Journal of Health Informatics

J Health Inform. 2026;18:e1801
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1801

ARTIGO COMPLETO
CBIS 2026

CNNs e vision transformers para classificação de nódulos pulmonares em TC com representação 2.5D

CNNs and vision transformers for classification of pulmonary nodules in CT with 2.5D representation

Moisés José Moreira Ribeiro[1] ✉, Pedro Moises de Sousa[1]

ORCID: 0009-0007-9375-992X · 0000-0003-4563-0033

[1] Universidade Federal de Viçosa, Campus Rio Paranaíba (UFV-CRP), Rio Paranaíba (MG), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1801

DOI
10.59681/2175-4411.v18.iEspecial.2026.1801

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
Moisés José Moreira Ribeiro
moises.ribeiro@ufv.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

Objetivo: Avaliar o impacto de diferentes arquiteturas e otimizadores na classificação binária de nódulos pulmonares em imagens de tomografia computadorizada. **Método:** Foram comparadas cinco arquiteturas (ResNet-50, ConvNeXt-Tiny, MobileNetV3-Large, ViT-B/16 e Swin-Tiny) a partir de uma representação 2.5D e os otimizadores Adam e SGD para treinamento. Além disso, foi realizada uma análise de interpretabilidade baseada em Grad-CAM. **Resultados:** O modelo Swin-Tiny com Adam obteve o melhor desempenho global, com F1-Score de 0,9690 e a maior sensibilidade (0,9609). A MobileNetV3-Large com SGD apresentou a maior especificidade (0,9832), com menor custo computacional, porém com sensibilidade inferior (0,8827). O ViT-B/16 apresentou desempenho inferior em comparação às demais arquiteturas. Mapas Grad-CAM indicaram que a ConvNeXt-Tiny apresentou ativações mais localizadas sobre as regiões de interesse clínico. A escolha do otimizador mostrou impacto relevante no desempenho dos modelos. **Conclusão:** Modelos hierárquicos e leves mostram potencial para aplicações em cenários de diagnóstico assistido.

Descritores: Inteligência Artificial; Neoplasias Pulmonares; Tomografia Computadorizada por Raios X.

ABSTRACT

Objective: To evaluate the impact of different architectures and optimizers on the binary classification of pulmonary nodules in computed tomography images. **Method:** Five architectures (ResNet-50, ConvNeXt-Tiny, MobileNetV3-Large, ViT-B/16, and Swin-Tiny) were compared using a 2.5D representation, alongside Adam and SGD optimizers for training. Additionally, an interpretability analysis based on Grad-CAM was conducted. **Results:** The Swin-Tiny model with Adam achieved the best overall performance, with an F1-Score of 0.9690 and the highest sensitivity (0.9609). MobileNetV3-Large with SGD yielded the highest specificity (0.9832) at a lower computational cost, but with lower sensitivity (0.8827). ViT-B/16 showed inferior performance compared to the other architectures. Grad-CAM maps indicated that ConvNeXt-Tiny presented more localized activations over clinically relevant regions. The choice of optimizer demonstrated a substantial impact on model performance. **Conclusion:** Hierarchical and lightweight models show potential for applications in computer-aided diagnosis scenarios.

Keywords: Artificial Intelligence; Lung Neoplasms; Tomography, X-Ray Computed.

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: Ribeiro MJM, Sousa PM. CNNs e vision transformers para classificação de nódulos pulmonares em TC com representação 2.5D. J Health Inform 2026;18:e1801. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1801

1 INTRODUÇÃO

Cânceres são compreendidos como um conjunto de doenças caracterizadas pela proliferação descontrolada de células anormais capazes de invadir tecidos e disseminar-se para outras partes do organismo por meio da corrente sanguínea ou do sistema linfático, processo conhecido como metástase [1]. Entre os diferentes tipos da doença, o câncer de pulmão destaca-se por apresentar elevada incidência e a maior taxa de mortalidade entre os cânceres mais incidentes. No mundo, estima-se que tenham ocorrido aproximadamente 2,5 milhões de novos casos e 1,8 milhão de mortes por câncer de pulmão em 2022, representando cerca de 18,7% das mortes por câncer [2]. No Brasil, as estimativas mais recentes do Instituto Nacional de Câncer (INCA) para o triênio 2026-2028 indicam elevada carga da doença no país, com destaque para tumores de traqueia, brônquio e pulmão entre os tipos de câncer mais incidentes [3].

Apesar dos avanços no tratamento oncológico, um dos principais desafios no enfrentamento da doença é o diagnóstico tardio. Em grande parte dos casos, o câncer de pulmão é identificado apenas em estágios avançados, quando as opções terapêuticas são mais limitadas e as chances de sobrevida do paciente são reduzidas. Dados da Cancer Research UK indicam que apenas 8,85% dos pacientes são diagnosticados em estágio inicial, evidenciando a necessidade de métodos que auxiliem na detecção precoce [4].

Nesse contexto, técnicas de Aprendizado de Máquina aplicadas à análise de imagens médicas têm demonstrado grande potencial para apoiar o diagnóstico clínico. Modelos de *Deep Learning*, especialmente redes neurais convolucionais (CNNs), têm apresentado bons resultados na identificação automática de padrões em exames radiológicos, contribuindo para detecção precoce de nódulos pulmonares [5,6]. Entretanto, o desempenho desses modelos pode variar significativamente de acordo com a arquitetura, estratégias de pré-processamento, qualidade e volume dos dados, além dos algoritmos de otimização empregados durante o treinamento [7].

