Classificação de Células Cervicais Cancerígenas com CNNs e ViTs em Ensemble para Auxílio ao Diagnóstico Médico

Marcelo Victor Lima[1], **Maria Helena Mesquita Britto**[2], **Laurindo Britto Neto**[1]

[1]Departamento de Computação – Universidade Federal do Piauí (UFPI),
CEP 64.049-550 – Teresina – PI – Brasil

[2]Centro Universitário Santo Agostinho, CEP 64.019-625 -Teresina - PI - Brasil

{marcelo.victor, laurindoneto}@ufpi.edu.br, mhrmesquita@hotmail.com

Abstract. *Cervical cancer is a global public health issue that requires effective screening methods. This work proposes an ensemble of EfficientViT, EVA-02, and EdgeNeXt for the automatic classification of cervical cells. The Herlev (917 images) and SIPaKMeD (4049 images) datasets were evaluated for binary and multiclass classification tasks. The methodology includes transfer learning, data augmentation, and 5-fold cross-validation. In the Herlev dataset, the results were 98.35% accuracy (binary) and 83.40% (7 classes). In SIPaKMeD, the model achieved 99.73% (binary), 98.96% (3 classes), and 98.08% (5 classes), reaching state-of-the-art performance. The results demonstrate the model's potential to assist in medical diagnosis, reducing the need for manual analysis.*

Resumo. *O câncer cervical é um problema global de saúde pública que exige métodos eficazes de triagem. Este trabalho propõe um ensemble de EfficientViT, EVA-02 e EdgeNeXt para classificação automática de células cervicais. Foram avaliados os conjuntos Herlev (917 imagens) e SIPaKMeD (4.049 imagens) em tarefas de classificação binária e multiclasse. A metodologia inclui transferência de aprendizado, aumento de dados e validação cruzada (5 folds). No conjunto Herlev, os resultados foram 98,35% de acurácia (binário) e 83,40% (7 classes). No SIPaKMeD, obteve-se 99,73% (binário), 98,96% (3 classes) e 98,08% (5 classes), atingindo o estado da arte. Os resultados demonstram o potencial do modelo para auxílio ao diagnóstico médico, reduzindo a necessidade de análise manual.*

1. Introdução

O câncer cervical é o quarto tipo de câncer mais prevalente entre mulheres [Ferlay et al. 2024a]. De acordo com a Organização Mundial da Saúde (OMS), aproximadamente 660 mil novos casos foram diagnosticados globalmente em 2022, resultando em cerca de 350 mil óbitos no mesmo período [Ferlay et al. 2024a]. A alta taxa de mortalidade está diretamente relacionada à falta de programas de triagem eficazes, especialmente em países de baixa e média renda, onde ocorrem cerca de 94% dos óbitos [Jiang et al. 2023]. No Brasil, a OMS estimou 18.700 novos casos em 2022, com previsão de aumento para 19.700 em 2025 [Ferlay et al. 2024b].

Essa patologia está fortemente associada ao Papilomavírus Humano (*Human Papillomavirus* – HPV) [Bedell et al. 2020], especialmente aos tipos de vírus de alto

risco, como o HPV-16 e o HPV-18, que provocam alterações significativas nos núcleos das células cervicais. Segundo as diretrizes mais recentes da OMS [Hou et al. 2022], recomenda-se a utilização de três métodos de triagem para a detecção precoce do câncer do colo do útero: o teste de HPV, a citologia (incluindo tanto o exame tradicional de Papanicolau quanto a citologia em base líquida) e a inspeção visual com ácido acético.

A técnica convencional de Papanicolaou constitui o método mais amplamente empregado para a detecção de anomalias cervicais em âmbito global. O procedimento envolve a coleta de células do colo do útero por meio de raspagem, seguida de análise microscópica, conforme ilustrado na Figura 1.

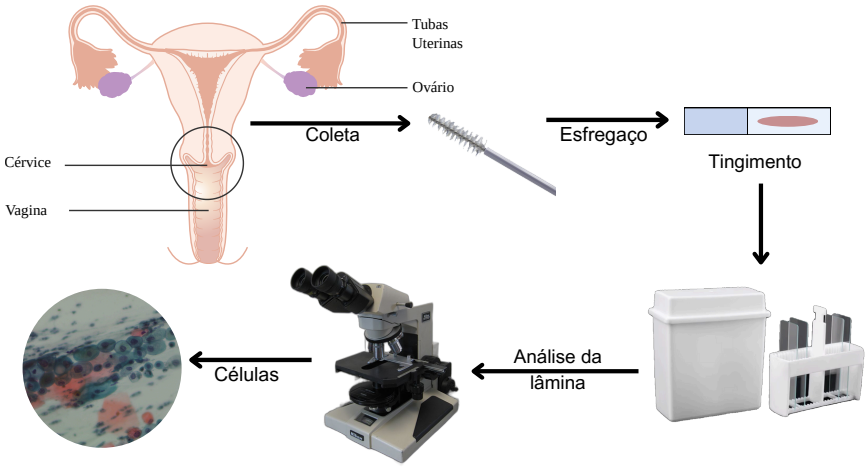

Figura 1. Ilustração do exame de citologia utilizando a técnica do Papanicolau convencional.

