Journal of Health Informatics

J Health Inform. 2026;18:e1851
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1851

ARTIGO COMPLETO
CBIS 2026

Detecção de massas mamárias estratificada por tamanho em mamografia: estudo comparativo entre YOLOv11m e RT-DETR-L

Size-stratified detection of breast masses in mammography: a comparative study of YOLOv11m and RT-DETR-L

Erick Keven da Silva Alves[1] ✉, Wellington Cezario Dos Santos[1], Sandro Gomes Paulino[1], Marcelo Costa Oliveira[1]

ORCID: 0009-0004-2571-4349 · 0009-0002-7241-0271 · 0009-0002-2273-4518 · 0000-0002-0825-6217

[1] Instituto de Computação, Universidade Federal de Alagoas (UFAL), Maceió (AL), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1851

DOI
10.59681/2175-4411.v18.iEspecial.2026.1851

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
Erick Keven da Silva Alves
eksa@ic.ufal.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: Comparar YOLOv11m (CNN one-stage) e RT-DETR-L (Transformer) na detecção de massas mamárias, estratificando o desempenho por tamanho de lesão, dimensão pouco explorada e relevante para o rastreamento no Sistema Único de Saúde (SUS), que tem déficits de cobertura. **Método:** Combinaram-se três datasets públicos (CDD-CESM low-energy, INbreast, VinDr-Mammo): 2.341 massas em 1.713 imagens. Faixas definidas por tercis empíricos (19,4 e 36,3 mm). Seis experimentos avaliados por mAP@0,5, FROC, Wilcoxon pareado e McNemar (correção de Bonferroni). **Resultados:** O RT-DETR-L obteve Recall superior nas três faixas ($p < 0{,}003$) e maior taxa de detecção por imagem ($p < 10^{-5}$). O YOLOv11m obteve F1 superior na faixa pequena ($p = 1{,}30 \times 10^{-4}$). O SAHI foi consistente com origem arquitetural das diferenças. **Conclusão:** As arquiteturas são complementares conforme o tamanho da lesão, sustentando a hipótese de estratégia híbrida de requisitos computacionais modestos, ainda pré-clínica e a validar com dados brasileiros.

Descritores: Mamografia; Diagnóstico por Computador; Aprendizado Profundo.

ABSTRACT

Objective: To compare YOLOv11m (one-stage CNN) and RT-DETR-L (Transformer) for breast mass detection, stratifying performance by lesion size, a dimension rarely assessed and relevant to screening in Brazil's public health system (SUS), which faces coverage gaps. **Method:** Three public datasets were combined (CDD-CESM low-energy, INbreast, VinDr-Mammo): 2,341 masses in 1,713 images. Bins defined by empirical terciles (19.4 and 36.3 mm). Six experiments evaluated with mAP@0.5, FROC, paired Wilcoxon and McNemar (Bonferroni correction). **Results:** RT-DETR-L achieved higher Recall in all three bins ($p < 0.003$) and a higher per-image detection rate ($p < 10^{-5}$). YOLOv11m achieved higher F1 in the small bin ($p = 1.30 \times 10^{-4}$). SAHI was consistent with an architectural origin of the differences. **Conclusion:** The architectures are complementary according to lesion size, supporting the hypothesis of a hybrid strategy with modest computational requirements, which remains pre-clinical and must be validated with Brazilian data.

Keywords: Mammography; Diagnosis, Computer-Assisted; Deep Learning.

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: Alves EKS, Dos Santos WC, Paulino SG, Oliveira MC. Detecção de massas mamárias estratificada por tamanho em mamografia: estudo comparativo entre YOLOv11m e RT-DETR-L. J Health Inform 2026;18:e1851. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1851

1 INTRODUÇÃO

O câncer de mama é a neoplasia maligna mais incidente entre mulheres, tanto no Brasil quanto no mundo. Para cada ano do triênio 2023–2025, o Instituto Nacional de Câncer (INCA) estimou 73.610 novos casos no país, correspondentes a uma taxa bruta de 66,54 casos por 100.000 mulheres [1]. O rastreamento mamográfico periódico é uma das principais estratégias de detecção precoce da doença: quando identificado em estágio localizado, o câncer de mama apresenta sobrevida relativa em cinco anos superior a 95% [2]. Entretanto, persistem desigualdades no acesso ao exame de mamografia no Brasil. Segundo a Pesquisa Nacional de Saúde de 2019, cujos dados são divulgados pelo INCA [3], 58,3% das mulheres de 50 a 69 anos haviam realizado mamografia nos dois anos anteriores à pesquisa. Essa proporção foi maior nas áreas urbanas do que nas rurais (60,5% e 41,6%, respectivamente) e variou conforme o nível de escolaridade, de 49,0% entre mulheres sem instrução ou com ensino fundamental incompleto a 77,8% entre aquelas com ensino superior completo. Esse cenário é agravado pela distribuição desigual dos radiologistas no território nacional: embora o Brasil conte com 20.453 especialistas certificados em Radiologia e Diagnóstico por Imagem, a densidade varia de 13,18 profissionais por 100.000 habitantes na região Sudeste a menos de 5 na região Norte e menos de 8 na região Nordeste. A concentração desses especialistas nas capitais é, ainda, quase quatro vezes maior do que nos municípios de menor porte [4].