Este trabalho propõe e avalia um conjunto de abordagens baseadas em *Deep Learning*, incluindo arquiteturas convolucionais e modelos baseados em Transformers, para classificação de nódulos pulmonares em imagens de tomografia computadorizada (TC). São comparadas cinco arquiteturas pré-treinadas, a saber: ResNet-50, ConvNeXt-Tiny, MobileNetV3-Large, ViT-B/16 e Swin-Tiny, utilizando representação 2.5D, e os otimizadores Adam e SGD para treinamento. Os modelos são treinados no conjunto de dados Lung Nodule Analysis 2016 (LUNA16) e validados externamente no IQ-OTH/NCCD, com análise de interpretabilidade por meio do Gradient-weighted Class Activation Mapping (Grad-CAM), que pode auxiliar na interpretação das predições pelo especialista. No contexto do Sistema Único de Saúde (SUS), tais abordagens podem contribuir para aplicações em diagnóstico assistido, especialmente em contextos com restrições de recursos. Nesse cenário, modelos leves como o MobileNetV3 destacam-se pelo potencial de aplicação em ambientes com recursos limitados, enquanto a interpretabilidade com o uso da Grad-CAM pode ajudar o especialista a compreender o comportamento do modelo, favorecendo a confiança no fluxo de triagem oncológica.

2 MÉTODOS

2.1 Conjunto de Dados

As imagens utilizadas neste estudo são provenientes do conjunto de dados público LUNA16 [8], derivado do repositório LIDC-IDRI e amplamente empregado como *benchmark* na tarefa de detecção de nódulos pulmonares. O conjunto é composto por 888 exames de TC de alta resolução, organizados originalmente em 10 subconjuntos (subset0 a subset9). O conjunto foi composto por 1.186 instâncias positivas provenientes do *annotations.csv*, nódulos validados por consenso radiológico, e igual número de instâncias negativas extraídas do *candidates_V2.csv* (class = 0), totalizando 2.372 *patches* balanceados. Por utilizar exclusivamente dados secundários, públicos e anonimizados, este trabalho está dispensado de submissão ao Comitê de Ética em Pesquisa (CEP), conforme as exceções éticas previstas na Resolução nº 510, de 7 de abril de 2016, do Conselho Nacional de Saúde.

2.2 Geração de Representações 2.5D

Os volumes originais de TC do conjunto de dados, armazenados em formato MHD/RAW, foram convertidos em uma representação 2.5D para permitir o uso de arquiteturas pré-treinadas em ImageNet via *transfer learning*. Essa abordagem preserva a informação contextual volumétrica ao extrair três cortes ortogonais centrados nas coordenadas do nódulo: axial (plano XY), coronal (plano XZ) e sagital (plano YZ), técnica consolidada por [9].

Para garantir que o recorte fosse extraído na localização anatômica correta, as coordenadas globais do nódulo (w), registradas em milímetros pelo scanner, foram convertidas para o espaço de voxels (v) através da equação: $v = round((w - o)/s)$ (a). Onde o representa a origem do volume e s o espaço entre voxels (*spacing*). Em casos em que o nódulo se localizava próximo às bordas do volume, aplicou-se um *padding* simétrico com valor de -1.000 unidades Hounsfield (UH) para evitar a perda de dados nos recortes de 224 x 224 pixels.

Finalmente, os valores de intensidade foram normalizados. Aplicou-se um redimensionamento (*clipping*) para janela de interesse entre -1.000 e 400 UH, seguido de um escalonamento linear para o intervalo

[0, 255]. Os três cortes foram então empilhados nos canais R, G e B, respectivamente, resultando na entrada final para a rede neural.

2.3 Particionamento, Pré-processamento e Aumento de Dados

O conjunto de dados gerado foi dividido nos conjuntos de treinamento (70% das imagens, totalizando 1.660 imagens), validação (15% das imagens, totalizando 354 imagens) e teste (15% das imagens, totalizando 358 imagens), com estratificação por classe para preservar a proporção entre nódulos e não nódulos em cada subconjunto. O conjunto de teste foi mantido isolado durante todo o desenvolvimento, sendo utilizado uma única vez, após a seleção do modelo final.

Todas as imagens possuem resolução 224 x 224 pixels, compatível com a entrada padrão dos modelos pré-treinados, não sendo necessário redimensionamento. A normalização foi realizada com a média e o desvio-padrão do ImageNet ([0,485; 0,456; 0,406] e [0,229; 0,224; 0,225], respectivamente), seguindo a prática de *transfer learning* a partir de pesos treinados nesse *corpus*.

Durante o treinamento, foram aplicadas técnicas de aumento de dados para reduzir o sobreajuste e aumentar a variabilidade das amostras vistas pelo modelo: espelhamento horizontal aleatório, espelhamento vertical aleatório e rotação aleatória de até 10°. Essas transformações são adequadas para imagens de TC pulmonar, dado que nódulos não apresentam orientação anatômica preferencial. Nos conjuntos de validação e teste, nenhum aumento de dados foi aplicado, utilizando-se apenas o redimensionamento e a normalização, de modo a garantir que a avaliação reflita o desempenho do modelo em condições representativas dos dados originais.

2.4 Arquiteturas Avaliadas

Foram avaliadas cinco arquiteturas pré-treinadas no ImageNet-1K, abrangendo CNNs e modelos baseados em atenção. Em todas, a camada de classificação final foi substituída por uma camada linear binária, com *fine-tuning* completo para os parâmetros.