Para a realização do exame, utiliza-se um espéculo vaginal, que permite a introdução de uma escova macia na cavidade vaginal, viabilizando a obtenção das células cervicais. Em seguida, as células coletadas são distribuídas de forma uniforme sobre uma lâmina de vidro, configurando o preparo citológico. Posteriormente, após a aplicação de reagentes corantes, os citopatologistas procedem à análise da amostra por meio de microscopia, o que possibilita a formulação do diagnóstico. Em decorrência de sangue, muco, inflamação e outros fatores, o exame convencional de Papanicolaou frequentemente resulta em amostras com baixa qualidade, apresentando imagens borradas e aumentando a probabilidade de erros na detecção [Jiang et al. 2023].

Na citologia em meio líquido, após a coleta das células do colo do útero, as amostras são transferidas para um meio líquido especial, em vez de serem disseminadas diretamente sobre uma lâmina de vidro, como ocorre no método convencional. A amostra líquida possibilita a formação de uma camada uniforme de células, aprimorando a qualidade da amostra e minimizando os artefatos decorrentes do processo de preparação.

Um dos desafios da triagem convencional é que a análise é feita manualmente por especialistas, tornando o processo dispendioso, lento e suscetível de erro. A identificação de alterações nas células cervicais é essencial para prevenir a progressão de células anormais em tumores malignos. Assim, um sistema automatizado e eficaz para auxiliar o diagnóstico médico, identificando precocemente essa patologia, poderia direcionar intervenções terapêuticas específicas, contribuindo para a redução da taxa de mortalidade.

Uma das primeiras tentativas de análise automatizada de células anormais foi o *Cytoanalyzer*, desenvolvido em 1956 [Tolles and Bostrom 1956]. No entanto, limitações tecnológicas e erros frequentes impediram sua adoção. Desde então, diversos sistemas automatizados foram criados, incluindo o *Genius Digital Diagnostics System* [Ikenberg et al. 2023], aprovado pela *Food and Drug Administration* (FDA).

Atualmente, com o avanço da inteligência artificial (IA) e o grande poder computacional para processamento de dados, são utilizadas abordagens baseadas em visão computacional e aprendizado de máquina (*Machine Learning* – ML) para o reconhecimento e classificação de células cervicais [Fang et al. 2024a].

Este trabalho propõe um modelo de *ensemble learning* para a classificação de células cervicais, combinando *EfficientViT*, *EVA-02* e *EdgeNeXt*, pré-treinados no ImageNet [Russakovsky et al. 2015]. Utilizou-se transferência de aprendizado e aumento de dados para mitigar desbalanceamentos e melhorar a generalização. A metodologia foi validada nos conjuntos Herlev [Jantzen et al. 2005] e SIPaKMeD [Plissiti et al. 2018], em tarefas de classificação binária e multiclasse. Após o *fine-tuning*, avaliou-se o desempenho individual dos modelos e do *ensemble*.

Este artigo está estruturado da seguinte forma: a Seção 2 aborda os trabalhos relacionados, a Seção 3 descreve os conjuntos de dados e métodos utilizados, e a Seção 4 detalha a abordagem proposta. Os resultados são apresentados na Seção 5 e analisados na Seção 6. Por fim, a Seção 7 apresenta as conclusões do estudo e sugere trabalhos futuros.

2. Trabalhos Relacionados

Diversas revisões de literatura recentes analisaram a aplicação de inteligência artificial na triagem e no diagnóstico do câncer cervical. Jiang et al. [Jiang et al. 2023] revisaram estudos de 2016 a 2022, destacando abordagens como *transfer learning*, *ensemble learning* e fusão híbrida de características, além de desafios como o desbalanceamento dos conjuntos de dados. Fang et al. [Fang et al. 2024a] analisaram artigos de 2017 a 2023 sobre segmentação e classificação de células cervicais, abordando o uso de CNNs e *Vision Transformers* (ViTs). Khare et al. [Khare et al. 2024] revisaram estudos entre 2013 e 2023, explorando IA aplicada a múltiplas modalidades de imagem e dados clínicos, além de sugerir aprendizado federado e meta-aprendizado. Wubineh et al. [Wubineh et al. 2024] revisaram pesquisas de 2016 a 2023 sobre segmentação e classificação de esfregaços de Papanicolaou, enquanto Vargas-Cardona et al. [Vargas-Cardona et al. 2024] analisaram estudos de 2009 a 2022 sobre IA no rastreamento do câncer cervical. Mathew e Cheriyan [Mathew and Cheriyan 2024] revisaram estudos de aprendizado de máquina para detecção e análise de fatores de risco do câncer cervical. Por fim, Wu et al. [Wu et al. 2024] discutiram avanços recentes no uso de IA para rastreamento da doença em países com infraestrutura de saúde limitada.

Além das revisões de literatura, três estudos recentes merecem destaque por suas contribuições significativas para a classificação de células cervicais. O estudo de S. et al. [S. et al. 2024] realizaram uma comparação entre modelos baseados em *ensemble* — especificamente, *AdaBoost*, *XGBoost*, *CatBoost* e *LightGBM* — para a classificação de células utilizando o conjunto de dados *SIPaKMeD*. Os autores relataram resultados expressivos, destacando, em particular, o desempenho do *XGBoost*, que atingiu 99,7% de acurácia, 96,4% de precisão, 97,5% de *recall* e 96,0% de *F1-score*.

