Journal of Health Informatics

J Health Inform. 2026;18:e1880
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1880

ARTIGO COMPLETO
CBIS 2026

Correção de viés na confiabilidade diagnóstica de sistemas de inteligência artificial para citologia cervical

Bias correction in the diagnostic reliability of artificial intelligence systems for cervical cytology

Arianny Raphaelly Paiva Leão[1] ✉, Gabrielle Ribeiro Sena[2], David Edson Ribeiro[1], Hélio Fernando Bentzen Pessoa Filho[1], Rosa Maria Oliveira Teixeira de Vasconcelos[1], Tiago Pessoa Ferreira de Lima[1]

ORCID: 0009-0000-1147-0751 · 0000-0002-8430-3599 · 0009-0006-1063-6813 · 0000-0002-3089-1639 · 0000-0002-6719-962X · 0000-0002-1148-4288

[1] Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco, Palmares (PE), Brasil.

[2] Universidade de Pernambuco, Recife (PE), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1880

DOI
10.59681/2175-4411.v18.iEspecial.2026.1880

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Arianny Raphaelly Paiva Leão
arpl@discente.ifpe.edu.br

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Introdução: O desempenho de sistemas de diagnóstico citológico é frequentemente comprometido por classes desbalanceadas. **Objetivo:** Avaliar estratégias para mitigar esse viés. **Método:** Unificaram-se dois conjuntos de imagens padronizados pelo Sistema Bethesda e empregou-se o particionamento estratificado por classe para prevenir vazamento de informações. Foram avaliadas três arquiteturas de rede, e adotaram-se estratégias de reamostragem e perda focal para corrigir o viés de probabilidade prévia. **Resultados:** A correção do viés foi associada ao aumento da sensibilidade para detecção de lesões de alto grau, de 44,2% para 91,9%, com F1-Macro de 0,94 e área sob a curva ROC de 0,99. Análises por métodos de interpretabilidade confirmaram a consistência morfológica das predições e permitiram identificar vieses de atalho durante o pré-processamento. **Conclusão:** A mitigação do viés de probabilidade prévia é determinante para a eficácia clínica do sistema, sobrepondo-se à escolha da arquitetura computacional.

Descritores: Neoplasias do Colo do Útero; Aprendizado Profundo; Inteligência Artificial.

ABSTRACT

Introduction: The performance of cytological diagnostic systems is often compromised by class imbalance. **Objective:** To evaluate strategies to mitigate this bias. **Method:** Two image datasets standardized by the Bethesda System were combined, and a class-stratified partitioning scheme was applied to prevent information leakage. Three network architectures were evaluated, and resampling strategies along with focal loss were employed to correct prior probability bias. **Results:** Correction of this bias was associated with an increased sensitivity for high-grade lesion detection from 44.2% to 91.9%, yielding a macro F1 of 0.94 and an area under the ROC curve of 0.99. Interpretability analyses confirmed the morphological consistency of predictions and enabled identification of shortcut biases introduced during preprocessing. **Conclusion:** Mitigation of prior probability bias is critical to the clinical effectiveness of the system, outweighing the choice of computational architecture.

Keywords: Uterine Cervical Neoplasms; Deep Learning; Artificial Intelligence.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Leão ARP, Sena GR, Ribeiro DE, Pessoa Filho HFB, de Vasconcelos RMOT, de Lima TPF. Correção de viés na confiabilidade diagnóstica de sistemas de inteligência artificial para citologia cervical. J Health Inform 2026;18:e1880. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1880

1 INTRODUÇÃO

O câncer do colo do útero permanece sendo um dos desafios mais críticos para a saúde pública no Brasil, com estimativas do Instituto Nacional de Câncer (INCA) apontando, para o triênio de 2026-2028, o número de casos estimados em 19.310, sendo o terceiro tipo de câncer mais incidente nas mulheres, com o risco estimado de 17,59 casos a cada 100 mil mulheres [1].

Apesar de ser uma patologia evitável por meio de programas de rastreio citopatológico, a eficácia do exame de Papanicolaou é limitada pela natureza subjetiva da análise manual, que está sujeita à fadiga do patologista e à variabilidade interobservador [2-4]. Nesse cenário, a integração de sistemas de Inteligência Artificial (IA) na citologia cervical tem sido explorada para automatizar a identificação de lesões precursoras e aumentar a precisão diagnóstica em larga escala [3].

A evolução tecnológica nesta área avançou de algoritmos baseados em características morfológicas extraídas manualmente para modelos de *Deep Learning* de ponta [2,4]. Atualmente, as Redes Neurais Convolucionais (CNNs) dominam a detecção citológica, porém enfrentam limitações em capturar dependências espaciais globais e relações contextuais de longo alcance na célula [2,4]. Para superar essas falhas, surgiram as arquiteturas baseadas em *Vision Transformers* (ViTs), que utilizam mecanismos de *self-attention* para modelar dependências estruturais globais, demonstrando superioridade em tarefas de classificação multiclasse [4-6].