Entre as CNNs, a ResNet-50 [10] utiliza conexões residuais (*skip connections*) para viabilizar o treinamento profundo sem degradação, sendo uma das *baselines* mais consolidadas em visão médica. O ConvNeXt-Tiny [11] moderniza o design convolucional incorporando convoluções *depthwise* 7x7, normalização por camada e ativação GELU, inspiradas em *Transformers*, mantendo desempenho competitivo mesmo na variante compactada. A MobileNetV3-Large [12] combina convoluções *depthwise* separáveis, atenção por canal (Squeeze-and-Excitation) e busca automatizada de arquitetura (NAS), sendo voltada para inferência eficiente com recursos limitados.

Entre os modelos baseados em atenção, o ViT-B/16 [13] divide a imagem em *patches* de 16x16 pixels tratados como *tokens* sequenciais, aplicando-se atenção multicabeça (*multi-head self-attention*) globalmente. O Swin Transformer Tiny [14] introduz uma variante hierárquica com atenção restrita a janelas locais deslocadas (*shifted windows*), reduzindo o custo quadrático de atenção global e aproximando a modelagem multiescala das CNNs.

2.5 Protocolos de Treinamento e Seleção de Modelos

Cada arquitetura foi treinada por 50 épocas com tamanho do lote (*batch_size*) de 32. Foram avaliados dois otimizadores: Adam (lr = $1x10^{-4}$, *weight decay* = $1x10^{-4}$) e SGD com *momentum* (lr = $1x10^{-2}$, *momentum* = 0,9, *weight decay* = $1x10^{-4}$). O escalonamento da taxa de aprendizado foi realizado com a estratégia *ReduceLROnPlateau* (fator 0,5, paciência de 5 épocas sem melhora na perda de validação).

O modelo salvo a cada experimento corresponde à época com maior F1-Score na validação, calculado tomando a classe normal como referência e sujeito à restrição de *recall* mínimo de 0,90 para essa classe, o que equivale a uma especificidade mínima de 0,90 para a classe nódulo. Esse critério, portanto, controla falsos positivos durante a seleção dos modelos; as métricas reportadas no conjunto de teste consideram a classe nódulo como positiva (Seção 2.6). Caso nenhuma época satisfaça essa restrição, é adotado como *fallback* o estado do modelo ao final do treinamento.

2.6 Avaliação dos Modelos

O desempenho dos modelos foi avaliado utilizando as métricas de acurácia, precisão, *recall*, especificidade e F1-Score. Essas métricas permitem analisar o desempenho do classificador considerando diferentes aspectos de classificação binária.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(b)}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{(c)}$$

$$\text{Recall (sensibilidade)} = \frac{TP}{TP + FN} \quad \text{(d)}$$

$$\text{Especificidade} = \frac{TN}{TN + FP} \quad \text{(e)}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{(f)}$$

Onde TP, TN, FP e FN representam verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respectivamente [15]. Em todas as métricas, a classe nódulo foi considerada a classe positiva.

Tabela 1. Métricas no conjunto de teste por experimento

Experimento	Acurácia	Sensibilidade	Especificidade	F1-Score
ResNet-50 + Adam	0,9553	0,9330	0,9777	0,9543
ResNet-50 + SGD	0,9441	0,9330	0,9553	0,9435
ConvNeXt-Tiny + Adam	0,9497	0,9385	0,9609	0,9432
ConvNeXt-Tiny + SGD	0,5782	0,2793	0,8771	0,3984
MobileNetV3-Large + Adam	0,8464	0,7654	0,9274	0,8328
MobileNetV3-Large + SGD	0,9330	0,8827	0,9832	0,9294
ViT-B/16 + Adam	0,8520	0,7765	0,9274	0,8399
ViT-B/16 + SGD	0,5391	0,2291	0,8492	0,3320
Swin-Tiny + Adam	0,9693	0,9609	0,9777	0,9690
Swin-Tiny + SGD	0,7989	0,6983	0,8994	0,7764

Fonte: Elaborado pelos autores (2026).

2.7 Análise de Explicabilidade com Grad-CAM

Para ajudar na compreensão das decisões das redes neurais, foi utilizada a técnica Grad-CAM [16]. Essa abordagem produz mapas de ativação que destacam as regiões da imagem que contribuíram de forma mais significativa para a decisão do modelo.

Os mapas de ativação foram gerados a partir da última camada convolucional das redes, permitindo avaliar qualitativamente se as regiões identificadas pelas CNNs coincidem com áreas clinicamente relevantes nas imagens de tomografia.

3 RESULTADOS

Esta seção apresenta os resultados dos 10 experimentos realizados com o objetivo de avaliar o impacto de diferentes arquiteturas de aprendizado profundo e otimizadores na tarefa de classificação binária de nódulos pulmonares em imagens de TC representadas no formato 2.5D. Para isso, são analisadas as curvas de aprendizado ao longo de 50 épocas, juntamente com as métricas finais de desempenho obtidas no conjunto de teste (acurácia, sensibilidade, especificidade e F1-Score), em uma discussão comparativa entre os grupos de arquiteturas avaliados. A Tabela 1 apresenta o resumo consolidado dos resultados, evidenciando diferenças relevantes tanto entre arquiteturas quanto entre otimizadores.