Outra abordagem promissora apresentada na literatura é a *Swin-GA-RF*, proposta por Alohali et al. [Alohali et al. 2024]. Essa metodologia integra o *Swin Transformer* para a extração de características, algoritmos genéticos (*genetic algorithm – GA*) para a seleção de atributos e, como classificador, o *Random Forest* (RF). Os autores aplicaram o conjunto de dados *SIPaKMeD* em experimentos que englobaram tanto a classificação binária (células normais e anormais) quanto a classificação em cinco classes. Ao comparar os resultados com os obtidos por redes CNN pré-treinadas e pelo próprio *Swin Transformer*, verificou-se que o emprego do otimizador Adam permitiu alcançar resultados superiores na classificação binária, atingindo acurácia de 99,012%, precisão de 99,015%, *recall* de 99,012% e *F1-score* de 99,011%. Na classificação em cinco classes, os resultados foram igualmente notáveis, apresentando acurácia de 98,808%, precisão de 98,812%, *recall* de 98,808% e *F1-score* de 98,808%.

Adicionalmente, o trabalho de Sholik et al. [Sholik et al. 2024] propõe a extração de características por meio de modelos pré-treinados das arquiteturas CNN e ViT, combinando-os para a obtenção de representações globais e locais. Em seguida, empregam-se técnicas de redução de dimensionalidade, tais como PCA e LDA para a extração de características discriminantes, seguidas da aplicação de classificadores, dentre os quais se destacam SVM, K-NN, MLP e regressão logística (*Logistic Regression* – LR). As redes pré-treinadas utilizadas nesse estudo incluem ResNet-50, VGG-16, DenseNet-121, Inception-V3, ViT-B16, ViT-B32, ViT-L16 e ViT-L32.

Essas abordagens foram avaliadas nos conjuntos *Herlev*, *SIPaKMeD* e *Mendeley LBC*. No *Herlev*, a classificação binária alcançou 97,83% de acurácia em todos os classificadores analisados. No *Mendeley LBC*, SVM, K-NN, MLP e LR atingiram 100%. Já no *SIPaKMeD*, para três classes (normais, anormais e benignas), SVM, K-NN e LR obtiveram 98,52% de acurácia.

As abordagens mencionadas destacam os avanços no uso de redes neurais e modelos híbridos na classificação de células cervicais, mas desafios persistem na diferenciação de subtipos celulares e no desbalanceamento dos dados. Este estudo visa aprimorar esses aspectos com um *ensemble* de *CNNs* e *ViTs*, fortalecendo a robustez da classificação e o desempenho no diagnóstico automatizado.

3. Materiais e Métodos

Esta seção apresenta os conjuntos de dados utilizados e os métodos de pré-processamento aplicados para o aumento de dados.

3.1. Conjuntos de Dados

Neste trabalho, foram utilizados os conjuntos de dados *Herlev* [Jantzen et al. 2005] e *SIPaKMeD* [Plissiti et al. 2018], ambos disponíveis publicamente e descritos a seguir.

3.1.1. Herlev

Em 2005, a Universidade do Hospital *Herlev* [Jantzen et al. 2005], na Dinamarca, disponibilizou um conjunto público com 917 imagens de esfregaços cervicais, divididas em sete classes (três normais e quatro anormais), conforme a Tabela 1. As imagens possuem resolução de 0,201 μm por pixel, como ilustrado nas Figuras 2 e 3.

Tabela 1. Classificação do conjunto de dados *Herlev* com 917 células (242 normais e 675 anormais).

Categoria	Tipo de célula	Quantidade
Normal	Superficial squamous epithelial	74
Normal	Intermediate squamous epithelial	70
Normal	Columnar epithelial	98
Anormal	Mild squamous non-keratinizing dysplasia	182
Anormal	Moderate squamous non-keratinizing dysplasia	146
Anormal	Severe squamous non-keratinizing dysplasia	197
Anormal	Squamous cell carcinoma in situ intermediate	150

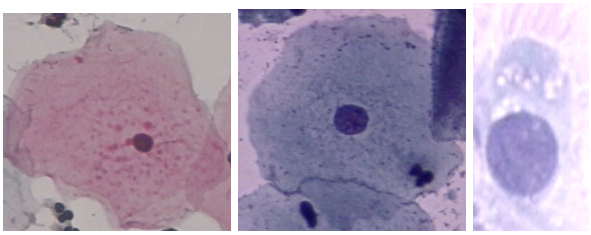

Figura 2. Células normais de cada classe da Tabela 1 [Jantzen et al. 2005].