A dificuldade em detectar a doença varia marcadamente com o tamanho da lesão: massas pequenas são desproporcionalmente perdidas no rastreamento, enquanto massas maiores são mais conspícuas, porém ainda sujeitas a erros de localização em parênquima denso. A maioria dos estudos de *Computer-Aided Detection* (CAD) baseados em aprendizado profundo avalia o desempenho de detectores de forma agregada, sem estratificação por tamanho. Ribli *et al.* [5] demonstraram que uma variante da *Faster* R-CNN superou o CAD comercial em dois *datasets* públicos. Rodriguez-Ruiz *et al.* [6] mostraram que um sistema de inteligência artificial (IA) autônomo atingiu sensibilidade equivalente à média de 101 radiologistas. McKinney *et al.* [7] demonstraram que sistemas de IA reduzem falsos negativos e aliviam a carga de dupla leitura. Apesar desses avanços, comparações controladas entre paradigmas arquiteturais distintos estratificadas por tamanho de lesão permanecem escassas na literatura.

A detecção de objetos em visão computacional é atualmente dominada por dois paradigmas. As redes neurais convolucionais (CNNs) de estágio único, representadas pela família *You Only Look Once* (YOLO), realizam extração hierárquica de características com âncoras locais [8]. Os detectores baseados em Transformers, introduzidos pelo *DEtection TRansformer* (DETR) [9], modelam dependências globais via *self-attention*, eliminando propostas manuais de região. O *Real-Time DEtection TRansformer* (RT-DETR) combina codificador híbrido CNN-Transformer com decodificador de consultas aprendíveis [10], alcançando desempenho competitivo em tempo real. As propriedades arquiteturais distintas (campos receptivos locais *versus* atenção global) sugerem comportamentos diferenciados conforme o tamanho do objeto-alvo, hipótese sistematicamente pouco investigada em imagem mamográfica.

A principal novidade deste trabalho reside na comparação sistemática entre paradigmas arquiteturais CNN e Transformer especificamente estratificada por faixas de tamanho de lesão calibradas em milímetros, lacuna recorrente na literatura de CAD mamográfico. Ao adotar tercis empíricos alinhados ao estadiamento *Tumour, Node, Metastasis* (TNM) e ao *Breast Imaging Reporting and Data System* (BI-RADS) como critério de estratificação, este trabalho aproxima a avaliação técnica da taxonomia clínica, permitindo identificar em qual regime de tamanho cada arquitetura é mais vantajosa. Complementarmente, propõe-se um *framework* de validação estatística formal (Wilcoxon pareado e McNemar com correção de Bonferroni) que vai além das métricas de *Average Precision* (AP) agregadas, uma análise *post hoc* com *Slicing Aided Hyper Inference* (SAHI) para isolar efeitos arquiteturais de efeitos de resolução, e uma análise de IA Explicável via Grad-CAM e mapas de atenção cruzada para interpretação mecanicista das diferenças observadas.

Para a Informática em Saúde, decorrem daí dois conhecimentos de alcance mais geral: a *mean Average Precision* (mAP) agregada, métrica predominante nos relatos de CAD, mostra-se insuficiente como critério isolado de avaliação e de aquisição desses sistemas; e, por a vantagem arquitetural depender da escala da lesão, sistemas de apoio ao diagnóstico em mamografia tendem a se beneficiar de concepções em cascata ou *ensemble*, com componentes especializados por faixa de tamanho, em vez de um modelo único. O próprio protocolo aqui empregado constitui, ainda, um *framework* validatório reutilizável por outros detectores e por futuras validações com dados nacionais.

Diante deste contexto, o objetivo principal deste trabalho foi comparar sistematicamente YOLOv11m e RT-DETR-L para detecção de massas mamárias em três faixas de tamanho empiricamente definidas, quantificando as diferenças arquiteturais com rigor estatístico e discutindo requisitos e limites para uma eventual implantação de estratégias híbridas em serviços de radiologia do SUS.

2 MÉTODOS

2.1 Datasets

Foram utilizados três *datasets* públicos de mamografia (Tabela 1): dois de *Full-Field Digital Mammography* (FFDM) e um de *Contrast-Enhanced Spectral Mammography* (CESM). Critério de inclusão uniforme: imagens com anotação de massa com tamanho computável em milímetros. Foram excluídas lesões não classificadas como massa (calcificações, assimetrias, distorções arquitetônicas).

Tabela 1. Características dos datasets utilizados.

