanicolaou (a saber, Herlev e SIPaKMeD) é que a CRIC Cervix possui as classificações celulares de acordo com a nomenclatura do Sistema Bethesda ([**?**])).

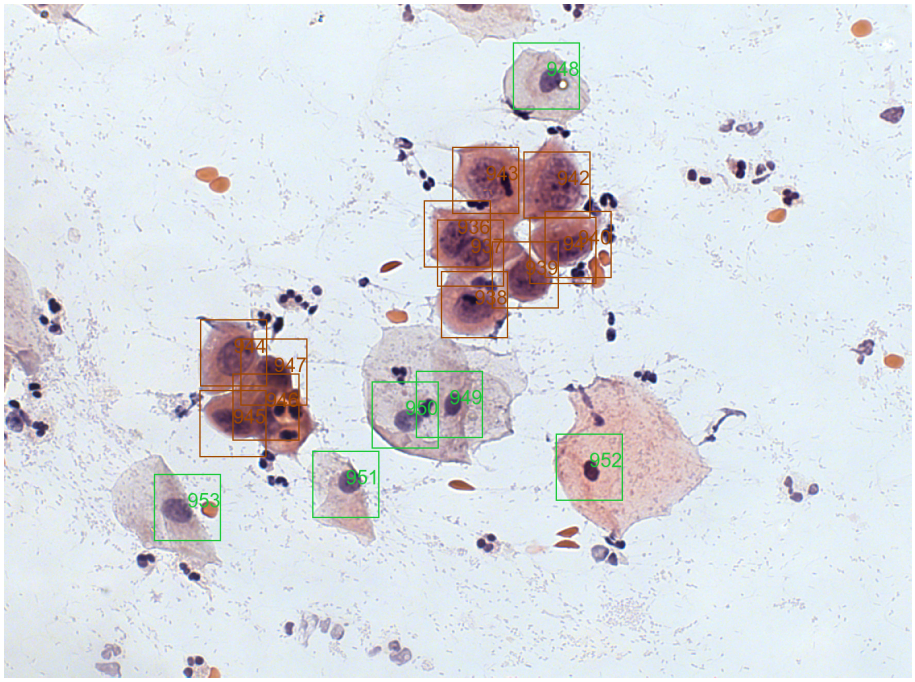

Figura 4. Imagem nº 383 da base de dados CRIC - lâmina de microscópio do colo do útero rotulada

Além disso, a CRIC Cervix é a única das demais bases de dados que apresenta imagens multicelulares - a fim de garantir um cenário real ao que é encontrado pelos profissionais citopatologistas durante a análise do esfregaço, sem imagens com células recortadas individualmente e/ou irrealisticamente escolhidas de acordo com a estrutura celular.

Foi proposto pelo grupo um método de *ensemble* (técnica de combinação de modelos) utilizando dez arquiteturas de CNNs para a classificação de células cervicais da base de dados CRIC Cervix, seguindo o Sistema Bethesda. O método apresenta uma acurácia de 95%, precisão, revocação e *F1-Score* de 85% para a classificação simultânea de seis classes. Embora o método represente o estado da arte entre os modelos que utilizam CNNs nesta base de dados, os resultados ainda podem ser otimizados ([Diniz et al. 2021])).

O diagnóstico citológico clínico correto do câncer anal como primeira instância diagnóstica pode permitir uma abordagem terapêutica mais rápida e direcionada e retardar o avanço da doença. Diante disso, a proposta deste trabalho é desenvolver um modelo de classificação de células cervicais que possa ser adaptado para a análise das células do canal anal. A ausência de consenso sobre a triagem de lesões anais em populações de risco resulta na falta de bases de dados disponíveis para o treinamento desse modelo, e a base de citologia anal ainda está em desenvolvimento pelo grupo de pesquisa.

Este trabalho de qualificação foi estruturado essencialmente em duas partes: (1) uma revisão sistemática sobre o tema datada de outubro de 2024 e, (2) com base nestes resultados, a delimitação do plano de ação.

O objetivo da revisão sistemática foi estudar como o modelo de ViT vem sendo empregado na classificação de células cervicais e anais, bem como investigar a possibilidade de transferência de aprendizado. 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.

4. Revisão Bibliográfica

A revisão sistemática foi realizada em outubro de 2024 para entender sobre o emprego de ViT em lesões cervicais e anais. 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 Aprendizado Profundo 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 e (4) Web of Science (Figura 5). 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.