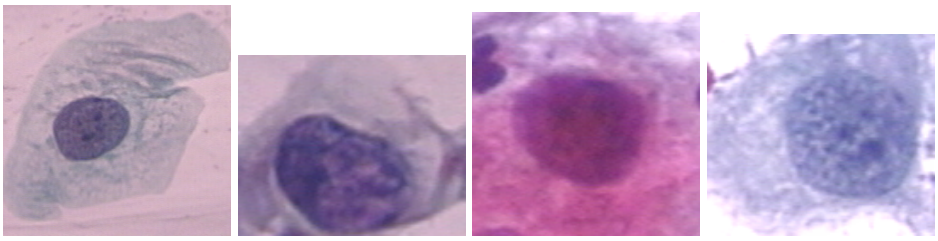

Figura 3. Células anormais de cada classe da Tabela 1 [Jantzen et al. 2005].

3.1.2. SIPaKMeD

O conjunto de dados SIPaKMeD [Plissiti et al. 2018] é composto por 4.049 imagens de células isoladas, cuidadosamente recortadas a partir de 966 imagens de *clusters* celulares provenientes de lâminas de exame de Papanicolau. As imagens foram obtidas por meio da câmera digital CCD Infinity 1 (Lumenera) ao microscópio óptico (OLYMPUS BX53F). A classificação das células foi realizada por citopatologistas especialistas, que categorizaram as amostras em cinco classes diferentes baseadas nas características morfológicas. As categorias, os tipos de células e suas respectivas distribuições são descritas pela Tabela 2 e ilustradas pela Figura 4.

Tabela 2. Conjunto de imagens SIPaKMeD com suas respectivas categorias.

Categoria	Tipo de célula	Número de Imagens	Número de Células
Normal	Superficial/Intermediate	126	831
Normal	Parabasal	108	787
Anormal	Koilocytotic	238	825
Anormal	Dyskeratotic	223	813
Benígno	Metaplastic	271	793
Total:		**966**	**4.049**

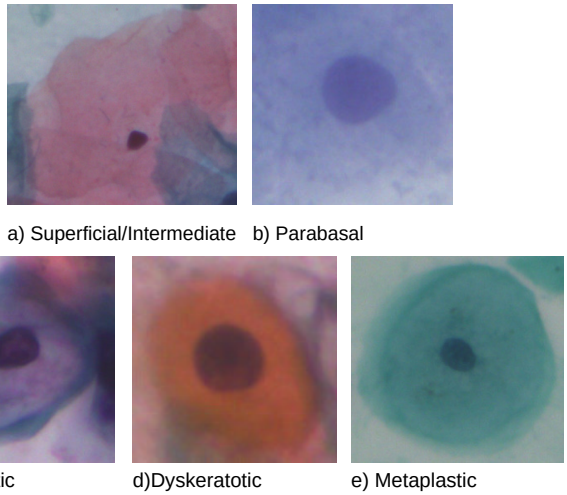

Figura 4. Células do conjunto SIPaKMeD. As figuras 'a' e 'b' representam células normais. Enquanto as figuras 'c', 'd' e 'e' representam células anormais [Plissiti et al. 2018].

3.2. Aumento de Dados

Como forma de diminuir o desbalanceamento entre as classes e para mitigar o *overfitting* dos modelos, aplicou-se uma estratégia de aumento de dados. Foram utilizados os seguintes procedimentos na fase de pré-processamento:

- **Para as imagens aumentadas:** *flip* horizontal e vertical; rotação aleatória; *CLAHE*; filtro gaussiano; aumento de nitidez; ajuste de brilho, contraste, saturação e matiz.
- **Para as demais imagens:** *flip* horizontal e vertical; rotação aleatória; ajuste de brilho, contraste, saturação, matiz e normalização.

A Tabela 3 resume a quantidade de imagens geradas (Aumento de Dados) e o total de imagens do conjunto de dados após o aumento de dados para cada tipo de experimento.

Tabela 3. Número de imagens geradas no aumento de dados para cada cenário.

Conjunto de Dados	Configuração	Aumento de Dados	Total de Imagens
Herlev	7 classes	462	1.379
	2 classes	433	1.350
SIPaKMeD	5 classes	106	4.155
	3 classes	865	4.914
	2 classes	813	4.862

4. Abordagem Proposta

Um dos critérios para a escolha dos modelos na abordagem proposta foi a relação entre a quantidade de parâmetros, o número de operações (*Multiply-Accumulate Operations* – MACs) e a acurácia obtida no conjunto ImageNet. Optou-se por modelos com custo computacional reduzido e acurácia superior a 80%. O *EfficientViT* possui alta eficiência computacional (1,6 GMACs); o EVA-02 alcançou 90% de acurácia no ImageNet; e o *EdgeNeXt* é capaz de capturar padrões multiescala por meio do mecanismo STDA.

O *EfficientViT* [Cai et al. 2022] é um modelo baseado no *Vision Transformer* (ViT) [Dosovitskiy et al. 2020] que inova ao empregar atenção linear em múltiplas es-

calas. Essa abordagem permite alcançar um desempenho satisfatório na classificação de imagens de alta resolução com um modelo de baixa complexidade.

De forma complementar, o *EVA-02* [Fang et al. 2024b] apresenta performance excepcional na classificação do ImageNet-1K, obtendo 90% de acurácia com 304 milhões de parâmetros. Esse desempenho é alcançado por meio de uma arquitetura *Transformer* otimizada, associada a um extenso pré-treinamento utilizando um codificador de visão *CLIP* [Radford et al. 2021] de grande escala, o qual é de código aberto e acessível.