Dataset	Modalidade	Origem	Espaçamento (mm/px)	Ref.
INbreast	FFDM	Portugal	DICOM (PixelSpacing)	[11]
CDD-CESM (LE)	CESM *low-energy*	Egito	*Fallback* 0,094	[12]
VinDr-Mammo	FFDM	Vietnã	metadata.csv ou 0,085	[13]

Fonte: Elaborado pelos autores (2026).

Para o INbreast, as massas foram extraídas dos arquivos XML e polígonos convertidos a retângulos envolventes mínimos. Para o CDD-CESM, utilizaram-se exclusivamente as imagens *low-energy* (LE), excluindo as recombinadas; segmentações manuais foram convertidas a *bounding boxes*. Para o VinDr-Mammo, apenas achados categorizados como *"Mass"* foram incluídos; registros *"No Finding"* foram excluídos. A coorte combinada totalizou **2.341 anotações de massa em 1.713 imagens** (CDD-CESM LE: 1.000; INbreast: 116; VinDr-Mammo: 1.225).

2.2 Pré-processamento

O *pipeline* foi aplicado identicamente a todos os *datasets*: (1) leitura DICOM com *pydicom*, aplicação de RescaleSlope/RescaleIntercept e inversão de MONOCHROME1; (2) normalização para uint8 [0–255] por percentil; (3) cálculo do tamanho em milímetros pela equação (a); (4) segmentação da região mamária por limiarização binária (limiar = 10) com margem de 10 px; (5) *Contrast Limited Adaptive Histogram Equalization* (CLAHE; grade 8 × 8, clip = 2,0); e (6) redimensionamento para 640×640 px com interpolação bilinear.

$$s\text{mm} = \max(w\text{px}, h\text{px}) \times d\text{mm/px} \quad \text{(a)}$$

2.3 Estratificação por Tamanho

As faixas foram definidas por **tercis empíricos (P33 = 19,4 mm; P67 = 36,3 mm)**, calculados sobre a coorte combinada antes da partição por paciente, garantindo contagens equilibradas e poder estatístico comparável entre estratos: ***small*** [0; 19,4) mm, ***medium*** [19,4; 36,3) mm e ***large*** [36,3; 150] mm. Esses limites coincidem com limiares do estadiamento TNM (T1c/T2 em 20 mm) e do sistema BI-RADS [2,14], conferindo interpretabilidade clínica.

2.4 Partição dos Dados

Divisão em treino (70%), validação (15%) e teste (15%) ao nível do paciente via GroupShuffleSplit, com ausência de sobreposição verificada formalmente (treino ∩ teste = ∅; *seed* = 42). Imagens com múltiplas massas foram atribuídas à faixa da maior lesão.

2.5 Modelos e Treinamento

Foram comparados **YOLOv11m** [8] e **RT-DETR-L** [10], ambos via Ultralytics com pesos pré-treinados no COCO e *fine-tuning* nos dados de mamografia. Os parâmetros de treinamento são apresentados na Tabela 2.

Aumentação idêntica: reflexão horizontal (prob. 0,5); reflexão vertical desativada (flipud = 0,0), pois a inversão craniocaudal não é anatomicamente equivalente; rotação ±10°; escala ±10%; *mosaic* desativado (*mosaic* = 0,0), preservando a escala real das lesões. Treinamentos independentes por faixa, totalizando seis experimentos; o *batch size* igualado a 8 elimina essa variável como confundidora.

Tabela 2. Parâmetros de treinamento dos modelos.

Parâmetro	YOLOv11m	RT-DETR-L
Peso base	yolo11m.pt	rtdetr-l.pt
Otimizador	SGD	AdamW
Taxa de aprendizado	0,01	0,0001
Weight decay	5×10^{-4}	1×10^{-4}
Warmup (épocas)	3	3
Batch size	8	8
Épocas máximas / paciência	100 / 20	100 / 20
Tamanho de entrada	640 × 640	640 × 640
Semente	42	42

Fonte: Elaborado pelos autores (2026).

2.6 Métricas, Análise Estatística e Post hoc

As métricas mAP@0,5, mAP@0,5:0,95, Precisão, *Recall* e F1 são reportadas no **conjunto de teste** (15%), via best.pt selecionado por mAP@0,5 de validação (conf = 0,25). A curva *Free-Response Receiver Operating Characteristic* (FROC) foi calculada com *Intersection over Union* (IoU) ≥ 0,3 (limiar mais permissivo que o padrão COCO, justificado pela resolução efetiva reduzida após redimensionamento para 640 px), em pontos 0,125 a 2,0 falsos positivos (FP) por imagem. Testes de **Wilcoxon pareados** foram aplicados a quatro métricas por faixa (12 testes); testes de **McNemar** avaliaram a detecção binária por faixa (3 testes). Correção de **Bonferroni**: $\alpha' \approx 0{,}0042$ (Wilcoxon) e $\alpha' \approx 0{,}0167$ (McNemar). O SAHI [15] (*slice*: 320×320 px; *overlap*: 0,2; *Non-Maximum Suppression*

(NMS) IoU: 0,5) foi aplicado *post hoc* na faixa pequena para isolar efeitos arquiteturais de efeitos de resolução.