3.1 Análise por Arquitetura

ResNet-50: A rede ResNet-50 apresentou comportamento consistente e estável nos dois otimizadores, com boa convergência durante o treinamento; o melhor *checkpoint* foi selecionado na época 23 com o otimizador Adam e na época 40 com o otimizador SGD. Com o Adam, atingiu F1 = 0,9543 e sensibilidade de 0,9330 no conjunto de teste, configurando-se como uma das *baselines* mais sólidas do experimento. A curva de treinamento revela boa generalização do experimento, com *loss* de validação acompanhando a de treinamento nas épocas iniciais e estabilizando em patamar aceitável. A matriz de confusão do teste confirma apenas 12 falsos negativos (6,70%) em 179 nódulos reais, desempenho clinicamente relevante.

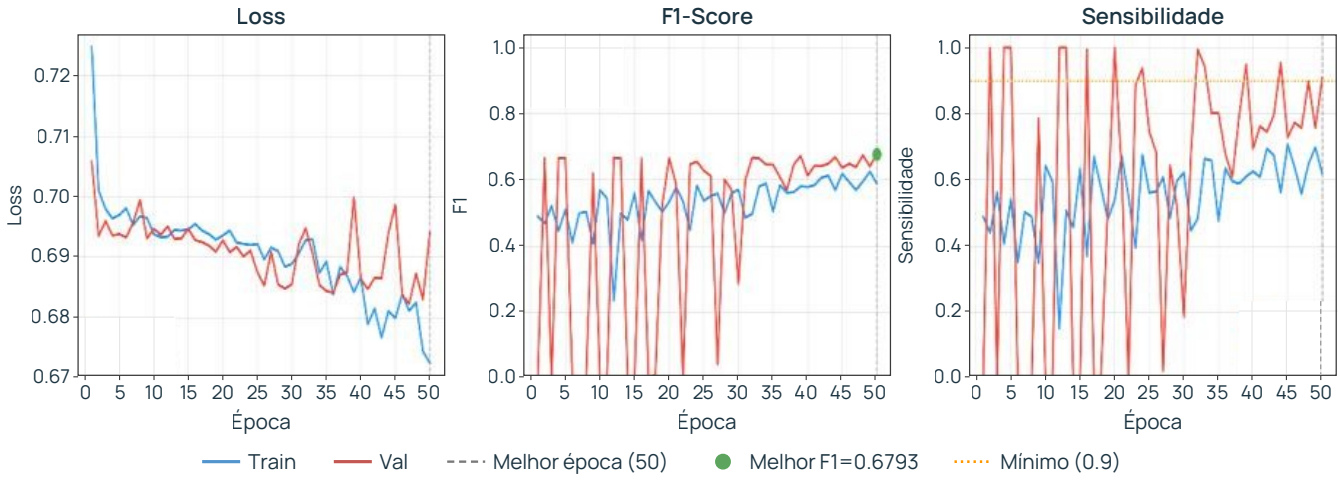
ConvNeXt-Tiny: A combinação ConvNeXt-Tiny + Adam apresentou convergência acelerada a partir da época 5. No teste, obteve F1 = 0,9492 e sensibilidade 0,9385. Em contraste marcante, a combinação com o otimizador SGD resultou em falha de convergência, conforme evidenciado pela Figura 1, que exibe a *loss* de validação estagnada próxima a 0,69 ao longo de quase todo o treinamento com oscilações severas das métricas de validação, e sensibilidade de apenas 0,28 no conjunto de teste (129 nódulos não detectados), indicando que o modelo apresentou um viés para classe normal sem discriminação real.

Esse comportamento sugere uma sensibilidade do ConvNeXt à escolha do otimizador, possivelmente relacionada às suas camadas LayerNorm, que tipicamente respondem melhor a otimizadores adaptativos como o Adam.

MobileNetV3-Large: A rede MobileNetV3-Large apresentou o comportamento mais sensível ao otimizador entre as CNNs avaliadas. Com o otimizador Adam, o modelo não atendeu à restrição de seleção descrita na Seção 2.5 em 36 das 50 épocas, com validação exibindo uma lacuna persistente de generalização. O *checkpoint* salvo correspondeu à época 9. No teste, a sensibilidade foi apenas 0,7654 (42 falsos negativos), com especificidade de 0,9274, indicando que a arquitetura deixou de detectar uma parcela relevante dos nódulos. Com o otimizador SGD, a dinâmica foi o oposto: o modelo convergiu progressivamente, com *checkpoint* selecionado na época 18 e, no conjunto de teste, sensibilidade de 0,8827 (21 falsos negativos) e especificidade de 0,9832, a maior de todo o experimento, com apenas 3 falsos positivos. Esse resultado posiciona a rede MobileNetV3 + SGD maior de todo o experimento, com apenas 3 falsos positivos. Esse resultado posiciona a rede MobileNetV3 + SGD como a configuração com melhor controle de falsos positivos

aliado à eficiência computacional. A Figura 2 reforça esse comportamento ao evidenciar estabilidade do treinamento e a confiabilidade de desempenho do