Além disso, o *EdgeNeXt* [Maaz et al. 2022] combina redes neurais convolucionais com ViTs, introduzindo o codificador de atenção transposta separável em profundidade (STDA). Este codificador segmenta os tensores de entrada em múltiplos grupos de canais e aplica convolução separável em profundidade em conjunto com mecanismos de *self-attention* ao longo das dimensões de canal, permitindo a ampliação implícita do campo receptivo e a codificação eficiente de características em múltiplas escalas.

A abordagem proposta neste trabalho fundamenta-se na utilização combinada desses três modelos pré-treinados, previamente adaptados aos conjuntos de imagens citológicas. A Tabela 4 detalha os hiperparâmetros utilizados no treinamento dos modelos. A decisão final do *ensemble* é baseada na média ponderada das probabilidades de cada modelo, em que os pesos são determinados pela acurácia de validação de cada arquitetura.

Para mitigar o sobreajuste, implementou-se uma estratégia de *early stopping* com um

Tabela 4. Hiperparâmetros utilizados para o treinamento e parâmetros dos modelos utilizados para o treinamento.

Parâmetros e hiperparâmetros	EfficientViT_b2	EdgeNeXt_base	EVA-02_small
Tamanho do lote	32	32	32
Otimizador	AdamW	AdamW	AdamW
Taxa de aprendizado inicial	0.0001	0.0001	0.0001
Épocas máximas	50	50	50
Épocas mínimas	20	20	20
Dimensões da imagem de entrada	(336x336x3)	(336x336x3)	(336x336x3)
Quantidade de parâmetros do modelo	21,8 M	17,9 M	21,7 M
GMACs	1,6	3,8	15,5

patience de 10 épocas, adotando como critério de parada uma variação mínima de $1 \cdot 10^{-5}$ na função *loss* de validação. Posteriormente, os modelos foram combinados em um esquema de *ensemble* para a classificação do conjunto de testes, permitindo a comparação entre os resultados individuais e os obtidos com a combinação dos modelos.

Nos experimentos, foram realizadas quatro rodadas de validação cruzada 5-fold (totalizando 20 execuções) para cada modelo, em ambos os conjuntos de dados, contemplando tarefas de classificação binária e multiclasse. Essa técnica foi adotada para garantir robustez estatística, especialmente em conjuntos pequenos como o Herlev (917 imagens). Em cada rodada, os dados foram redivididos aleatoriamente, assegurando variação nas imagens selecionadas para os *folds*, sendo quatro utilizados para treinamento e um para teste. Ao final de cada execução, computaram-se as métricas de Acurácia, Precisão, *Recall* e *F1-score*. Nos conjuntos de treinamento, 90% dos exemplos foram utilizados para

o treinamento efetivo e os 10% restantes foram reservados para validação durante o processo de ajuste dos modelos.

Para verificar diferenças estatísticas entre os modelos, utilizou-se o teste *Wilcoxon Signed-Rank* (WSR), apropriado para dados pareados e que não exigem distribuição normal. Segundo Bridge e Sawilowsky [Bridge and Sawilowsky 1999], testes não paramétricos como os de *Wilcoxon* apresentam vantagens em relação ao teste *t* de *Student*, especialmente em contextos com dados assimétricos, caudas pesadas ou amostras pequenas, oferecendo maior poder estatístico nessas situações.

5. Resultados

Nesta seção são apresentados os resultados obtidos a partir dos experimentos realizados nos conjuntos de dados *Herlev* e *SIPaKMeD*. Foram avaliadas diferentes configurações de classificação: para o conjunto Herlev, os experimentos consideraram 7 classes e 2 classes; para o conjunto *SIPaKMeD*, foram testadas configurações de 5, 3 e 2 classes.

5.1. Resultados no Conjunto de Dados *Herlev*

Os experimentos realizados com o conjunto *Herlev* foram conduzidos em duas configurações: classificando as imagens em 7 classes e em 2 classes. As Tabelas 5 e 6 apresentam os resultados dos experimentos para a abordagem proposta e os demais modelos utilizados como pilares para ela, destacando os melhores resultados.

Tabela 5. Desempenho no conjunto *Herlev* para 2 classes

Métrica	Abordagem Proposta	*EfficientViT_b2*	*EVA02_small*	*EdgeNeXt_base*
Acurácia	**98,35±0,61%**	97,41±0,74%	96,71± 1,07%	97,91±0,66%
Precisão	**98,37±0,61%**	97,42±0,73%	96,76±1,05%	97,94±0,64%
Recall	**98,35±0,61%**	97,41±0,74%	96,71±1,07%	97,91±0,66%
F1-Score	**98,35±0,61%**	97,41±0,74%	96,71±1,07%	97,91±0,66%

Tabela 6. Desempenho no conjunto *Herlev* para 7 classes

Métrica	Abordagem Proposta	*EfficientViT_b2*	*EVA02_small*	*EdgeNeXt_base*
Acurácia	**83,40±1,59%**	80,90±1,76%	78,31±2,29%	80,42±2,05%
Precisão	**83,59±1,60%**	81,17±1,90%	78,94±2,27%	80,54±2,03%
Recall	**83,42±1,58%**	80,91±1,76%	78,33±2,31%	80,43±2,04%
F1-Score	**83,34±1,62%**	80,80±1,87%	78,24±2,21%	80,21±2,07%

5.2. Resultados no Conjunto de Dados *SIPaKMeD*

Para o conjunto *SIPaKMeD*, os experimentos foram conduzidos com três configurações: 5 classes, 3 classes (células normais, anormais e benignas) e 2 classes. As mesmas métricas de desempenho aplicadas ao conjunto *Herlev* foram computadas no SIPaKMeD. As Tabelas 7, 8 e 9 apresentam os resultados obtidos, destacando os melhores resultados.