2.7 Aspectos Éticos e Infraestrutura

Esta pesquisa utilizou exclusivamente *datasets* públicos anonimizados, disponibilizados pelos respectivos autores para fins científicos, sem envolvimento de coleta direta com seres humanos. A aprovação por Comitê de Ética em Pesquisa (CEP) foi dispensada com fundamento na Resolução nº 510/2016 do Conselho Nacional de Saúde [16], que exclui da exigência de apreciação ética pesquisas realizadas com informações de acesso público cujos dados não permitam identificar os sujeitos. Os experimentos foram conduzidos em GPU NVIDIA RTX 3070 (8 GB VRAM), Python 3.11 e Ultralytics; reprodutibilidade garantida por *seed* = 42 global, cudnn.deterministic = True, benchmark = False.

3 RESULTADOS

Todos os seis experimentos convergiram antes de 100 épocas via *early stopping* (paciência = 20). Notavelmente, o RT-DETR-L na faixa pequena convergiu já na época 6, com mAP@0,5 de validação de 0,227, contra a época 43 do YOLOv11m. Essa convergência prematura indica que o Transformer saturou o sinal disponível nas imagens com massas pequenas antes de extrair representações discriminativas suficientes, antecipando o padrão de desempenho detalhado a seguir.

A Tabela 3 apresenta as métricas no conjunto de teste. O achado central é uma **inversão de vantagem arquitetural dependente do tamanho**: o YOLOv11m demonstrou F1 e Precisão superiores na faixa *small*, enquanto o RT-DETR-L obteve *Recall* e F1 superiores nas faixas *medium* e *large*. Mecanisticamente, massas pequenas ocupam poucos *tokens* no *feature map* após *downsampling* para 640 px; o *self-attention* do RT-DETR dilui o sinal radiômico ao distribuí-lo por correlações espaciais de longo alcance inadequadas para objetos compactos, enquanto o campo receptivo local do YOLOv11m é mais adequado para padrões de pequena extensão e alto contraste.

Os testes de Wilcoxon pareados (Tabela 4) confirmaram estatisticamente essa inversão: 5 dos 12 testes atingiram significância após Bonferroni. O RT-DETR-L foi significativamente superior em *Recall* nas três faixas e em F1 na faixa grande; o YOLOv11m foi significativamente superior em F1 na faixa pequena ($p = 1{,}30 \times 10^{-4}$). Nenhuma diferença significativa foi detectada em mAP@0,5 ou mAP@0,5:0,95, evidenciando que métricas de AP agregadas são insuficientes para revelar diferenças arquiteturais clinicamente relevantes, o que reforça a necessidade de estratificação por tamanho e de métricas complementares ao mAP.

Tabela 3. Desempenho por faixa e modelo (conjunto de teste).

Faixa	Modelo	mAP@0,5	mAP@0,5:0,95	Precisão	*Recall*	F1
small	YOLOv11m	**0,466**	**0,242**	**0,339**	0,376	**0,347**
	RT-DETR-L	0,463	0,210	0,132	**0,569**	0,204
medium	YOLOv11m	0,629	0,315	**0,453**	0,531	0,475
	RT-DETR-L	0,610	0,305	0,436	**0,615**	**0,486**
large	YOLOv11m	**0,608**	**0,306**	0,353	0,386	0,358
	RT-DETR-L	0,575	0,282	**0,514**	**0,523**	**0,506**

Negrito: melhor valor descritivo por faixa.

Fonte: Elaborado pelos autores (2026).

Tabela 4. Testes de Wilcoxon pareados por imagem (apenas teste).

Faixa	Métrica	Δ	*p*-valor	Sig.*	n
small	mAP@0,5	−0,003	1,000	Não	126
	mAP@0,5:0,95	−0,033	1,000	Não	126
	Recall	+0,193	$8{,}53 \times 10^{-5}$	Sim	126
	F1	−0,143	$1{,}30 \times 10^{-4}$	Sim	126
medium	mAP@0,5	−0,020	1,000	Não	343
	mAP@0,5:0,95	−0,010	1,000	Não	343
	Recall	+0,084	0,0021	Sim	343
	F1	+0,012	1,000	Não	343
large	mAP@0,5	−0,034	0,527	Não	282
	mAP@0,5:0,95	−0,024	0,361	Não	282
	Recall	+0,138	$1{,}39 \times 10^{-5}$	Sim	282
	F1	+0,149	$2{,}12 \times 10^{-7}$	Sim	282

Δ > 0: RT-DETR-L superior; Δ < 0: YOLOv11m superior. *Após Bonferroni (α/12 ≈ 0,0042).