Contudo, a adoção clínica desses modelos enfrenta barreiras relacionadas à opacidade dos algoritmos ("caixas-pretas") e à falta de robustez metodológica. A literatura alerta para falhas comuns em IA médica, como o vazamento de dados (*data leakage*), que ocorre quando o isolamento das amostras não respeita o nível do paciente, resultando em métricas de acurácia artificialmente infladas [7,8]. Além disso, a falta de interpretabilidade impede que o citopatologista valide visualmente as decisões do modelo, tornando os sistemas de IA Explicável (XAI) fundamentais para a transparência diagnóstica [9-11].

Diante deste cenário, o presente trabalho propõe o desenvolvimento de um sistema CAD (*Computer-Aided Diagnosis*) híbrido que integra CNNs e ViT em um *Stacking Ensemble* explicável. A proposta utiliza as bases de dados Herlev e SIPaKMeD [12,13], adotando uma divisão estratificada por classe para evitar o vazamento de dados [8] e o mapeamento biológico harmonizado ao Sistema Bethesda [14].

2 MÉTODOS

Esta pesquisa adotou abordagem experimental quantitativa, implementada em Python 3.11, PyTorch 2.5.1 e scikit-learn 1.7.2. Os experimentos foram realizados em computador com Intel Core i5-12450HX, 8 GB de RAM e NVIDIA GeForce RTX 3050 (6 GB).

2.1 Fonte de Dados

Foram utilizados dois *datasets* públicos de referência na abordagem automatizada de câncer cervical. O SIPaKMeD [12] contém 4.049 imagens de células individuais segmentadas a partir de esfregaços de Papanicolaou, organizadas em cinco classes morfológicas, sendo duas normais (superficial-intermediária e parabasal), duas anormais (koilocítica e disceratótica) e uma benigna (metaplástica). O Herlev [13] contém 917 imagens em sete classes baseadas no grau de displasia. Essa combinação foi selecionada por introduzir diferenças reais de resolução, intensidade de coloração e protocolo de aquisição, reproduzindo as condições de um laboratório que recebe lâminas de múltiplas unidades de saúde. O SIPaKMeD utiliza coloração de Papanicolaou, enquanto o Herlev emprega Hematoxilina e Eosina (H&E). Ambos os *datasets* foram obtidos em suas versões oficiais integrais, sem exclusão de imagens, dado que nenhum arquivo corrompido foi identificado durante a inspeção do conjunto.

2.2 Unificação Pelo Sistema Bethesda

Os rótulos originais de ambas as bases de dados (totalizando 12 classes morfológicas) foram mapeados para as categorias primárias do Sistema Bethesda (TBS) [14]: *Negative for Intraepithelial Lesion or Malignancy* (NILM), *Low-grade Squamous Intraepithelial Lesion* (LSIL) e *High-grade Squamous Intraepithelial Lesion* (HSIL). A adoção desta separação fundamenta-se nos três princípios bases do TBS: a terminologia deve ser (1) clinicamente relevante, (2) razoavelmente reprodutível e (3) refletir o entendimento biológico mais atual da neoplasia cervical.

Adicionalmente, o TBS estabelece uma categorização geral focada em separar de forma clara aquilo que é "normal" de "anormal", com o objetivo de auxiliar clínicos na triagem para o manejo dos pacientes. Em conformidade com essa diretriz, as células superficiais, intermediárias, colunares, as células metaplásicas e parabasais foram alocadas no grupo de normalidade (NILM).

2.3 Estratificação

A estratificação foi conduzida diretamente sobre os rótulos do Sistema Bethesda, garantindo que a distribuição de classes no treinamento reflita a prevalência real e prevenindo o *Prior Probability Shift* [15]. O SIPaKMeD foi particionado em nível de célula, e o Herlev submetido a estratificação proporcional padrão, dado que nenhum dos conjuntos dispõe de metadados de identificação individual que permitissem o isolamento por paciente. Essa abordagem assegura a separação estrita entre partições, evitando o vazamento de padrões morfológicos e de coloração entre os

conjuntos de treino, validação e teste, principal fonte de otimismo inflacionado em métricas de validação em IA médica [7,8]. A proporção resultante foi de 73,1/13,5/13,4%, diferindo marginalmente do esquema planejado de 70/15/15%, desvio esperado e metodologicamente justificável pela estratificação proporcional sobre classes desbalanceadas.

2.4 Pré-processamento

O *pipeline* não incluiu equalização de contraste, preservando a heterogeneidade natural entre os laboratórios para avaliar a capacidade de generalização intrínseca das arquiteturas sob *domain shift* real [16]. O redimensionamento para 224×224 pixels foi realizado via PadToSquare em modo reflexivo, suprimindo o *Shortcut Learning* induzido por gradientes de alta frequência espacial que preenchimentos constantes introduzem após a normalização ImageNet [9].