Conforme a leitura dos artigos, os critérios de exclusão foram sendo delimitados em: (1) documentos duplicados; (2) literatura cinza - documentos que não passaram por canais formais de publicação científica, os quais, em geral, não são revisados por pares e apresentam menor rigor metodológico; (3) artigos cuja menção a ViT em células-alvo foi extremamente superficial, não sendo essas células empregadas no desenvolvimento do modelo; (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.

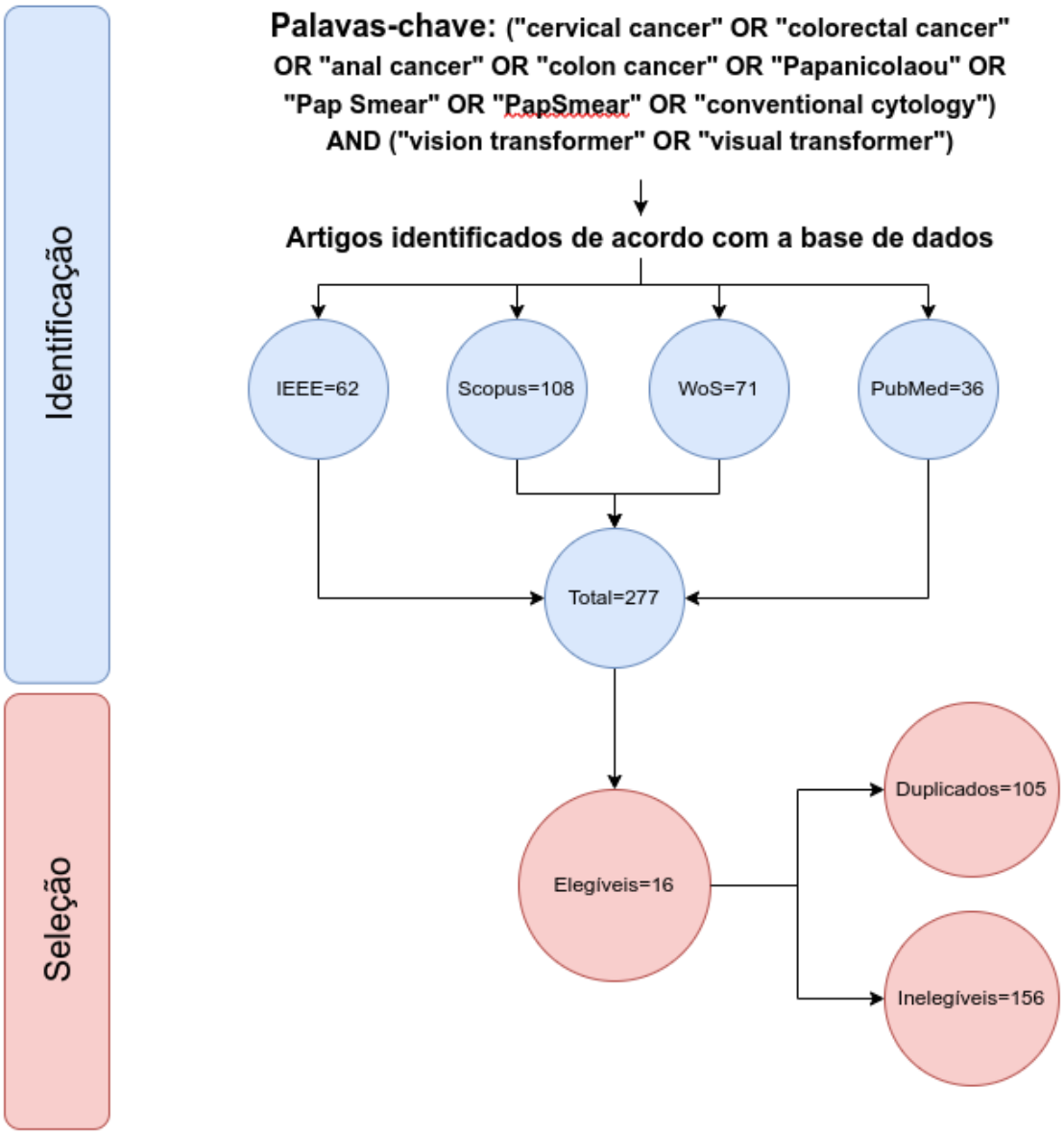

Figura 5. Descrição da Revisão Bibliográfica

Ao todo, foram importados 277 artigos, sendo eles: IEEE – 62, Pubmed – 36, Scopus – 108, Web of Science – 71 (Figuras 6 e 7). Dos artigos identificados, 16 foram selecionados para leitura com base nos critérios de inclusão e exclusão estabelecidos.