Fonte: Elaborado pelos autores (2026).

Tabela 5. Teste de McNemar para detecção binária por imagem (IoU ≥ 0,3).

Faixa	Taxa DETR	Taxa YOLO	Δ	*p*-valor	Sig.*
small	0,643	0,429	+0,214	$3{,}72 \times 10^{-6}$	Sim
medium	0,703	0,586	+0,117	$1{,}72 \times 10^{-6}$	Sim
large	0,716	0,454	+0,262	$1{,}87 \times 10^{-14}$	Sim

*Após Bonferroni ($\alpha/3 \approx 0{,}0167$).

Fonte: Elaborado pelos autores (2026).

O teste de McNemar (Tabela 5) revelou o achado mais consistente: o RT-DETR-L detectou significativamente mais imagens com massa nas três faixas após correção de Bonferroni, com vantagens de 21,4 pontos percentuais (pp) na faixa pequena ($p = 3{,}72 \times 10^{-6}$), 11,7 pp na média ($p = 1{,}72 \times 10^{-6}$) e 26,2 pp na grande ($p = 1{,}87 \times 10^{-14}$). Esse resultado, não capturado pelo Wilcoxon sobre mAP, demonstra que a métrica binária por imagem é mais sensível para revelar diferenças clinicamente relevantes entre arquiteturas: ao maximizar a detecção verdadeira, o Transformer reduz falsos negativos, porém ao custo de maior número de falsos positivos, especialmente na faixa pequena, onde a Precisão colapsa para 0,132 contra 0,339 do YOLOv11m.

A análise FROC e o experimento SAHI (Tabela 6) reforçam e qualificam esse padrão. Na faixa pequena, o YOLOv11m manteve sensibilidade FROC superior a baixos níveis de FP/imagem (0,228 *vs.* 0,059 a 0,125 FP/img), coerente com seu F1 superior. Nas faixas média e grande o RT-DETR-L dominou em todos os pontos operacionais. O SAHI melhorou marginalmente o *Recall* do YOLOv11m (+4,3 pp) à custa de Precisão (−9,8 pp), enquanto a vantagem de detecção binária do RT-DETR-L persistiu após *tiling* (73,0% *vs.* 46,8%), sustentando que as diferenças entre as arquiteturas **não se explicam apenas pela resolução de entrada**. Note-se que, por ausência de imagens negativas no conjunto de teste, o FROC mede FP excedentes em imagens positivas e não é diretamente comparável a estudos com FROC clínico completo.

A análise de **IA Explicável** (Figura 1) fornece interpretação mecanicista das diferenças quantitativas. Os mapas Grad-CAM [17] do YOLOv11m foram gerados via última camada Conv2d do *backbone*; os mapas de atenção cruzada do RT-DETR-L foram extraídos via *forward hook* no módulo MSDeformAttn do *decoder*, processados por escala (P3, P4, P5) e fundidos por média após *Gaussian splat* e suavização bicúbica. Três casos representativos da faixa *medium* ilustram os padrões típicos: no **Caso 1** (verdadeiro positivo em ambos), o Grad-CAM cobre difusamente o parênquima denso, enquanto a atenção do RT-DETR ancora seletivamente sobre a lesão; no **Caso 2** (RT-DETR detecta, YOLO falha), a atenção cruzada foca com precisão a massa inferior (conf = 0,86), enquanto o Grad-CAM dispersa pela borda mamária; no **Caso 3** (falso positivo do YOLO), a ativação convolucional sobre tecido fibroglandular denso gera detecção espúria, enquanto o RT-DETR ancora corretamente na região anotada. Esses padrões são consistentes com a hipótese de que o *self-attention* do Transformer constrói representações contextuais mais seletivas para lesões de maior extensão espacial.