Tabela 7. Desempenho no conjunto *SIPaKMeD* para 2 classes.

Métrica	Abordagem Proposta	*EfficientViT_b2*	*EVA02_small*	*EdgeNeXt_base*
Acurácia	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%
Precisão	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%
Recall	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%
F1-Score	**99,73 ± 0,13%**	99,57 ± 0,26%	99,40 ± 0,25%	99,54 ± 0,19%

Tabela 8. Desempenho no conjunto *SIPaKMeD* para 3 classes.

Métrica	Abordagem Proposta	*EfficientViT_b2*	*EVA02_small*	*EdgeNeXt_base*
Acurácia	**98,96 ± 0,19%**	98,67 ± 0,32%	98,20 ± 0,39%	98,45 ± 0,33%
Precisão	**98,97 ± 0,19%**	98,68 ± 0,31%	98,20 ± 0,39%	98,46 ± 0,32%
Recall	**98,96 ± 0,19%**	98,67 ± 0,32%	98,20 ± 0,39%	98,45 ± 0,33%
F1-Score	**98,96 ± 0,19%**	98,67 ± 0,32%	98,20 ± 0,39%	98,45 ± 0,33%

Tabela 9. Desempenho no conjunto *SIPaKMeD* para 5 classes.

Métrica	Abordagem Proposta	*EfficientViT_b2*	*EVA02_small*	*EdgeNeXt_base*
Acurácia	**98,08 ± 0,54%**	97,36 ± 0,72%	96,84 ± 0,57%	97,15 ± 0,44%
Precisão	**98,09 ± 0,54%**	97,37 ± 0,73%	96,87 ± 0,57%	97,17 ± 0,44%
Recall	**98,08 ± 0,54%**	97,36 ± 0,72%	96,84 ± 0,57%	97,15 ± 0,45%
F1-Score	**98,08 ± 0,54%**	97,35 ± 0,73%	96,84 ± 0,58%	97,15 ± 0,44%

6. Discussão

Os resultados mostram que a complexidade da tarefa de classificação tem impacto direto nas métricas de desempenho. Embora a redução do número de classes tenda a melhorar as métricas, ela pode comprometer a distinção entre subtipos celulares devido à perda de informações relevantes, limitando a aplicabilidade clínica. A Tabela 10 resume os melhores resultados da proposta em comparação com estudos da literatura.

Tabela 10. Comparação da abordagem proposta com os principais estudos da literatura em cada cenário do experimento realizado

Conjunto de Dados	Modelo	Acurácia	Precisão	*Recall*	*F1-score*	Divisão de Dados
Herlev (2 classes)	Abordagem proposta	**98,35**	98,37	**98,35**	**98,38**	Validação Cruzada *5-fold*
	Sholik et al. [Sholik et al. 2024]	97,83	**98,75**	92,86	95,52	*Holdout*
Herlev (7 classes)	Abordagem proposta	**83,40**	**83,59**	**83,42**	**83,34**	Validação Cruzada *5-fold*
SIPaKMeD (2 classes)	Abordagem proposta	**99,73**	**99,73**	**99,73**	**99,73**	Validação Cruzada *5-fold*
	Swin-GA-RF [Alohali et al. 2024]	99,01	99,01	99,01	99,01	*Holdout*
SIPaKMeD (3 classes)	Abordagem proposta	**98,96**	**98,97**	**98,96**	**98,96**	Validação Cruzada *5-fold*
	Sholik et al. [Sholik et al. 2024]	98,52	98,81	97,78	98,28	*Holdout*
SIPaKMeD (5 classes)	Abordagem proposta	98,08	98,09	98,08	98,08	Validação Cruzada *5-fold*
	Swin-GA-RF [Alohali et al. 2024]	98,81	**98,81**	**98,81**	**98,81**	*Holdout*
	XGBoost [S. et al. 2024]	**99,70**	96,40	97,50	96,00	Validação Cruzada *K-fold*

Uma diferença relevante entre os estudos comparados está no método de validação adotado. A abordagem proposta utilizou validação cruzada 5-fold, garantindo maior robustez e reduzindo o impacto da divisão inicial dos dados. Já os estudos de Sholik et al.[Sholik et al. 2024] e *Swin-GA-RF*[Alohali et al. 2024] adotaram *holdout*, mais sensível a essa divisão, o que torna a proposta mais confiável para aplicações reais.

Para o conjunto *Herlev* em duas classes (Tabelas 5 e 10), a abordagem proposta superou Sholik et al. [Sholik et al. 2024], com acurácia média de 98,35% contra 97,83%, resultado atribuído à robustez do *ensemble*, que permitiu capturar representações mais robustas das células anormais e normais.