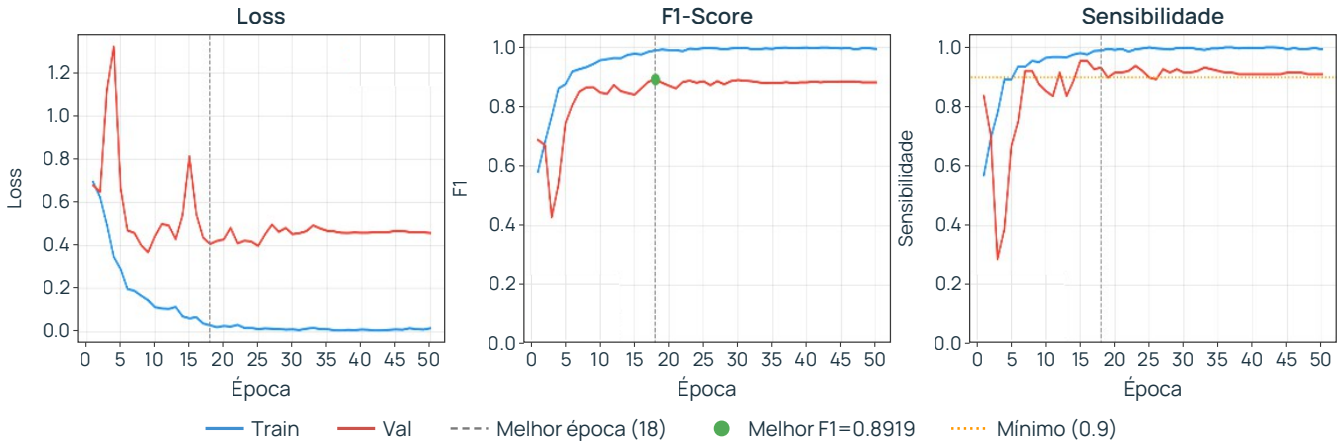
Figura 1. Evolução da loss, F1-Score e sensibilidade da rede ConvNeXt-Tiny + SGD



As métricas de validação desta figura foram calculadas com a classe normal como positiva (Seção 2.5); a curva de sensibilidade de validação corresponde à especificidade para a classe nódulo.

Fonte: Elaborado pelos autores (2026).

Figura 2. Curvas de treinamento da MobileNetV3-Large + SGD



As métricas de validação desta figura foram calculadas com a classe normal como positiva (Seção 2.5); a curva de sensibilidade de validação corresponde à especificidade para a classe nódulo.

Fonte: Elaborado pelos autores (2026).

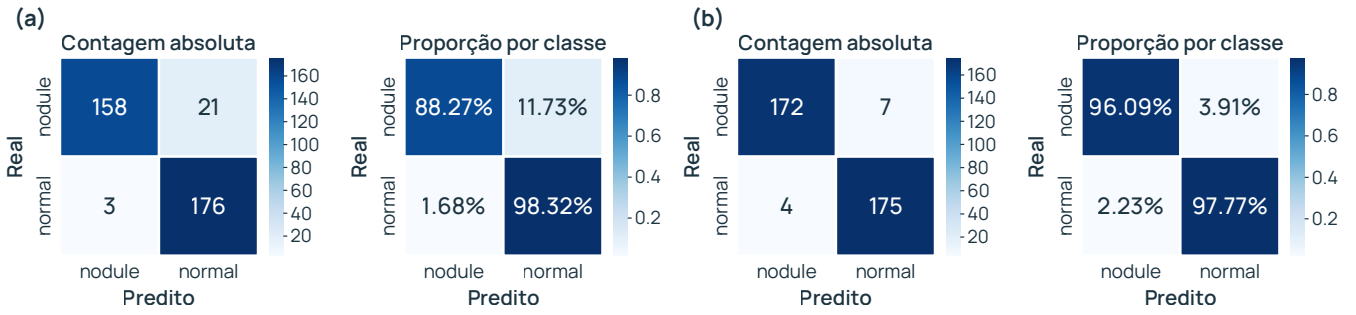
modelo com o otimizador SGD.

As curvas de aprendizado mostram uma convergência progressiva, com a perda de treinamento decaindo rapidamente para valores próximos a zero, enquanto a perda de validação se estabiliza sem apresentar aumento significativo, indicando ausência de *overfitting*. De modo complementar, as métricas de validação exibem um crescimento consistente ao longo das épocas, com pequena lacuna entre treino e validação, sugerindo boa capacidade de generalização.

A matriz de confusão no conjunto de teste confirma esses achados, evidenciando elevado desempenho discriminativo, conforme apresentado na Figura 3a.

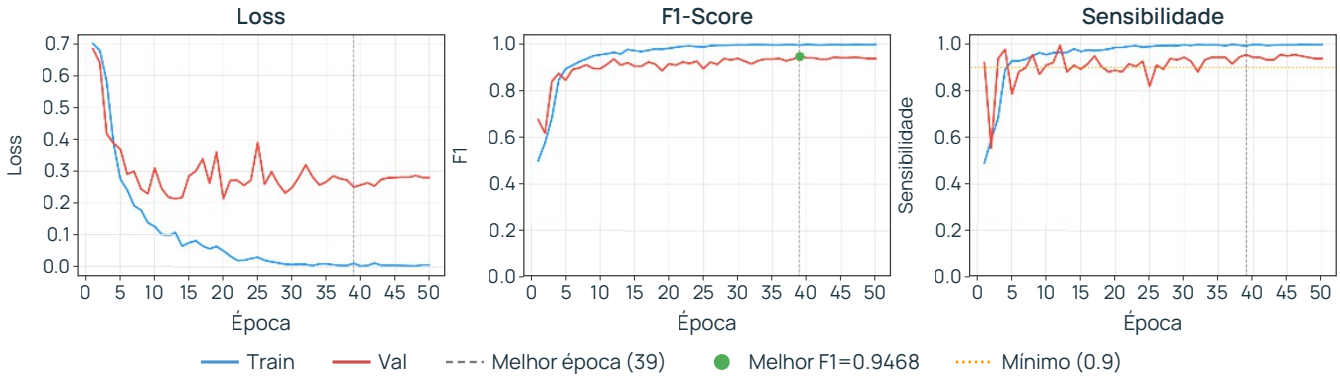
O modelo classificou corretamente 158 nódulos e 176 exames normais, com apenas 21 falsos negativos e 3 falsos positivos, resultando em sensibilidade de 0,8827 e especificidade de 0,9832. Esses resultados demonstram que o modelo mantém excelente controle sobre falsos alarmes, característica relevante em aplicações clínicas, embora 21 nódulos (11,73%) não tenham sido detectados. Em conjunto, as evidências indicam que a utilização do otimizador SGD proporcionou uma otimização mais estável e generalizável para a rede MobileNetV3-Large, consolidando seu desempenho superior ao uso do Adam neste estudo.