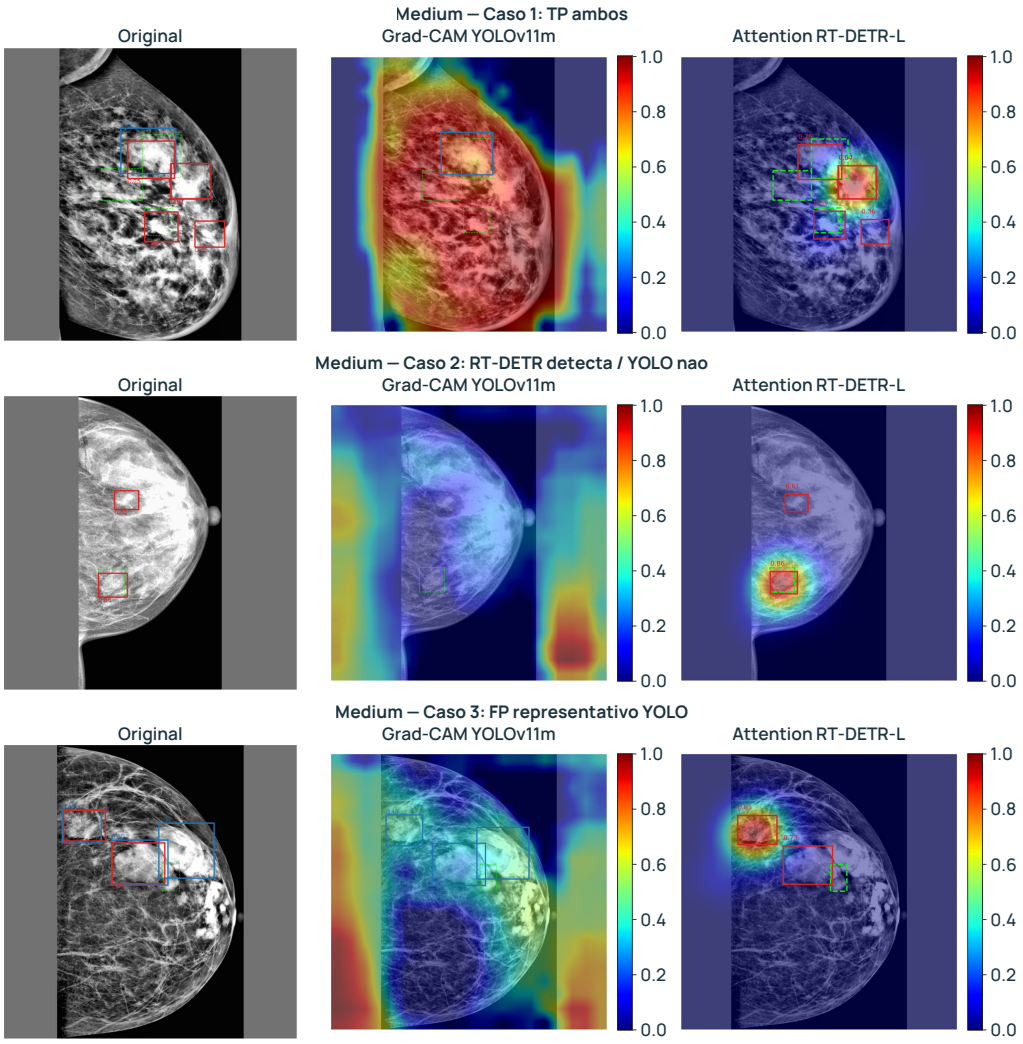
Tabela 6. Sensibilidade FROC e impacto do SAHI na faixa pequena (IoU ≥ 0,3).

Faixa	Modelo	@ 0,125	@ 0,5	@ 1,0	@ 2,0	Detecção SAHI (%)
small	YOLOv11m	**0,228**	**0,397**	**0,456**	**0,529**	46,8%
	RT-DETR-L	0,059	0,176	0,338	0,449	**73,0%**
medium	YOLOv11m	0,362	0,564	0,627	0,689	N/A
	RT-DETR-L	**0,411**	**0,616**	**0,662**	**0,719**	N/A
large	YOLOv11m	0,345	0,485	0,595	0,688	N/A
	RT-DETR-L	**0,473**	**0,643**	**0,673**	**0,702**	N/A

Detecção SAHI: taxa de detecção binária por imagem com SAHI (IoU ≥ 0,3). N/A: SAHI aplicado apenas na faixa *small*. Negrito: melhor valor por coluna e faixa.

Fonte: Elaborado pelos autores (2026).

Figura 1. IA Explicável: painéis comparativos Grad-CAM (YOLOv11m) vs. mapa de atenção cruzada (RT-DETR-L) para a faixa medium (19,4–36,3 mm). Cada painel: imagem original | Grad-CAM YOLOv11m | Mapa de Atenção RT-DETR-L.



Fonte: Elaborado pelos autores (2026).

4 DISCUSSÃO

Antes das implicações específicas destes dois modelos, o padrão de significância observado sugere quatro conhecimentos de alcance mais geral para a Informática em Saúde. Primeiro, nenhum dos seis testes sobre mAP@0,5 e mAP@0,5:0,95 atingiu significância, contra cinco dos seis testes em *Recall* e F1 e três dos três testes de McNemar: métricas de AP agregadas podem mascarar diferenças que só se revelam quando o desempenho é estratificado por tamanho, razão pela qual o relato estratificado deveria compor a prática padrão de avaliação de sistemas de CAD mamográfico, ao lado das métricas globais. Segundo, a taxa de detecção binária por imagem mostrou-se mais informativa que a AP para o cenário de triagem, por corresponder à decisão de rechamar ou não a paciente. Terceiro, se a vantagem arquitetural depende da escala do alvo, a seleção de detectores para apoio ao diagnóstico deveria ser condicionada à distribuição de tamanhos do achado de interesse, e não a posições em *rankings* de *benchmarks* gerais. Quarto, a escolha do ponto operacional é uma decisão de serviço e não apenas de modelo: a queda de Precisão do RT-DETR-L a 0,132 na faixa *small* implicaria carga de rechamada possivelmente insustentável em serviços com fila de espera. Essas quatro implicações são recomendações metodológicas sugeridas pelos dados, e não conclusões causais sobre as arquiteturas comparadas, conforme a ressalva detalhada nas limitações.