2.5 Arquiteturas de Aprendizado Profundo

Foram avaliadas três arquiteturas pré-treinadas no ImageNet-1K em regime de *fine-tuning* completo, sem congelamento de camadas, com todos os parâmetros atualizados durante o treinamento. A ResNet-50 [2,11] opera por convolução hierárquica com conexões residuais, sendo especialmente adequada para a identificação de texturas nucleares e bordas citoplasmáticas em imagens de alta variabilidade local. A EfficientNet-B0 [11,17] aplica escalonamento composto simultâneo de largura, profundidade e resolução da rede, oferecendo maior eficiência paramétrica sob restrições computacionais sem sacrificar a capacidade representacional. O ViT-B/16 [5, 6,11] divide a imagem em *patches* de 16×16 pixels e computa atenção global entre todos eles via mecanismo *self-attention*, capturando dependências morfológicas de longo alcance que arquiteturas convolucionais modelam de forma implícita. A inclusão do ViT teve caráter exploratório-comparativo dado que sua superioridade diagnóstica sobre CNNs em imagens citológicas ainda não está consolidada na literatura [3]. A otimização utilizou AdamW com *Focal Loss* [18] e WeightedRandomSampler, com hiperparâmetros individualizados por arquitetura via otimização bayesiana; o tamanho de *batch* otimizado foi de 16 imagens para todas as arquiteturas. O treinamento foi conduzido por até 50 épocas, com agendador *ReduceLROnPlateau* configurado para maximização do F1-Score de validação (fator de decaimento de 0,5 e paciência de 5 épocas) e *Early Stopping* com paciência de 15 épocas.

2.6 Tratamento do Desbalanceamento das Classes

Conforme destacado por revisões sistemáticas recentes na área de citologia cervical [2,3], o desbalanceamento intrínseco de classes é um fator crítico que frequentemente leva ao enviesamento do modelo e à superestimação do desempenho em classes majoritárias. Para mitigar esse efeito, foram adotadas duas estratégias combinadas: WeightedRandomSampler, com pesos de amostragem inversamente proporcionais à frequência de cada classe, e *Focal Loss* [18], que reduz a contribuição de amostras fáceis e concentra o gradiente em exemplos morfologicamente ambíguos. O parâmetro γ da *Focal Loss* foi calibrado via otimização bayesiana, com valores buscados no intervalo (1,0; 3,0), convergindo para γ = 1,19 nas três arquiteturas.

2.7 Otimização Bayesiana em Cascata

A busca de hiperparâmetros foi conduzida por otimização bayesiana em duas fases, utilizando o algoritmo *Tree-structured Parzen Estimator* (TPE) com poda por mediana (Optuna), maximizando o F1-Macro de validação sobre 50% do conjunto de treinamento. Na Fase 1, todos os hiperparâmetros (taxa de aprendizado, decaimento de pesos, tamanho de lote, γ da *Focal Loss* e parâmetros de aumento de dados) foram otimizados sobre a ResNet-50 em 30 tentativas com limite de 8 épocas cada; essa arquitetura foi escolhida por sua estabilidade de convergência e menor custo computacional em relação ao ViT-B/16. Na Fase 2, esses parâmetros foram fixados e transferidos para EfficientNet-B0 e ViT-B/16, restringindo a busca à taxa de aprendizado e ao decaimento de pesos em 15 tentativas cada, com limite de 8 e 12 épocas respectivamente. Além de reduzir o custo total da busca, o compartilhamento dos parâmetros entre arquiteturas induz uma calibração estruturada no *ensemble*, favorecendo a diversidade complementar de erros entre redes heterogêneas [19].

2.8 Stacking Ensemble Oof

Os três modelos base foram treinados em K=5 *folds* com validação cruzada estratificada por grupo (StratifiedGroupKFold) sobre o conjunto de treinamento (n=3.631), gerando probabilidades *out-of-fold* (OOF) normalizadas via *softmax* para cada amostra sem exposição no *fold* de validação. Todos os componentes foram fixados com seed=42 para garantir reprodutibilidade. Esse procedimento elimina o vazamento de dados e a superestimação de desempenho que ocorreriam se as predições de treino fossem usadas diretamente como características do metamodelo, atendendo às recomendações de confiabilidade em IA médica [7,8]. As probabilidades OOF de dimensão 3.631 × 9 (3 modelos × 3 classes) alimentaram um *Grid Search* com validação cruzada de 5 *folds*, avaliando três metamodelos candidatos: Floresta Aleatória, Regressão Logística e *Gradient Boosting*, com o F1-Macro como critério de seleção. O metamodelo final selecionado, no caso Floresta Aleatória, foi avaliado exclusivamente no conjunto de teste isolado, o qual permaneceu inacessível durante todas as etapas prévias de treinamento e seleção.