Para o conjunto *Herlev* em sete classes (Tabelas 6 e 10) a ab proposta obteve os melhores resultados com acurácia média de 83,40% e com uma maior diferença entre os outros modelos. A diferença de desempenho em relação à configuração binária destaca a complexidade adicional envolvida na diferenciação entre múltiplos tipos celulares.

Em relação ao conjunto *SIPaKMeD* binário (Tabelas 7 e 10), a abordagem proposta atingiu o estado da arte com média de 99,73% em todas as métricas. Em

comparação, o modelo *Swin-GA-RF* [Alohali et al. 2024] obteve 99,01% em todas as métricas, sendo superado. Essa superioridade pode estar associada ao uso de modelos com mecanismos distintos de aprendizado, combinando características extraídas por CNNs e ViTs, o que favorece uma maior generalização dos padrões celulares.

Para a classificação de três classes do conjunto *SIPaKMeD* (Tabelas 8 e 10) a proposta apresentou os melhores resultados, atingindo 98,96% em todas as métricas, superando Sholik et al. [Sholik et al. 2024], cujo melhor resultado de acurácia foi de 98,52%. Isso sugere que a fusão de diferentes tipos de extração de características realizada no *ensemble* pode ser mais eficaz na distinção entre células normais, anormais e benignas.

Os experimentos no conjunto *SIPaKMeD* com 5 classes (Tabelas 9 e 10) mostraram que a proposta alcançou 98,08% em todas as métricas. O modelo *Swin-GA-RF* [Alohali et al. 2024] obteve métricas ligeiramente superiores (98,81%), mas com uma diferença inferior a 0,75%. Entretanto, esse resultado foi obtido por meio do *holdout*, que pode gerar avaliações mais otimistas ou sensíveis à divisão dos dados. Já o *XGBoost* [S. et al. 2024] apresentou a maior acurácia entre os estudos (99,70%), porém com os piores resultados nas outras métricas, indicando menor capacidade de generalização. Dessa forma, a proposta demonstrou maior equilíbrio em todas as métricas, sendo mais confiável para aplicações médicas, nas quais a estabilidade dos resultados é fundamental.

Do ponto de vista clínico, a métrica mais crítica para a triagem automatizada do câncer cervical é o *Recall*, uma vez que falsos negativos podem resultar em diagnósticos tardios, aumentando a taxa de mortalidade. Nos experimentos realizados, a abordagem proposta obteve um *Recall* superior aos modelos comparados na maioria dos cenários, reduzindo significativamente a probabilidade de falha na detecção de células anormais. Além disso, a estabilidade proporcionada pela validação cruzada 5-fold reforça a robustez do modelo, sugerindo um desempenho mais confiável em ambientes clínicos reais. A elevada acurácia alcançada na classificação binária evidencia o potencial da abordagem proposta para a triagem automatizada de exames citológicos, minimizando a necessidade de revisão manual por especialistas. No entanto, o desempenho inferior em tarefas com um maior número de classes indica desafios persistentes na diferenciação de subtipos celulares, o que ressalta a importância de futuras melhorias nos modelos para assegurar um diagnóstico mais preciso e detalhado, especialmente em contextos clínicos complexos.

Por fim, a Tabela 11 apresenta os p-valores obtidos nos testes de hipóteses WSR, comparando a abordagem proposta com os melhores modelos de referência em cada cenário experimental. Em todos os casos, observaram-se diferenças estatisticamente significativas com nível de confiança de 99% em todas as métricas computadas, o que evidencia a superioridade da abordagem proposta em relação aos demais métodos analisados.

Tabela 11. P-valores obtidos pelo WSR, com nível de confiança de 99%, comparando a abordagem proposta com os melhores modelos em cada cenário.

Conjunto	Comparação	Acurácia	Precisão	*Recall*	*F1-score*
Herlev (2 classes)	*EdgeNeXt_base*	$9{,}92 \cdot 10^{-4}$	$3{,}29 \cdot 10^{-4}$	$9,92 \cdot 10^{-4}$	$6,53 \cdot 10^{-4}$
Herlev (7 classes)	*EfficientViT_b2*	$2{,}39 \cdot 10^{-4}$	$1,40 \cdot 10^{-4}$	$6,4 \cdot 10^{-5}$	$7,5 \cdot 10^{-5}$
SIPaKMeD (2 classes)	*EfficientViT_b2*	$2,848 \cdot 10^{-3}$	$4,003 \cdot 10^{-3}$	$2,482 \cdot 10^{-3}$	$3,495 \cdot 10^{-3}$
SIPaKMeD (3 classes)	*EfficientViT_b2*	$9{,}28 \cdot 10^{-4}$	$1,116 \cdot 10^{-3}$	$2,136 \cdot 10^{-3}$	$1,477 \cdot 10^{-3}$
SIPaKMeD (5 classes)	*EfficientViT_b2*	$1{,}31 \cdot 10^{-4}$	$2{,}0 \cdot 10^{-6}$	$4{,}0 \cdot 10^{-6}$	$2{,}0 \cdot 10^{-6}$

7. Conclusões

O *ensemble* proposto mostrou-se eficaz na análise automatizada de imagens citológicas, especialmente no enfrentamento da heterogeneidade dos dados e das variações na qualidade das imagens. Os experimentos indicaram que a combinação de modelos pré-treinados eleva a acurácia geral, superando abordagens individuais, com destaque para a classificação binária, em que a proposta atingiu o estado da arte.