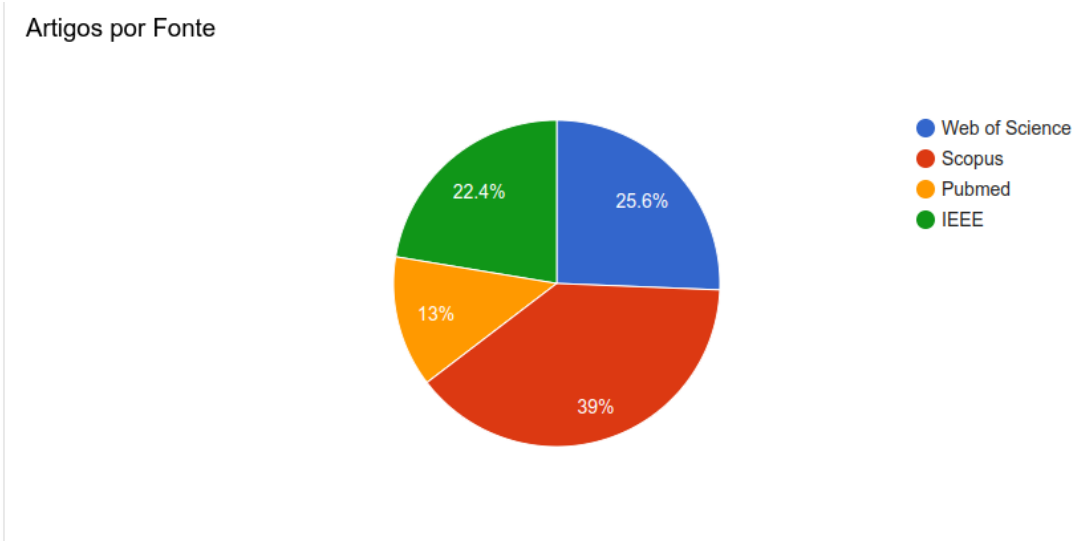

Figura 6. Artigos selecionados por fonte

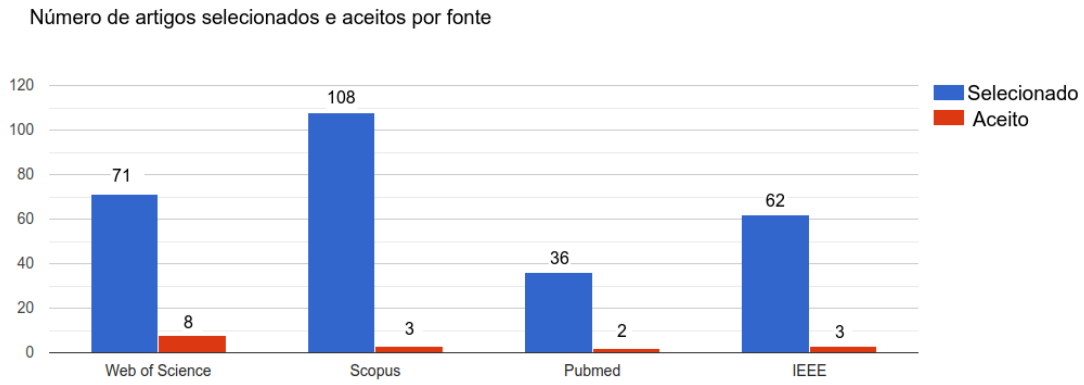

Figura 7. Artigos selecionados e aceitos por fonte

Os estudos selecionados foram publicados entre os anos de 2022 e 2024 (Figura 2.4).