2.9 Validação Estatística

Para evitar as armadilhas comuns de avaliação pontual em IA médica [7], a validade estatística dos resultados foi quantificada por duas abordagens complementares. O *Bootstrap* com 2.000 iterações (seed=42), implementado com NumPy, gerou intervalos de confiança de 95% pelo método percentil, tendo como métrica primária a sensibilidade da classe HSIL e como métrica secundária o F1-Macro. Complementarmente, foram calculadas sensibilidade, especificidade, valor preditivo positivo e negativo para a classe HSIL, conferindo interpretabilidade clínica aos resultados. A comparação pareada dos padrões de erro entre modelos foi conduzida pelo Teste de McNemar com correção de continuidade de Edwards, implementado com SciPy, adotando nível de significância α=0,05 [20].

2.10 Explicabilidade (Xai)

Aplicou-se Grad-CAM [21] com implementação via *hooks* de gradiente, tendo como camada alvo a terceira convolução do último bloco residual da ResNet-50 e a primeira convolução do último bloco de características da EfficientNet-B0. Para o ViT-B/16, os gradientes foram extraídos da camada de normalização (LayerNorm 1) anterior ao último bloco do *encoder*; os 196 *tokens* de *patch* foram isolados do *token* CLS e remodelados para grade 14×14. A fidelidade das explicações foi auditada quantitativamente pelo protocolo *Remove And Debias* (ROAD) [22]. O procedimento foi direcionado ao ViT-B/16, devido à complexidade de sua atenção global em contextos clínicos, consistindo em ROADMostRelevantFirst (remoção dos 20% de pixels mais notáveis) e ROADLeastRelevantFirst (remoção dos 80% dos pixels menos notáveis). A fidelidade foi quantificada pela confiança média de predição após remoção dos pixels: valores próximos de zero no MostRelevantFirst indicam alta dependência causal das regiões destacadas, enquanto valores elevados no LeastRelevantFirst confirmam independência do modelo em relação a artefatos de fundo.

2.11 Aspectos Éticos

Este estudo utilizou exclusivamente *datasets* públicos estabelecidos na literatura científica: o SIPaKMeD [12], disponibilizado pelos autores para uso em pesquisa científica, e o Herlev [13], disponibilizado para fins acadêmicos desde sua publicação original. Ambos os conjuntos são compostos por imagens de lâminas citológicas sem qualquer identificação pessoal dos pacientes. Não houve coleta de dados de seres humanos ou acesso a prontuários clínicos no âmbito desta pesquisa. O trabalho enquadra-se na modalidade de desenvolvimento tecnológico de software, sem necessidade de aprovação pelo Comitê de Ética em Pesquisa.

3 RESULTADOS

3.1 Distribuição do Dataset Após Particionamento Estratificado

A aplicação do particionamento estratificado, com estratificação sobre os rótulos Bethesda, resultou na distribuição apresentada na Tabela 1. A proporção final de 73,1%/13,5%/13,4% (treino/validação/teste), levemente distinta dos 70%/15%/15% planejados, é um procedimento necessário para mitigar o vazamento de dados (*data leakage*) [7,8].

3.2 Desempenho dos Modelos Base e do Stacking Ensemble

Em experimento preliminar conduzido sem estratificação por classe e sem as estratégias de mitigação de desbalanceamento, a sensibilidade para HSIL atingiu 44,2%. A Tabela 2 apresenta as métricas dos três modelos base e do *Stacking Ensemble* OOF no conjunto de teste (N=664), com intervalos de confiança (IC) de 95% calculados pelo método de *Bootstrap* [23]. A Tabela 3 detalha as métricas por classe do *Stacking Ensemble*, destacando a classe HSIL.

A Figura 1 detalha os acertos e os falsos negativos por meio da matriz de confusão do *Stacking Ensemble* no conjunto de testes. E a Figura 2 (Curva ROC multiclasse) demonstra visualmente a capacidade discriminativa do *ensemble* (área sob a curva, AUC = 0,9965).

A Tabela 4 apresenta os resultados do Teste de McNemar com correção de continuidade de Edwards entre todos os pares de modelos base.

3.3 Métricas de Fidelidade Xai (Grad-cam/road)

A avaliação quantitativa de fidelidade causal pelo protocolo ROAD produziu os seguintes escores médios para a arquitetura ViT-B/16 sobre o conjunto de teste: ROADMostRelevantFirst = −0,35; ROADLeastRelevantFirst = −0,09.

Em paralelo, a avaliação espacial qualitativa (Grad-CAM) confirmou o aprendizado de características biológicas fidedignas pelo *ensemble*. A Figura 3 apresenta os mapas de ativação extraídos da arquitetura ResNet-50 para amostras representativas de cada classe, ilustrando a coerência morfológica do modelo, que focaliza ativamente o núcleo celular em casos de displasia de alto grau.