Adicionalmente, a maior discrepância no desempenho entre tarefas binárias e multiclasse no conjunto *Herlev* indica a complexidade da diferenciação entre subtipos celulares, sugerindo que métodos híbridos e técnicas avançadas de normalização e pré-processamento podem ajudar a superar esse desafio.

Para trabalhos futuros, recomenda-se explorar técnicas como aprendizado federado para aumentar a privacidade dos dados clínicos, bem como o uso de métodos de interpretabilidade para tornar os modelos mais transparentes e confiáveis para aplicações médicas, aumentando a confiança dos profissionais da saúde na adoção desses modelos. Além disso, a validação em conjuntos de dados mais diversificados e cenários clínicos reais será essencial para confirmar a robustez e aplicabilidade da abordagem.

Referências

Alohali, M. A., El-Rashidy, N., Alaklabi, S., Elmannai, H., Alharbi, S., and Saleh, H. (2024). Swin-GA-RF: genetic algorithm-based Swin Transformer and random forest for enhancing cervical cancer classification. *Frontiers in Oncology*, 14.

Bedell, S. L., Goldstein, L. S., Goldstein, A. R., and Goldstein, A. T. (2020). Cervical cancer screening: Past, present, and future. *Sexual Medicine Reviews*, 8(1):28–37.

Bridge, P. and Sawilowsky, S. (1999). Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research. *J CLIN EPIDEMIOL*.

Cai, H., Li, J., Hu, M., Gan, C., and Han, S. (2022). Efficientvit: Multi-scale linear attention for high-resolution dense prediction. *arXiv preprint arXiv:2205.14756*.

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

Fang, M., Liao, B., Lei, X., and Wu, F.-X. (2024a). A systematic review on deep learning based methods for cervical cell image analysis. *Neurocomputing*, 610:128630.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., and Cao, Y. (2024b). Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171.

Ferlay, J., Colombet, M., Soerjomataram, I., and D. M., P. (2024a). Global cancer observatory: Cancer today. Acessado em 17 de abril de 2024.

Ferlay, J., Laversanne, M., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., and Bray, F. (2024b). Global cancer observatory: Cancer tomorrow (version 1.1). Acessado em 27 de dezembro de 2024.

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.

Ikenberg, H., Lieder, S., Ahr, A., Wilhelm, M., Schön, C., and Xhaja, A. (2023). Comparison of the hologic genius digital diagnostics system with the thinprep imaging system—a retrospective assessment. *Cancer Cytopathology*, 131(7):424–432.

Jantzen, J., Norup, J., Dounias, G., and Bjerregaard, B. (2005). Pap-smear benchmark data for pattern classification. *NiSIS*, pages 1–9.

Jiang, P., Li, X., Shen, H., Chen, Y., Wang, L., Chen, H., Feng, J., and Liu, J. (2023). A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. *Artif. Intell. Rev.*, 56(Suppl 2):2687–2758.

Khare, S. K., Blanes-Vidal, V., Booth, B. B., Petersen, L. K., and Nadimi, E. S. (2024). A systematic review and research recommendations on artificial intelligence for automated cervical cancer detection. *WIREs Data Min. Knowl. Discov.*, 14(6):e1550.

Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S. W., Anwer, R. M., and Khan, F. S. (2022). Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *CADL*. Springer.

Mathew, B. and Cheriyan, R. (2024). A detailed review on classification and risk factor analysis of cervical cancer using artificial intelligence. In *IATMSI*, pages 1–6. IEEE.

Plissiti, M. E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., and Charchanti, A. (2018). Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *ICIP*, pages 3144–3148. IEEE.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252.

S., V., N., M. D., G., M., D., V. K., C., S., and C., A. M. (2024). Predicting cervical cancer using advanced machine learning algorithms. In *ICSCSS*, pages 1600–1604.

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 *IAICT*, pages 290–296. IEEE.

Tolles, W. E. and Bostrom, R. C. (1956). Automatic screening of cytological smears for cancer: The instrumentation. *Ann. N. Y. Acad. Sci.*, 63(6):1211–1218.

Vargas-Cardona, H. D., Rodriguez-Lopez, M., Arrivillaga, M., Vergara-Sanchez, C., García-Cifuentes, J. P., Bermúdez, P. C., and Jaramillo-Botero, A. (2024). Artificial intelligence for cervical cancer screening: Scoping review, 2009–2022. *International Journal of Gynecology & Obstetrics*, 165(2):566–578.

Wu, T., Lucas, E., Zhao, F., Basu, P., and Qiao, Y. (2024). Artificial intelligence strengthenes cervical cancer screening–present and future. *Cancer Biol. Med.*, 21(10):864.

Wubineh, B. Z., Rusiecki, A., and Halawa, K. (2024). Segmentation and classification techniques for pap smear images in detecting cervical cancer: A systematic review. *IEEE Access*, 12:118195–118213.