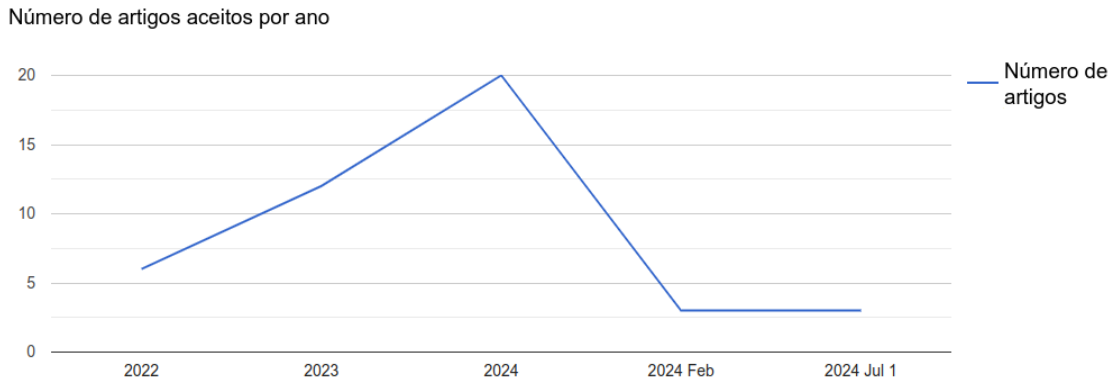

Figura 8. Artigos aceitos por ano

4.1. Resultados e Discussão

Estudos recentes exploram diferentes variações do ViT e sua integração com CNNs para aprimorar a classificação de imagens cervicais e a detecção de câncer. As abordagens analisadas combinam as características locais extraídas pelas CNNs com a habilidade dos ViTs em modelar dependências globais, buscando aumentar tanto a precisão quanto a eficiência computacional.

As bases de dados utilizadas nos artigos analisados foram: SIPaKMeD, Herlev e CRIC, sendo as duas primeiras de imagens com células unicelulares, ou seja, o número de imagens é igual ao número de células (Tabela 1).

Tabela 1. Bases de Dados e Classificações Celulares

Base de Dados	Número de Imagens/Células	Classificações Celulares
SIPaKMeD	4.049/4.049**	(a) Superficial-Intermediário, (b) Parabasal, (c) Koilocitótica, (d) Disqueratótica, (e) Metaplásica
Herlev	917/917**	(a) Células Normais Superficiais, (b) Células Intermediárias Normais, (c) Células Parabasais Normais, (d) Lesão Intraepitelial de Baixo Grau (LSIL), (e) Lesão Intraepitelial de Alto Grau (HSIL), (f) Carcinoma In Situ, (g) Carcinoma Escamoso Invasivo
CRIC	400/11.534	(a) Células escamosas atípicas, não podem excluir uma lesão de alto grau (ASC-H); (b) células escamosas atípicas de significado indeterminado, possivelmente não neoplásicas (ASC-US); (c) lesão intraepitelial escamosa de alto grau (HSIL); (d) lesão intraepitelial escamosa de baixo grau (LSIL); (e) negativo para lesão intraepitelial ou malignidade (NILM); (f) carcinoma espinocelular (CEC)

As métricas de desempenho com destaque dos artigos (acima de 90%) foram divididas em duas tabelas, de acordo com as bases de dados empregadas - SIPaKMeD e Herlev ou CRIC (Tabelas 2 e 3).

Tabela 2. 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%

Nota: O método CervixFuzzyFusion empregou células ”segm.”(previamente segmentadas à classificação) ou ”não segm.”(não segmentadas), conforme indicado entre parênteses.

”Acur.”é a abreviatura para Acurácia, ”Prec.” para Precisão, ”Rev.” para Revocação e ”F1” para *F1-Score*.

Tabela 3. 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]	DIFF (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%

Nota: O modelo ViT-PSO-SVM empregou classificação binária (Herlev - Binário) entre lesão ou não lesão nas imagens da base de dados, conforme indicado entre parênteses.

Alguns modelos se destacaram em relação às métricas de desempenho, indicando potencial para aplicação em diagnóstico de citologia convencional cervical.

Em relação a base de dados SipakMed, cabe destacar alguns artigos:

MaxCerVixT – Alcançou acurácia de 99,02%, precisão de 99,03%, revocação de 99,04% e *F1-score* de 99,02% na detecção de câncer cervical ([Pacal 2024])). Esse modelo propôs uma versão otimizada e leve da arquitetura MaxViT, especificamente adaptada para imagens de citologia com resolução limitada. Entre as principais modificações estruturais foram a substituição dos blocos MBConv por blocos ConvNeXtv2, que oferecem maior eficiência na extração de características com menos parâmetros, e o uso de camadas MLP baseadas em *Global Response Normalization*, que melhoram a seletividade e a estabilidade do treinamento. O MaxCerVixT também incorporou atenção multi-eixo local e global por meio de mencanismos de janela e grade respectivamente, promovendo um aprendizado mais robusto de padrões morfológicos das células. Sua alta acurácia e rapidez na inferência destacam o modelo como uma alternativa eficaz e viável.