4 DISCUSSÃO

O achado central deste trabalho demonstrou que o desempenho clínico do *ensemble* era limitado por um viés sistemático de representação, o *Prior Probability Shift* [15], e não exclusivamente pela capacidade das arquiteturas. A estratificação direta sobre os rótulos do Sistema Bethesda [14] estabilizou a prevalência em torno de 11,1% e elevou o *Recall* de HSIL para 91,89% (IC

Tabela 1. Distribuição das amostras por classe Bethesda após particionamento estratificado

Partição	HSIL (%)	LSIL (%)	NILM (%)	N total
Treino	345 (9,5%)	1.334 (36,7%)	1.952 (53,8%)	3.631
Validação	74 (11,0%)	226 (33,7%)	371 (55,3%)	671
Teste	74 (11,1%)	260 (39,2%)	330 (49,7%)	664
Total	493 (9,9%)	1.820 (36,6%)	2.653 (53,4%)	4.966

Fonte: Elaborado pelos autores (2026).

Tabela 2. Desempenho comparativo no conjunto de teste

Modelo	Acurácia	F1-Macro (IC 95%)	*Recall* Macro	AUC Macro
ResNet-50	95,93%	0,9431 (0,9199–0,9629)	0,9394	0,9938
EfficientNet-B0	94,28%	0,9220 (0,8946–0,9453)	0,9202	0,9928
ViT-B/16	95,33%	0,9384 (0,9139–0,9601)	0,9455	0,9966
Stacking OOF	95,63%	0,9403 (0,9167–0,9617)	0,9462	0,9965

Fonte: Elaborado pelos autores (2026).

Tabela 3. Métricas por classe do Stacking Ensemble OOF no conjunto de teste

Classe	*Precision*	Recall (IC 95%)	F1-Score	*Support*
HSIL	0,87	0,9189 (0,8553–0,9744)	0,89	74
LSIL	0,96	0,95	0,95	260
NILM	0,98	0,97	0,97	330
Stacking OOF	0,93	0,9462	0,94	664

Fonte: Elaborado pelos autores (2026).

Tabela 4. Teste de McNemar

Par	n_{01}	n_{10}	p-valor	Significativo (α=0,05)
ResNet-50 vs EfficientNet	22	11	0,082	Não
ResNet-50 vs ViT	17	13	0,584	Não
EfficientNet vs ViT	12	19	0,281	Não

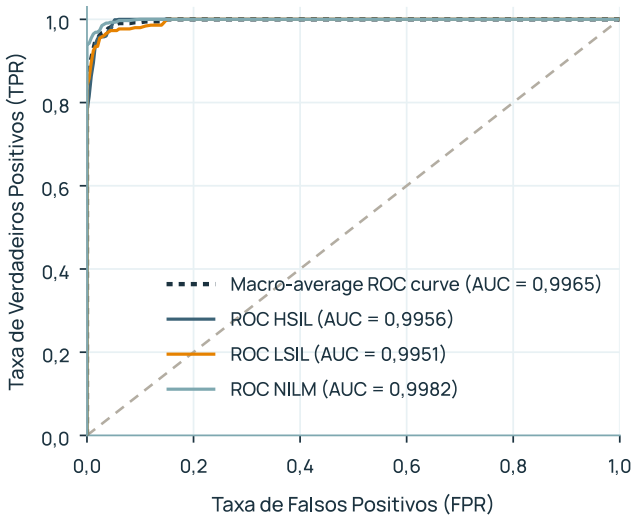
Fonte: Elaborado pelos autores (2026).

Figura 1. Matriz de confusão do Stacking Ensemble no conjunto de teste.



Fonte: Elaborado pelos autores (2026).

Figura 2. Gráfico de curva ROC do modelo multiclasse



Fonte: Elaborado pelos autores (2026).

95%: 85,53%–97,44%). Embora a contribuição isolada de cada componente não tenha sido quantificada formalmente, este resultado sugere que o alinhamento entre a distribuição de treinamento e a prevalência epidemiológica real é determinante para a eficácia clínica do modelo, corroborando revisões que identificam a representatividade dos *datasets* como principal barreira de generalização [2,3].