Quatro vieses algorítmicos principais devem ser considerados para uso responsável dos modelos. **(1) Viés de domínio:** os *datasets* provêm de Portugal, Egito e Vietnã, sem dados de equipamentos da rede pública brasileira, representando risco de *domain shift* em mamógrafos e populações do SUS; a mitigação requer coleta colaborativa para *fine-tuning* local. **(2) Viés de anotação:** heterogeneidade de convenções (XML/polígono, CSV/máscara, DICOM) introduz inconsistências no *ground truth*, mitigadas pela padronização via retângulo envolvente mínimo e critério único de inclusão. **(3) Viés de calibração espacial:** o espaçamento *fallback* do CDD-CESM (0,094 mm/px) pode introduzir imprecisão na definição de faixas para casos fronteiriços. **(4) Ausência de calibração de confiança:** sem *temperature scaling*, as probabilidades de detecção

não são interpretáveis como probabilidades clínicas, limitando a integração em fluxos assistenciais reais.

A estratégia híbrida proposta (RT-DETR-L para triagem de alta sensibilidade, seguido de YOLOv11m para refinamento em massas pequenas) é clinicamente motivada, mas permanece uma hipótese de projeto ainda não validada em cenário assistencial. Os requisitos computacionais observados são modestos: os seis experimentos foram treinados e avaliados em GPU de médio porte (RTX 3070, 8 GB VRAM), sem infraestrutura de *data center*. Isso é relevante para o SUS, embora não constitua, por si, evidência de viabilidade de implantação, a qual dependeria de integração a *workstations* DICOM existentes, de desempenho verificado em equipamentos e populações locais e de avaliação de custo operacional, dimensões não medidas neste estudo. Em 2022 foram realizadas 3.856.595 mamografias de rastreamento em mulheres no SUS [18], cifra que dimensiona a escala do problema e o potencial de impacto de ferramentas de apoio ao diagnóstico; extrapolar, porém, ganhos de sensibilidade na faixa *small* (correspondente à categoria T1 do TNM, ≤ 20 mm [2]) para benefício populacional não é sustentado pelos dados aqui apresentados, obtidos sem imagens negativas e sem dados brasileiros. Eventual implantação exigiria ainda atenção a *data drift* e protocolos de atualização periódica em conformidade com a Lei Geral de Proteção de Dados (LGPD).

As principais limitações do estudo são: (1) FROC calculado sem imagens negativas, não refletindo cenário de triagem real; (2) INbreast sub-representado (116 anotações); (3) escopo restrito a massas, sem calcificações ou distorções arquitetônicas; (4) ausência de validação prospectiva com dados e equipamentos brasileiros; (5) regra de atribuição de faixa pela maior lesão; (6) ausência de calibração de confiança antes de uso clínico; e (7) treinamento único por configuração, com semente fixa: o Wilcoxon pareado e o McNemar medem variabilidade entre imagens, não entre inicializações, de modo que os testes não separam o efeito arquitetural da variabilidade entre treinos, o que exigiria repetições com sementes distintas.

5 CONCLUSÃO

Este estudo respondeu ao objetivo de comparar YOLOv11m e RT-DETR-L para detecção de massas mamárias estratificada por tamanho, demonstrando que as arquiteturas apresentam perfis complementares e estatisticamente distintos conforme a faixa de lesão. O RT-DETR-L alcançou *Recall* significativamente superior nas três faixas ($p < 0{,}003$) e taxa de detecção binária superior em todas ($p < 10^{-5}$, McNemar), sendo o componente preferencial para triagem de alta sensibilidade. O YOLOv11m alcançou F1 significativamente superior na faixa pequena ($p = 1{,}30 \times 10^{-4}$), sendo mais adequado para refinamento em massas incipientes, onde o controle de falsos positivos é determinante. A análise SAHI indicou que essas diferenças não decorrem apenas da resolução de entrada, e a análise de IA Explicável revelou que o *self-attention* do Transformer ancora seletivamente sobre lesões de maior extensão espacial, enquanto o campo receptivo local da CNN é mais adequado para padrões compactos.