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 proposta integrou a arquitetura convolucional EfficientNet-B0 e ViT. As características extraídas por ambas as redes foram combinadas e processadas por uma Máquina de Aprendizado Extremo, um classificador de camada única conhecido por sua alta capacidade de generalização e velocidade de treinamento. Essa combinação híbrida explora as forças complementares das arquiteturas CNN e ViT, resultando em uma abordagem eficiente, leve e altamente precisa.

HViT-Cerv – Este modelo híbrido, que combina CNN e ViT, obteve média de acurácia, precisão, revocação e *F1-score* em torno de 98% ([Emara et al. 2024])). O artigo propôs três abordagens principais: modelos CNN tradicionais, uma arquitetura Vision Transformer (ViT-Cerv), e um modelo híbrido chamado HViT-Cerv. Este último obteve melhor desempenho, e foi avaliado utilizando as CNNs EfficientNetB0, DenseNet121, Xception e ResNet50, sendo que a fusão entre a EfficientNetB0 e o ViT obteve o melhor resultado (descrito acima).

CervixFuzzyFusion – O modelo obteve acurácia de 98,13% na classificação de células segmentadas e 96,13% em imagens não segmentadas, com altos valores de precisão e *F1-score* em ambos os casos ([Hemalatha et al. 2023])). Sua arquitetura combina características locais extraídas por uma CNN (DenseNet201) e características globais obtidas por um modelo de ViT, com mecanismos de *Shifted Patch Tokenization* (SPT) e *Locality Self-Attention* (LSA). As características fundidas são refinadas por um processo de seleção baseado em lógica *fuzzy* (FFS, em inglês *Fuzzy Feature Selection*), que elimina atributos redundantes e melhora a eficiência do modelo. Essa abordagem híbrida se destaca por dispensar a necessidade de segmentação manual e por sua capacidade de generalizar bem mesmo em conjuntos de dados limitados.

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

3D CNN-ViT com classificador KELM - Esse modelo combinou a capacidade de extração de características espaciais e temporais da 3D CNN com a habilidade do ViT de capturar representações visuais complexas. Sua arquitetura incluiu 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 foram então concatenadas por uma Rede de Pirâmide

de Características 3D - uma CNN, aprimoradas por um bloco de exaltação 3D, outra CNN e, finalmente, classificadas por uma Máquina de Aprendizado Extremo com *Kernel*. Os resultados demonstraram uma alta precisão de 98,6% ([Abinaya and Sivakumar 2024])).

ViT-PSO-SVM – Alcançando acurácia de 97,78% no conjunto Herlev e até 99,21% no conjunto SIPaKMeD, o modelo propôs uma abordagem híbrida que integrou o ViT para extração de características locais e globais, o algoritmo *Particle Swarm Optimization* (PSO) para seleção das características mais relevantes e redução de dimensionalidade, e um classificador *Support Vector Machine* (SVM) na etapa final de decisão ([AlMohimeed et al. 2024])). Essa combinação permitiu ao modelo manter alta precisão, mesmo em cenários com dados limitados ou classes múltiplas, superando redes convolucionais e *Trannsfomers* isolados. Além disso, o uso de mapas de ativação (Grad-CAM) contribuiu para a interpretabilidade do modelo, na geração de mapas de calor que mostram as regiões da imagem que mais influenciaram a classificação.

CNN+ViT com PCA e LDA – Neste estudo, os autores avaliaram diversas combinações de arquiteturas para extração de características, utilizando modelos de CNNs (ResNet-50, DenseNet-121, VGG-16 e Inception-V3) para capturar informações locais e modelos de ViTs (ViT-B16, ViT-L16, ViT-B32 e ViT-L32) para capturar informações globais. Cada combinação foi testada individualmente, com as características extraídas sendo concatenadas e normalizadas. Para mitigar a alta dimensionalidade e eliminar redundâncias, aplicou-se uma etapa dupla de redução com Análise de Componentes Principais (ACP) seguida de Análise Discriminante Linear (ADL). As características reduzidas foram classificadas utilizando algoritmos tradicionais como SVM, K-NN, MLP e regressão logística. A melhor performance foi obtida com o uso da combinação CNN+ViT seguida por ACP, ADL e classificador SVM, atingindo acurácia de 97,83% e precisão de 98,75% no conjunto Herlev ([Sholik et al. 2024])). No entanto, o artigo não especifica qual par de modelos (CNN e ViT) foi responsável por esse resultado, reportando os dados apenas de forma agregada. Ainda assim, os resultados evidenciam a eficácia da fusão entre características locais e globais para representar células cervicais em imagens de Papanicolau.