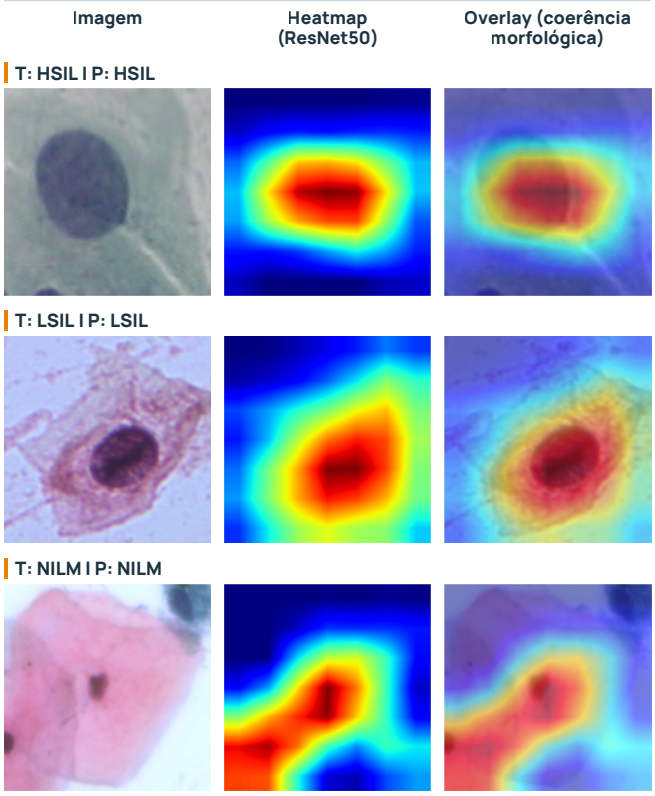
A interpretabilidade dos modelos revelou nuances cruciais sobre o comportamento das arquiteturas. A validação visual via Grad-CAM [21] demonstrou que a ResNet-50 obteve sucesso em identificar características morfológicas, focalizando o núcleo hipercromático e expandido em casos de HSIL e avaliando corretamente a relação núcleo-citoplasma (N/C) em amostras NILM, o que ancora o aprendizado aos critérios patológicos do Sistema Bethesda [14].

A ausência de significância estatística no Teste de McNemar [20] reflete a estratégia de Otimização Bayesiana em Cascata. Ao compartilhar configurações de aumento de dados entre as redes, induziu-se uma convergência parcial nas invariâncias aprendidas, reduzindo a descorrelação de erros entre os modelos base [19,24]. Nesse cenário, o ganho do *Stacking* não decorre da diversidade de erros individuais, mas da calibração probabilística do metamodelo sobre o espaço conjunto de predições, mecanismo que confere maior consistência às estimativas de confiança.

Diferentemente das CNNs, o ViT carece de *inductive bias* espacial local, como localidade e correspondência translacional, tornando-o mais dependente de grandes volumes de dados para aprender invariâncias morfológicas [4,5]. Observou-se, de forma consistente com essa predição teórica, uma sensibilidade anômala do mecanismo de atenção a artefatos de pré-processamento geométrico. O uso de preenchimento reflexivo criou transições de alta frequência que o modelo interpretou como regiões de interesse, evidenciando costuras artificiais nas extremidades em vez do núcleo celular. Este viés de atalho não invalida a capacidade preditiva do ViT (F1-Macro de 0,9384), mas aponta uma limitação das ferramentas de XAI baseadas em gradiente ao mapear o raciocínio global por *patches* [21]. Isso reforça que métricas de acurácia elevadas são condição necessária, mas não suficiente, para a adoção clínica segura de sistemas de IA.

Por fim, os erros do *Stacking Ensemble* para HSIL merecem interpretação clínica explícita. A *Precision* de 0,87 implica aproximadamente 13% de falsos positivos, casos NILM ou LSIL classificados como HSIL e encaminhados desnecessariamente para colposcopia. Em contexto de triagem oncológica, esse perfil é clinicamente aceitável [1]: o custo do encaminhamento desnecessário é mensurável, reversível e operacionalmente gerenciável, enquanto o falso negativo dessa classe representa uma lesão precursora não sinalizada

Figura 3. Mapas de ativação Grad-CAM (ResNet-50) ilustrando o foco atencional do modelo no núcleo hipercromático (HSIL), no núcleo atípico (LSIL) e na relação núcleo-citoplasma (NILM).



Fonte: Elaborado pelos autores (2026).

com risco de progressão para carcinoma invasivo [1, 25]. Isso evidencia que as decisões de design do sistema devem ser guiadas por consequências clínicas, e não apenas por métricas globais de desempenho.

A avaliação sobre *datasets* digitalizados com anotações especializadas pré-existentes garante a validade interna do estudo, mas não substitui a validação prospectiva em fluxo clínico real, etapa necessária para que os ganhos de sensibilidade observados se traduzam em impacto assistencial mensurável. A ausência de metadados de identificação individual em ambos os *datasets* impediu o isolamento em nível de paciente, introduzindo risco residual de *data leakage* por possível coocorrência de imagens de uma mesma lâmina em partições distintas. Adicionalmente, embora a combinação de *datasets* de origens distintas reproduza condições próximas às de um laboratório, essa heterogeneidade constitui também uma fonte de viés de domínio não controlado. Estudos futuros devem explorar técnicas de normalização de coloração e adaptação de domínio para mitigar esse viés e avaliar a generalização em populações com distribuições distintas. Além disso, uma limitação metodológica adicional deste estudo é a aplicação simultânea de múltiplas intervenções, como a correção do *Prior Probability Shift*, otimização de hiperparâmetros e *ensembling*. A ausência de um estudo de ablação (*ablation study*) formal no escopo atual impede a quantificação isolada

da contribuição de cada componente para a métrica final e a atribuição de causalidade a um único fator.