Figura 3. Matriz de confusão do teste (a) MobileNetV3-Large + SGD; (b) Swin-Tiny + Adam



Fonte: Elaborado pelos autores (2026).

Figura 4. Curvas de treinamento do Swin-Tiny + Adam



As métricas de validação desta figura foram calculadas com a classe normal como positiva (Seção 2.5); a curva de sensibilidade de validação corresponde à especificidade para a classe nódulo.

Fonte: Elaborado pelos autores (2026).

ViT-B/16: O ViT-B/16 foi a arquitetura com pior desempenho geral e maior tempo de treinamento (aproximadamente 118 min com o otimizador Adam e 119 min com o otimizador SGD). Com o Adam, a convergência foi lenta e gradual; o *checkpoint* foi selecionado apenas na época 40, com alta instabilidade nas primeiras 20 épocas, refletindo a conhecida dificuldade do ViT puro em conjuntos de dados pequenos. A matriz de confusão do teste revelou 40 falsos negativos (22,35%), o que é clinicamente inaceitável para rastreamento de nódulos. Com o otimizador SGD, o modelo não convergiu, oscilando aleatoriamente em torno de 50% de acurácia ao longo de todo o treinamento; a *loss* de validação estabilizou próxima a 0,69 com sensibilidade de apenas 0,2291 no teste. Esses resultados confirmam que o ViT-B/16 teve um desempenho subótimo no contexto dos nossos testes, o que pode sugerir uma inadequação da arquitetura ao volume de dados do teste, por exemplo.

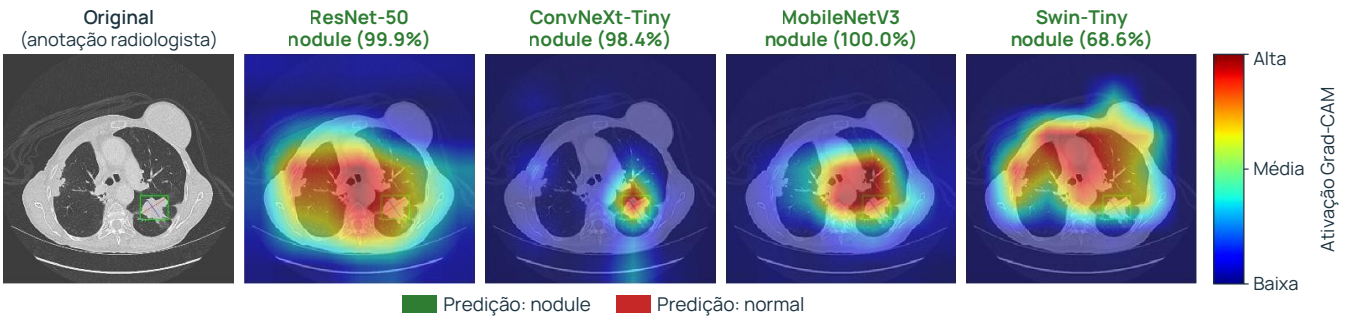
Swin-Tiny: A rede Swin-Tiny + Adam foi o melhor modelo individual do experimento, atingindo F1 = 0,9690, acurácia de 0,9693, sensibilidade de 0,9609 e especificidade de 0,9777. As curvas de treinamento, apresentadas na Figura 4, exibem convergência mais suave e estável ao longo das 50 épocas com as métricas de validação crescendo consistentemente até a seleção do *checkpoint* na época 39.

A matriz de confusão do teste, ilustrada na Figura 3b, revela apenas 7 falsos negativos (3,91%) e 4 falsos positivos (2,23%), resultado notavelmente equilibrado entre as duas classes e o mais próximo de um desempenho clinicamente ideal entre todos os experimentos. O desempenho superior do Swin-Tiny em relação à rede ViT-B/16 confirma que a injeção de hierarquia e localidade via janelas deslocadas (*shifted windows*) é essencial para viabilizar Transformers em conjuntos de dados médicos de pequeno porte. Com o SGD, o Swin-Tiny também sofreu convergência lenta e instável, atingindo F1 = 0,7764 com alta oscilação ao longo do treinamento.