Os modelos citados acima demonstram alto potencial para o diagnóstico de citologia cervical, com ótimos resultados observados através das métricas de desemepenho, e estratégias inovadoras de fusão e otimização de características.

Foram identificados dois estudos que empregaram ViT sobre a base de dados CRIC. O primeiro, CVM-Cervix, é um método híbrido para classificação de imagens de células cervicais usando uma combinação de CNN - para a extração de características locais, o ViT na captura de características globais, e um Perceptron Multicamadas para a classificação final. O método foi treinado em uma base de dados composta pela união dos datasets CRIC e SIPaKMeD, totalizando 8838 imagens divididas em 11 classes. As imagens foram aumentadas e normalizadas na etapa de pré-processamento. Na avaliação, o CVM-Cervix obteve uma acurácia de 91,72% no conjunto de teste, com precisão média de 91,80%, revocação de 91,60% e *F1-Score* de 91,70%, superando modelos tradicionais como DenseNet, ResNet e o ViT sem agregação a outro modelo ([Liu et al. 2022])).

Já o segundo artigo utilizou o modelo Deep Integrated Feature Fusion (abreviado nas Tabelas 2.2 e 2.3 como DIFF), que propõe uma arquitetura híbrida com dois

ramos paralelos: um baseado em CNN para extração de características locais e outro baseado em ViT para captura de informações globais. A integração entre essas duas representações é realizada por meio de blocos específicos de fusão, denominados DIFF blocks. Esses blocos combinam interativamente os mapas de características locais e globais por concatenação, convoluções 1×1 e 3×3, e conexões residuais, promovendo uma fusão mais eficaz entre os dois tipos de informação. O método obteve um desempenho de destaque na base CRIC, com acurácia de 85,06%, precisão de 85,47%, revocação de 85,23% e *F1-Score* de 85,12% na tarefa de classificação multiclasse com seis categorias ([Fang et al. 2024])).

4.2. 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, nove 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])).

Diante desses resultados, evidencia-se a necessidade de empregar métodos de classificação mais eficazes para a base de dados CRIC, utilizando o Sistema Bethesda, com o objetivo de desenvolver um modelo de triagem que proporcione maior robustez e segurança ao paciente.

5. Experimentos

5.1. Base de Dados

Os métodos propostos serão avaliados utilizando a base de dados **CRIC Cervix** (`https://database.cric.com.br`), acessada em 7 de maio de 2025. Essa coleção contém 400 imagens de lâminas de esfregaço cervical convencional (exame de Papanicolau), adquiridas por microscopia de campo claro com objetiva de 40x e ocular de 10x. As imagens foram capturadas por uma câmera digital Zeiss AxioCam MRc acoplada a um microscópio Zeiss AxioImager Z2, utilizando o software Axio Vision Zeiss.

A base conta com um total de 11.534 células rotuladas, representadas por caixas delimitadoras (*bounding boxes*) de 100 *pixels* $\times$ 100 *pixels*. Quatro citopatologistas realizaram a anotação e revisão das células, utilizando a nomenclatura do Sistema Bethesda para classificação de lesões celulares.

Como mencionado anteriormente, dentre as 11.534 células rotuladas, 4.755 foram classificadas como positivas para lesão celular e 6.779 como negativas. As imagens já se encontram segmentadas, com cada célula devidamente demarcada e rotulada. A partir dessas segmentações, as células foram extraídas por meio de regiões de interesse de 70 *pixels* $\times$ 70 *pixels*, centradas nas coordenadas do núcleo celular. Essa dimensão foi definida com base em estudos prévios, os quais indicam que ela é adequada para a representação morfológica e a extração eficiente de características do núcleo celular. Para manter a consistência no tamanho das amostras, 425 células foram descartadas por estarem fora dos limites definidos.

5.2. Extração de Atributos Texturais

Embora as CNNs sejam capazes de aprender representações espaciais e texturais diretamente das imagens, este trabalho adota uma abordagem alternativa ao integrar a extração manual de atributos texturais no processo de classificação. A principal motivação decorre da análise dos artigos incluídos na revisão bibliográfica, os quais utilizaram predominantemente modelos *end-to-end*, nos quais toda a representação é aprendida automaticamente, sem qualquer intervenção explícita na seleção ou extração de características. Essa predominância de abordagens puramente profundas sugere a possibilidade de que informações relevantes — especialmente texturais — estejam sendo subutilizadas.