5 CONCLUSÃO

Este trabalho desenvolveu e validou um sistema CAD para classificação de células cervicais. Embora a ResNet-50 tenha apresentado o pico de desempenho isolada em F1-Macro (0,9431), a integração de arquiteturas CNN e ViT em um *Stacking Ensemble* (F1-Macro de 0,9403) atende aos requisitos de alta sensibilidade diagnóstica ao oferecer redundância preditiva contra falsos negativos. Os resultados demonstram que a correção do *Prior Probability Shift* e um isolamento estrito estão fortemente associados à obtenção de desempenhos clinicamente relevantes, sugerindo que esses fatores exercem um forte papel na eficácia do sistema. Análises de interpretabilidade (Grad-CAM- e protocolo ROAD) indicaram que as decisões do modelo se fundamentam em estruturas morfológicas relevantes, aumentando a confiança clínica. A execução em hardware comercial de médio porte e o uso de bases públicas sugerem viabilidade de implantação em laboratórios do Sistema Único de Saúde (SUS). Trabalhos futuros deverão contemplar estudos de ablação, avaliação de técnicas de normalização de domínio e testes de usabilidade com citopatologistas para validação em ambiente clínico.

Contribuição dos autores. Os autores contribuíram igualmente em todas as etapas de construção do artigo: concepção e delineamento do estudo; coleta, análise e interpretação dos dados; redação e revisão crítica do manuscrito; e aprovação da versão final.

Agradecimentos. Os autores agradecem ao Instituto Federal de Pernambuco (IFPE) campus Palmares pelo custeio e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento, concedidos por meio do Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI).

Financiamento. O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio de bolsa do Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI).

Disponibilidade de dados e materiais. Os conjuntos de dados analisados são públicos e estão disponíveis em: SipakMed, https://www.kaggle.com/datasets/prahladmehandiratta/cervical-cancer-largest-dataset-sipakmed; Herlev, https://www.kaggle.com/datasets/yuvrajsinhachowdhury/herlev-dataset.

Conflitos de interesse. Os autores não declararam conflitos de interesse.

Declaração de uso de IA. Os autores declaram que utilizaram modelos de linguagem de inteligência artificial (Gemini e Claude) apenas como ferramentas de apoio à redação e revisão do manuscrito (clareza, gramática e estilo) e para auxílio na organização e formatação do texto. Não foi empregada IA para gerar, manipular ou interpretar dados, resultados, tabelas ou conclusões.

REFERÊNCIAS

1. Instituto Nacional de Câncer. Estimativa 2026: incidência de câncer no Brasil [Internet]. Rio de Janeiro: INCA; 2026 [citado 15 abr 2026]. Disponível em: https://www.gov.br/inca/pt-br/assuntos/noticias/2026/inca-estima-781-mil-novos-casos-de-cancer-por-ano-no-brasil-entre-2026-e-2028.
2. Jiang P, Li X, Shen H, Chen Y, Wang L, Chen H, et al. A Survey on Deep Learning-based Cervical Cytology Screening: from Cell Identification to Whole Slide Image Analysis. Research Square [Preprint]. 2023. doi: https://doi.org/10.21203/rs.3.rs-2680912/v1
3. Valles-Coral MA, Pinedo L, Rodríguez C, Rodríguez D, Sánchez-Dávila K, Arévalo-Fasanando L, et al. Application of artificial intelligence in cervical cytology: a systematic review of deep learning models, datasets, and reported metrics. Front Big Data. 2025;8:1678863. doi: https://doi.org/10.3389/fdata.2025.1678863
4. Al-Hejri AM, Al-Tam RM, Sable AH, Almuhaya B, Alshamrani SS, Alshmrany KM. A hybrid vision transformer with ensemble CNN framework for cervical cancer diagnosis. BMC Med Inform Decis Mak. 2025;25(1):411. doi: https://doi.org/10.1186/s12911-025-03250-x
5. Lima MV, Britto MHM, Britto Neto L. Classificação de Células Cervicais Cancerígenas com CNNs e ViTs em Ensemble para Auxílio ao Diagnóstico Médico. In: Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2025). [s.l.]: Sociedade Brasileira de Computação - SBC; 2025. p. 497-508. doi: https://doi.org/10.5753/sbcas.2025.7467
6. AlMohimeed A, Shehata M, El-Rashidy N, Mostafa S, Samy Talaat A, Saleh H. ViT-PSO-SVM: Cervical Cancer Predication Based on Integrating Vision Transformer with Particle Swarm Optimization and Support Vector Machine. Bioengineering (Basel). 2024;11(7):729. doi: https://doi.org/10.3390/bioengineering11070729
7. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3(3):199-217. doi: https://doi.org/10.1038/s42256-021-00307-0
8. Prayitnaningsih S, Syarifuddin OA, Dhani FK, Novita HD, Samsu N, Sasongko MB, et al. Development and Validation of a Novel Deep Learning-Based Model for Detection of Diabetic Kidney Disease from Retinal Imaging Using a Weighted Loss Method. Clin Ophthalmol. 2026;20:586474. doi: https://doi.org/10.2147/OPTH.S586474
9. Attallah O. Light-XAI: a CADx for explainable cervical cancer detection via attention-based lightweight convolutional neural networks and layer-wise feature fusion. BioData Min. 2026;19(1):26. doi: https://doi.org/10.1186/s13040-026-00540-6
10. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med. 2022;140:105111. doi: https://doi.org/10.1016/j.compbiomed.2021.105111