3.2 Análise Qualitativa com Grad-CAM

O ViT-B/16 foi excluído desta análise qualitativa em razão do desempenho insatisfatório observado durante o treinamento. Mesmo com o otimizador Adam, o modelo registrou 40 falsos negativos (22,35%) no conjunto de teste, índice clinicamente inferior para rastreamento de nódulos pulmonares. Sua inclusão nos mapas Grad-CAM não acrescentaria valor interpretativo relevante.

Figura 5. Mapas de ativação Grad-CAM, caso 000223_05_03_105



Fonte: Elaborado pelos autores (2026).

Para os quatro modelos restantes, foram gerados mapas de ativação Grad-CAM sobre amostras do conjunto IQ-OTH/NCCD [17], utilizado como validação externa. Essas imagens não participaram do treinamento e validação das redes, ou seja, os modelos não viram essas imagens anteriormente. Em cada painel, a imagem original apresenta a anotação do radiologista (retângulo verde), seguida pelos mapas de cada arquitetura.

Caso 000223_05_03_105 (nódulo confirmado, Figura 5): caso com nódulo único de localização periférica no hemitórax esquerdo. Todos os modelos classificaram corretamente, com probabilidades preditas de 99,9% (ResNet-50), 98,4% (ConvNeXt-Tiny), 100% (MobileNetV3) e 68,6% (Swin-Tiny). O ConvNeXt-Tiny produziu o mapa mais preciso do painel, com núcleo de alta ativação diretamente sobre a lesão anotada. A ResNet-50 apresentou ativação coerente, porém com dispersão para a região mediastinal. A MobileNetV3 exibiu ativação ampla cobrindo o hemitórax esquerdo sem concentração clara sobre o nódulo. O Swin-Tiny, com a menor probabilidade predita do grupo (68,6%), apresentou ativação difusa sem discriminação espacial precisa sobre a lesão.

Considerando o caso em conjunto, o ConvNeXt-Tiny destacou-se pela maior coerência entre os mapas de ativação e as anotações dos radiologistas, produzindo o foco mais localizado sobre a lesão mesmo com menores probabilidades preditas. A ResNet-50 apresentou comportamento intermediário, com mapa coerente, porém com maior dispersão. A MobileNetV3, apesar das altas probabilidades preditas, exibiu ativações espacialmente abrangentes sem foco preciso sobre a lesão, limitação relevante para aplicações clínicas que exigem rastreabilidade da decisão. O Swin-Tiny, melhor modelo em F1-Score, apresentou ativação consistentemente difusa, sugerindo que seu desempenho quantitativo superior deriva de características globais da imagem não associadas diretamente à região da lesão. Esses achados reforçam a importância de complementar métricas quantitativas com análises de interpretabilidade em diagnóstico assistido por computador, onde a confiabilidade do processo de decisão é tão relevante quanto a acurácia do modelo.

4 DISCUSSÃO

O desempenho superior do Swin-Tiny (F1 = 0,9690) frente ao ViT-B/16 evidencia que a introdução de hierarquia e localidade via janelas deslocadas é determinante para viabilizar Transformers em conjuntos de dados médicos de porte moderado, embora a diferença em relação à ResNet-50 e à ConvNeXt-Tiny + Adam seja pequena (F1 de 0,9543 e 0,9492, respectivamente) e não tenha sido avaliada estatisticamente. O ViT puro, ao aplicar atenção global sem viés indutivo espacial, tende a exigir volumes de dados maiores para generalizar, o que explicaria os 40 falsos negativos (22,35%) registrados no conjunto de teste.

O impacto do otimizador revelou-se crítico quanto à escolha da arquitetura. O Adam favoreceu arquiteturas com camadas LayerNorm (Swin-Tiny, ConvNeXt-Tiny), cujos gradientes são mais sensíveis à escala adaptativa, enquanto o SGD proporcionou convergência mais estável para a MobileNetV3-Large, rede que combina convoluções *depthwise* e atenção por canal, respondendo melhor a uma taxa de aprendizado fixa com *momentum*. A falha completa de convergência do ConvNeXt-Tiny + SGD (sensibilidade de 0,28) e do ViT-B/16 + SGD reforça que a compatibilidade entre otimizador e normalização interna da rede deve ser tratada como hiperparâmetro crítico em estudos futuros.

A dissociação entre F1-Score elevado e coerência dos mapas Grad-CAM é um achado clinicamente relevante. O Swin-Tiny, melhor modelo em métricas quantitativas, apresentou ativações consistentemente difusas na validação externa (IQ-OTH/NCCD), sugerindo que seu desempenho pode derivar de características globais da imagem não diretamente associadas ao nódulo; em contraste, o ConvNeXt-Tiny produziu os focos mais precisos sobre as lesões anotadas pelos radiologistas, mesmo com probabilidades preditas inferiores. Em comparação com trabalhos recentes no LUNA16, o Swin-Tiny com Adam superou a arquitetura híbrida Cross-ViT proposta por Zhu e Fei [18], que combina CNN e *Vision Transformer* por meio de atenção cruzada.

Enquanto a Cross-ViT atingiu F1 de 92,96%, o Swin-Tiny alcançou F1 de 96,90%, sem recorrer à fusão explícita

de múltiplas *branches*. Contudo, enquanto aquele estudo não realizou validação externa, a análise Grad-CAM no IQ-OTH/NCCD evidenciou limitações de coerência espacial que métricas isoladas não capturam, reforçando que comparações baseadas exclusivamente em *benchmarks* internos podem superestimar a capacidade de generalização em dados não vistos, aspecto especialmente crítico em aplicações clínicas reais.