A inclusão de atributos texturais extraídos manualmente, como aqueles derivados de GLCM e LBP, busca enriquecer a representação das imagens com informações complementares que podem não ser facilmente captadas por uma CNN. Esses atributos oferecem uma descrição direta de padrões como contraste, homogeneidade e entropia, que podem auxiliar o modelo a discriminar entre classes sutilmente distintas, como é o caso da morfologia de células cervicais. Além disso, essa abordagem favorece a interpretabilidade do modelo, uma vez que permite identificar quais características específicas da textura estão influenciando as decisões do classificador — aspecto especialmente relevante em aplicações biomédicas, onde a transparência é frequentemente desejável.

Portanto, a opção pela extração manual de atributos texturais representa não apenas uma forma de potencializar o desempenho do modelo em cenários com dados limitados, mas também uma contribuição metodológica relevante, dado que esse tipo de abordagem permanece pouco explorado na literatura atual sobre classificação de células por meio de aprendizado profundo.

Para viabilizar a extração de atributos texturais, todas as imagens foram convertidas para escala de cinza. Em seguida, foram extraídas quatro propriedades da matriz de coocorrência de níveis de cinza (GLCM): contraste, correlação, energia e homogeneidade. Também foram gerados histogramas do padrão binário local (LBP), com 10 *bins* uniformes, e extraídos 13 atributos de Haralick utilizando a biblioteca `mahotas`. Todos esses atributos foram concatenados em um vetor de características com 27 valores por imagem (4 GLCM + 10 LBP + 13 Haralick) e armazenados em arquivos `.csv`, juntamente com o identificador da célula e seu respectivo rótulo.

5.3. Modelos

Com base na revisão bibligráfica, os modelos selecionados para execução na base CRIC são o *MaxCerVixT* ([Pacal 2024])), e ViT combinado com a CNN EfficientNetB0 ([Khowaja et al. 2023], [Emara et al. 2024])). Inicialmente, o foco será na obtenção do

melhor desempenho com o modelo ViT + EfficientNetB0; posteriormente, o modelo de *MaxCerVixT* será avaliado conforme a disponibilidade de tempo.

Os experimentos foram organizados em três etapas: treino, validação e teste. O conjunto de treino é composto por 70% das células (3.300 positivas e 4.505 negativas), enquanto os conjuntos de validação e teste utilizam, respectivamente, 15% das amostras (709 positivas e 942 negativas para validação; 704 positivas e 949 negativas para teste).

A classificação das imagens citopatológicas será realizada de forma progressiva, buscando otimizar os parâmetros dos modelos. Inicialmente, será utilizada uma abordagem binária: negativo para lesão celular (NILM) e positivo para lesão (ASC-US, ASC-H, LSIL, HSIL, CEC). Em seguida, será adotada uma classificação em três classes: negativo (NILM), atipia indefinida ou limítrofe (ASC-US e ASC-H), e lesão confirmada (LSIL, HSIL e CEC). Por fim, será empregada a classificação completa com as seis categorias do Sistema Bethesda.

6. Resultados e Discussão

7. Conclusões

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.

Albuquerque, A. and Fontes, F. (2025). Recent guidelines on anal cancer screening: A systematic review. *Journal of Lower Genital Tract Disease*, 29(2):180–185.

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.

Bernal, M., Bonono, C.-R., El-Halabi, S., Martí, J., Mwaka, A. D., Vassilakos, P., Jeronimo, J., and Gülmezoglu, A. M. (2023). Who recommendations and good practice statements on screening for cervical cancer. *International Journal of Gynecology & Obstetrics*, 160(S1):5–17.

Bora, K., Chowdhury, M., Mahanta, L. B., Kundu, M., and Nasipuri, M. (2017). Pap smear image classification using convolutional neural network. In *Procedia Computer Science*, volume 132, pages 1063–1070. Elsevier.

Bray, F., Ferlay, J., Soerjomataram, I., and et al. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Campos, I. S. and Carvalho, L. R. B. (2024). Dificuldades para o rastreio do câncer do colo de útero no brasil: uma revisão de literatura. *Cadernos de Ensino e Pesquisa em Saúde*, 4(2):20–32.

Chandwani, V., Saraswathi, K., Rao, G. B., and Sivastava, V. (2021). Liquid based cytology versus conventional cytology for evaluation of cervical pap smear. *Journal of Pharmaceutical Research International*, 33(45B):27–33.