11. Panboonyuen T. Seeing Isn't Always Believing: Analysis of Grad-CAM Faithfulness and Localization Reliability in Lung Cancer CT Classification. In: 2026 18th International Conference on Knowledge and Smart Technology (KST). [s.l.]: IEEE; 2026. p. 80-6. doi: https://doi.org/10.1109/kst67832.2026.11432175
12. Plissiti M, Dimitrakopoulos P, Sfikas G, Nikou C, Krikoni O, Charchanti A. Sipakmed: A New Dataset for Feature and Image Based Classification of Normal and Pathological Cervical Cells in Pap Smear Images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). [s.l.]: IEEE; 2018. p. 3144-8. doi: https://doi.org/10.1109/icip.2018.8451588
13. Jantzen J, Norup J, Dounias G, Bjerregaard B. Pap-smear benchmark data for pattern classification. In: Proceedings of Nature inspired Smart Information Systems (NiSIS); 2005; Albufeira. p. 1-9.
14. Nayar R, Wilbur DC, editores. The Bethesda System for Reporting Cervical Cytology. 3. ed. Cham: Springer International Publishing; 2015. doi: https://doi.org/10.1007/978-3-319-11074-5
15. Huang MY, Qin J, Huang CY. Efficient data integration under prior probability shift. Biometrics. 2024;80(2):ujae035. doi: https://doi.org/10.1093/biomtc/ujae035
16. Coskun G, Akuner MC, Kaplanoglu E. Robust Cell-Level Classification for Liquid-Based Cervical Cytology Using Deep Transfer Learning: A Multi-Source Study Addressing Scanner-Induced Domain Shifts. Bioengineering (Basel). 2026;13(3):289. doi: https://doi.org/10.3390/bioengineering13030289
17. Kurita Y, Meguro S, Tsuyama N, Kosugi I, Enomoto Y, Kawasaki H, et al. Accurate deep learning model using semi-supervised learning and Noisy Student for cervical cancer screening in low magnification images. PLoS One. 2023;18(5):e0285996. doi: https://doi.org/10.1371/journal.pone.0285996
18. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). [s.l.]: IEEE; 2017. p. 2999-3007. doi: https://doi.org/10.1109/iccv.2017.324
19. Wenzel F, Snoek J, Tran D, Jenatton R. Hyperparameter ensembles for robustness and uncertainty quantification [Internet]. arXiv:2006.13570 [Preprint]. 2020 jun [citado 29 abr 2026]. Disponível em: https://arxiv.org/abs/2006.13570.
20. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 1998;10(7):1895-923. doi: https://doi.org/10.1162/089976698300017197
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis. 2020;128(2):336-59. doi: https://doi.org/10.1007/s11263-019-01228-7
22. Rong Y, Leemann T, Borisov V, Kasneci G, Kasneci E. A Consistent and Efficient Evaluation Strategy for Attribution Methods [Internet]. In: Proceedings of the 39th International Conference on Machine Learning (ICML); 2022 jul 17-23; Baltimore. p. 18770-18795. Disponível em: https://arxiv.org/abs/2202.00449.
23. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Boston, MA: Springer US; 1993. doi: https://doi.org/10.1007/978-1-4899-4541-9
24. Gontijo-Lopes R, Dauphin Y, Cubuk ED. No one representation to rule them all: overlapping features of training methods. In: International Conference on Learning Representations (ICLR 2022). 2022. [citado 2026 out 18]. Disponível em: https://arxiv.org/abs/2110.12899
25. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301-9. doi: https://doi.org/10.1038/s41591-019-0508-1

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.

Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.

SBIS Sociedade Brasileira de Informática em Saúde