No contexto do SUS, a MobileNetV3-Large com SGD apresenta-se como alternativa concreta para ambientes com recursos computacionais limitados: além do menor custo de treinamento (~30 minutos), sua alta especificidade (0,9832) reduz falsos alarmes e exames confirmatórios desnecessários. No entanto, sua sensibilidade inferior (0,8827), com 21 nódulos não detectados no teste, limita seu uso isolado em triagem oncológica, onde a minimização de falsos negativos é prioritária; somada às ativações Grad-CAM espacialmente amplas, essa limitação aponta para a necessidade de validação clínica prospectiva antes de qualquer implantação assistencial real.

5 CONCLUSÕES

Este trabalho avaliou cinco arquiteturas de aprendizado profundo combinadas a dois otimizadores para classificação binária de nódulos pulmonares em representação 2.5D. O Swin-Tiny com Adam obteve o melhor desempenho global (F1 = 0,9690), enquanto a MobileNetV3-Large com SGD destacou-se pela maior especificidade (0,9832) e menor custo computacional, sendo uma alternativa para ambientes com recursos limitados, ainda com sensibilidade inferior (0,8827). A análise Grad-CAM evidenciou que métricas elevadas não garantem coerência de ativação clínica, reforçando a importância da explicabilidade em diagnóstico assistido. O ViT-B/16 mostrou-se inadequado para o volume de dados utilizado sem estratégias adicionais de regularização. Como trabalhos futuros, propõe-se ampliar o conjunto de dados para representação de populações brasileiras, investigar pré-treinamento específico em imagens médicas para o ViT-B/16 e conduzir validação clínica prospectiva com radiologistas no contexto do SUS.

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.

Financiamento. Os autores não declararam fonte de financiamento.

Disponibilidade de dados e materiais. Os autores não declararam a disponibilidade dos dados e materiais da pesquisa.

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

Declaração de uso de IA. Os autores declaram que ferramentas de inteligência artificial generativa (Claude IA e Google Gemini) foram utilizadas como auxílio na revisão gramatical e estruturação textual de partes deste manuscrito. Todo o conteúdo científico, incluindo concepção do estudo, metodologia, análise de dados e interpretação dos resultados, é de inteira responsabilidade dos autores.

REFERÊNCIAS

1. Instituto Nacional de Câncer. O que é câncer? [Internet]. Rio de Janeiro: INCA; 2020 [citado 2026 abr 15]. Disponível em: https://www.inca.gov.br/o-que-e-cancer
2. International Agency for Research on Cancer. Lung cancer [Internet]. Lyon: IARC; 2025 [citado 2026 abr 15]. Disponível em: https://www.iarc.who.int/cancer-type/lung-cancer/
3. Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2026: incidência de câncer no Brasil [Internet]. Rio de Janeiro: INCA; 2026 [citado 2026 abr 15]. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa
4. Cancer Research UK. Cancer statistics for the UK [Internet]. London: Cancer Research UK; 2018 [citado 2026 abr 18]. Disponível em: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk
5. Chaturvedi P, Jhamb A, Vanani M, Nemade V. Prediction and Classification of Lung Cancer Using Machine Learning Techniques. IOP Conf. Ser.: Mater. Sci. Eng. 2021;1099(1): 012059. doi: https://doi.org/10.1088/1757-899x/1099/1/012059
6. Lee JH, Sun HY, Park S, Kim H, Hwang EJ, Goo JM, et al. Performance of a Deep Learning Algorithm Compared with Radiologic Interpretation for Lung Cancer Detection on Chest Radiographs in a Health Screening Population. Radiology. 2020;297(3):687-96. doi: https://doi.org/10.1148/radiol.2020201240
7. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. J Big Data. 2019;6(1):60. doi: https://doi.org/10.1186/s40537-019-0197-0
8. Van Ginneken B, Jacobs C. LUNA16 Part 1/2 [Internet]. Geneva: Zenodo; 2019 [citado 2026 abr 25]. Disponível em: https://doi.org/10.5281/zenodo.3723295
9. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging. 2016;35(5): 1285-98. doi: https://doi.org/10.1109/TMI.2016.2528162
10. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [s.l.]: IEEE; 2016. p. 770-8. doi: https://doi.org/10.1109/cvpr.2016.90

11. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [s.l.]: IEEE; 2022. p. 11966-76. doi: https://doi.org/10.1109/cvpr52688.2022.01167

12. Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, et al. Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). [s.l.]: IEEE; 2019. p. 1314-24. doi: https://doi.org/10.1109/iccv.2019.00140

13. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR); 2021; [virtual]. [S.l.: s.n.]; 2021.

14. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). [s.l.]: IEEE; 2021. p. 9992-10002. doi: https://doi.org/10.1109/iccv48922.2021.00986

15. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies. 2011;2(1):37-63.

16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). [s.l.]: IEEE; 2017. p. 618-26. doi: https://doi.org/10.1109/iccv.2017.74

17. Alyasriy H, Al-Huseiny M. The IQ-OTH/NCCD lung cancer dataset [Internet]. London: Mendeley Data; 2023 [citado 2026 abr 27]. Versão 4. Disponível em: https://doi.org/10.17632/bhmdr45bh2.4

18. Zhu Q, Fei L. Cross-ViT based benign and malignant classification of pulmonary nodules. PLoS One. 2025;20(2):e0318670. doi: https://doi.org/10.1371/journal.pone.0318670

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