Chen, R., Lu, M., Zhang, J., Zhan, X., Zhao, Y., Wang, J., Zhu, Y., and Wang, L. (2022). Transpath: Transformer-based self-supervised learning for pathology image analysis. *Medical Image Analysis*, 77:102371.

Claro, I. B. et al. (2021). Análise dos motivos de insatisfatoriedade dos exames histopatológicos do colo do útero no sistema Único de saúde, brasil, 2014 a 2017. *Revista Brasileira de Cancerologia*, 67(3):e–081299.

Diniz, D., Rezende, M., Campos, A. G., and Souza, M. J. (2021). A deep learning ensemble method to assist cytopathologists in pap test image classification. *Journal of Imaging*, 7(7):111.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dyer, C. E. F., Jin, F., Hillman, R. J., Nyitray, A. G., Roberts, J. M., Law, C., Grulich, A. E., and Poynten, I. M. (2025). Self-collected versus clinician-collected anal swabs for anal cancer screening: A systematic review and meta-analysis. *International Journal of Cancer*, 156(1):79–90.

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.

Hemalatha, K., Vetriselvi, V., Dhandapani, M., and Gladys, A. A. (2023). Cervixfuzzyfusion for cervical cancer cell image classification. *Biomedical Signal Processing and Control*, 85:104920.

Hopp, A. M., Puzyrenko, A., and Giorgadze, T. (2023). Comprehensive review of anal cytology. *Acta Cytologica*, 67(2):185–194.

Hou, X., Shen, G., Zhou, L., Li, Y., Wang, T., and Ma, X. (2022). Artificial intelligence in cervical cancer screening and diagnosis. *Frontiers in Oncology*, 12:851367.

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.

Klas, J. V., Rothenberger, D. A., Wong, W. D., and Madoff, R. D. (1999). Malignant tumors of the anal canal: the spectrum of disease, treatment, and outcomes. *Cancer*, 85(8):1686–1693.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105. Curran Associates, Inc.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

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).

Li, T., Zhang, X., Yang, L., Wang, R., and Huang, X. (2022). Transformer-based architecture for classification of cervical cell images. *Computer Methods and Programs in Biomedicine*, 220:106828.

Liang, J., Wang, D., and Ling, X. (2021). Image classification for soybean and weeds based on vit. In *MMSE 2021 - Journal of Physics: Conference Series*, volume 2002, page 012068, Shanghai, China. IOP Publishing. https://doi.org/10.1088/1742-6596/2002/1/012068.

Liu, B., Tang, R., Chen, Y., Yu, J., Guo, H., and Zhang, Y. (2019). Feature generation by convolutional neural network for click-through rate prediction. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*, pages 1119–1129, New York, NY, USA. Association for Computing Machinery.

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.

Manickadevi, M. S., Hemalatha, S. V., and Thangamani, M. (2022). Comparative study of the cytologic diagnosis, specimen adequacy, sensitivity, and cost effectiveness of liquid-based cytology with that of conventional pap tests. *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, 11(2):474–478.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

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.

Nayar, R. and Wilbur, D. (2015). *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. Springer, New York.

Pacal, I. (2024). Maxcervixt: A novel lightweight vision transformer-based approach for precise cervical cancer detection. *Knowledge-Based Systems*, 289:111482.

Phung, V. and Rhee, E. (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9:4500.

Plissiti, M. E., Nikou, C., and Charchanti, A. (2011). Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):233–241.

Robb, B. W. and Mutch, M. G. (2006). Epidermoid carcinoma of the anal canal. *Clinics in Colon and Rectal Surgery*, 19(2):108–115.

Ryan, D. P., Compton, C. C., and Mayer, R. J. (2000). Carcinoma of the anal canal. *The New England Journal of Medicine*, 342(11):792–800.

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

Souza, E. P. P., Mustafa, M. M., and Sena, A. B. (2021). Vantagens e desvantagens da citologia convencional e da citologia em meio líquido na prática clínica: uma revisão integrativa. *Research, Society and Development*, 10(14):e245101422350.

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.

Tanum, G., Tveit, K. M., and Karlsen, K. O. (1991). Diagnosis of anal carcinoma – doctor's finger still the best. *Oncology*, 48(5):383–386.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc.

Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., and Chang, E. I.-C. (2020). Deep learning of feature representation with multiple instance learning for medical image analysis. *Neurocomputing*, 188:68–76.

Zhang, L., Lu, L., Nogues, I., Summers, R. M., Liu, S., and Yao, J. (2017). Deeppap: deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1633–1643.