Os resultados sustentam a hipótese de uma arquitetura híbrida de CAD clinicamente motivada e com requisitos computacionais modestos, obtidos sem infraestrutura de *data center*. Trata-se, contudo, de um estudo pré-clínico e gerador de hipóteses, conduzido exclusivamente com dados públicos internacionais. A replicação com dados brasileiros, a calibração de confiança dos modelos, a avaliação de *domain shift* e o estabelecimento de protocolos de governança sob a LGPD constituem etapas necessárias e prévias a qualquer afirmação sobre contribuição efetiva para a equidade do rastreamento mamográfico no Brasil ou sobre viabilidade de implantação em serviços de radiologia do SUS. Trabalhos futuros devem investigar *copy-paste augmentation*, intervalos de confiança para FROC por *bootstrap* e integração em fluxos DICOM reais.

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 à Pró-Reitoria Estudantil (PROEST) da Universidade Federal de Alagoas pelo auxílio financeiro para a participação no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026).

Financiamento. Este trabalho foi financiado pelo Programa Pesquisa para o SUS (PPSUS), por meio do Ministério da Saúde, CNPq, SESAU-AL e FAPEAL, Chamada FAPEAL nº 02/2025, Processo nº E: 60030.0000001374/2025.

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

Conflitos de interesse. Os autores declaram não haver conflitos de interesse.

Declaração de uso de IA. Declaramos que utilizamos a ferramenta Claude (Anthropic) para apoio pontual na revisão gramatical e na estruturação do manuscrito. A ferramenta não foi utilizada para geração de resultados, desenvolvimento do código experimental ou análise estatística. Os autores assumem total responsabilidade pelo conteúdo final do manuscrito.

REFERÊNCIAS

1. Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2023: incidência de câncer no Brasil [Internet]. Rio de Janeiro: INCA; 2022 [citado em 28 jul 2026]. 160 p. Disponível em: https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil
2. Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, et al., editores. AJCC Cancer Staging Manual. 8th ed. New York: Springer International Publishing; 2017. doi: https://doi.org/10.1007/978-3-319-40618-3
3. Instituto Nacional de Câncer José Alencar Gomes da Silva. Rastreamento do câncer de mama na população-alvo [Internet]. Rio de Janeiro: INCA; 2023 [citado 2026 jul 28]. Dados da Pesquisa Nacional de Saúde 2019 (IBGE). Disponível em: https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/dados-e-numeros/rastreamento-do-cancer-de-mama-na-populacao-alvo
4. Chojniak R, Cerri GG. Radiology in Brazil and the global resolution on strengthening medical imaging: a regional responsibility. Radiol Bras. 2026;59:e20250106. doi: https://doi.org/10.1590/0100-3984.2025.0106
5. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep. 2018;8(1):4165. doi: https://doi.org/10.1038/s41598-018-22437-z
6. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916-22. doi: https://doi.org/10.1093/jnci/djy222
7. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94. doi: https://doi.org/10.1038/s41586-019-1799-6
8. Jocher G, Qiu J, Chaurasia A. Ultralytics YOLO [software]. Versão v8.3.0. [s.l.]: Zenodo; 2024. doi: https://doi.org/10.5281/zenodo.13858602
9. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. In: Lecture Notes in Computer Science. Cham: Springer International Publishing; 2020. p. 213-29. doi: https://doi.org/10.1007/978-3-030-58452-8_13
10. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs Beat YOLOs on Real-time Object Detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [s.l.]: IEEE; 2024. p. 16965-74. doi: https://doi.org/10.1109/cvpr52733.2024.01605
11. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol. 2012;19(2):236-48. doi: https://doi.org/10.1016/j.acra.2011.09.014
12. Khaled R, Helal M, Alfarghaly O, Mokhtar O, Elkorany A, El Kassas H, et al. Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. Sci Data. 2022;9(1):122. doi: https://doi.org/10.1038/s41597-022-01238-0
13. Nguyen HT, Nguyen HQ, Pham HH, Lam K, Le LT, Dao M, et al. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Sci Data. 2023;10(1):277. doi: https://doi.org/10.1038/s41597-023-02100-7
14. American College of Radiology. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. 5th ed. Reston (VA): American College of Radiology; 2013.
15. Akyon FC, Onur Altinuc S, Temizel A. Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In: 2022 IEEE International Conference on Image Processing (ICIP). [s.l.]: IEEE; 2022. p. 966-70. doi: https://doi.org/10.1109/icip46576.2022.9897990
16. Brasil. Ministério da Saúde. Conselho Nacional de Saúde. Resolução nº 510, de 7 de abril de 2016. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais [Internet]. Diário Oficial da União. 2016 maio 24 [citado 2026 out 18];(98 Seção 1):44-6. Disponível em: https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2016/resolucao-no-510.pdf
17. 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
18. Instituto Nacional de Câncer José Alencar Gomes da Silva. Mamografias no SUS [Internet]. Rio de Janeiro: INCA; 2023 [citado 2026 jul 28]. Fonte primária: Ministério da Saúde, Sistema de Informações Ambulatoriais do SUS (SIA/SUS), 2022. Disponível em: https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/dados-e-numeros/mamografias-no-sus

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