ISSN 2596-2701 - Revista de Ciência da Computação - ReCiC v. 8 n. 1 - jan/dez (2026)
Dossiê Temático - Computação Aplicada à Saúde / Saúde Digital

Segmentação de Lesões Mamárias com nnU-Net: Evidência Negativa do Crescimento Guloso e Limites do Pós-processamento

Harold Ivan Angulo Bustos
Universidade do Estado do Rio Grande do Norte (UERN)
Brasil
haroldivan@uern.br

Francisco Chagas de Lima Júnior
Universidade do Estado do Rio Grande do Norte (UERN)
Brasil
limajunior@uern.br

Jorge Luiz Silva Braz
Universidade do Estado do Rio Grande do Norte (UERN)
Brasil
jorgebraz@alu.uern.br

Francisco Fernandes do Vale Segundo
Universidade do Estado do Rio Grande do Norte (UERN)
Brasil
valesegundo@alu.uern.br

***Resumo* —Este trabalho avalia pós-processamento da nnU-Net na segmentação de lesões mamárias no INbreast, sob a restrição de não retreinar a rede. A hipótese inicial, recuperar margens por crescimento guloso sobre sete mapas de evidência, não se sustentou nas 107 imagens com massa (50 pacientes); correções incrementais e um *grid search* de 100 combinações tampouco superaram o núcleo bruto (V3.2: Dice $0{,}7010$; $\Delta = -0{,}0216$). A premissa de expansão universal era inadequada na maior parte das instâncias: BEM+SUPER $= 89/107$ ($83{,}2\%$; Seção IV). Em validação cruzada aninhada por paciente, a triagem por direção de erro não supera o baseline (Dice $0{,}7242$; $\Delta = +0{,}0016$; IC 95% $[-0{,}0041;\ +0{,}0071]$; $p = 0{,}141$) e fica abaixo de baselines triviais (limiar de probabilidade e erosão). O ganho aparente do limiar ($\Delta = +0{,}0435$) é apenas interno: a rede foi avaliada nas próprias imagens de treino, sem validade externa demonstrada.**

***Palavras-chave* —Inteligência artificial, nnU-Net, INbreast, Aprendizado de máquina, Segmentação de lesões, Mamografia, Saúde digital.**

I. Introdução

Uma nnU-Net [1] treinada para detectar lesões mamárias acerta, de forma recorrente, a região central da massa, a parte de maior contraste e mais homogênea, e deixa de fora a margem: espículas, halos difusos e extensões de baixo contraste. A morfologia dessa margem associa-se à suspeição clínica; o problema de Computação é o viés residual da própria rede, que distorce a fronteira.

Duas explicações, coerentes com a literatura de treino e com o padrão observado, orientaram o início do trabalho. Funções de perda como Dice e entropia cruzada binária tratam cada *pixel* com peso igual: o núcleo, de contraste mais forte, domina o gradiente, e a borda difusa contribui pouco para a atualização dos pesos. Além disso, o desbalanceamento de classe (a lesão é uma fração pequena da imagem) favorece modelos conservadores, que preferem errar por falta a errar por excesso. Esses fatores oferecem uma explicação plausível para o comportamento observado, mas não foram isolados experimentalmente neste estudo.

A correção na origem seria retreinar a rede com uma perda que penaliza mais a subsegmentação, como Tversky, Focal Tversky ou Boundary Loss [2]–[4]. Não havia, contudo, janela para retreino. A restrição de projeto **(R1)**, não retreinar a nnU-Net, fixou o escopo em pós-processamento da máscara já produzida. A hipótese inicial foi de *subsegmentação sistemática*: dado o núcleo, quanto da extensão real da lesão se recupera por processamento clássico de imagem? A primeira família de soluções (V3.1–V3.3) foi um crescimento guloso por fusão de sete mapas de evidência.

A avaliação nas 107 imagens com massa do INbreast [5] (50 pacientes; Dataset002_INBreast16) mostrou que essa família piora o Dice em relação ao núcleo bruto e que nenhuma de 100 combinações de hiperparâmetros supera o baseline. O diagnóstico caso a caso estabeleceu que a premissa de expansão universal era falsa na maior parte das imagens (Figura 1, centro).

Esse resultado negativo deslocou a pergunta: em vez de consertar o crescimento, passou-se a estimar a *direção* do erro residual (sub versus supersegmentação) com um classificador probabilístico, ainda sem retreinar a nnU-Net.

Este trabalho contribui com (i) evidência negativa de que o crescimento guloso multi-evidência, nesta formulação, não recupera a margem; (ii) o diagnóstico de que a expansão universal era inadequada em $89/107$ imagens (BEM+SUPER); (iii) a demonstração, em CV aninhada por paciente, de que

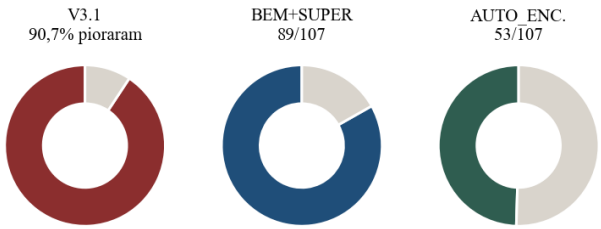

Fig. 1. Frações ($N = 107$): 90,7% pioraram na V3.1; BEM+SUPER = 83,2%; AUTO_ENCOLHER = 53/107 na CV aninhada.

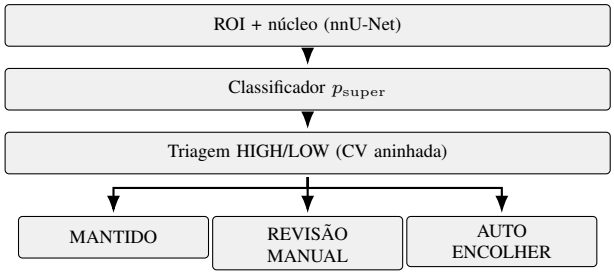

Fig. 2. Pipeline de triagem: o encolhimento só ocorre em AUTO_ENCOLHER.

a triagem por direção de erro não supera o baseline nem baselines triviais (limiar e erosão). O limiar de probabilidade aparece como comparação interna, não como método validado: os números da Tabela I são *internos* (a nnU-Net treinou nestas imagens). A leitura radiológica permanece indispensável.

O restante do artigo organiza-se como segue. A Seção II discute trabalhos relacionados. A Seção III descreve os mapas de evidência, o crescimento guloso, a calibração e o *pipeline* de triagem. A Seção IV registra o diagnóstico que alterou a pergunta de pesquisa. A Seção V apresenta os experimentos. A Seção VI discute os resultados. A Seção VII lista limitações. A Seção VIII indica trabalhos futuros. A Seção IX conclui. A Figura 2 resume o fluxo de triagem final.

II. Trabalhos Relacionados

A. nnU-Net, perdas assimétricas e Attention U-Net

A nnU-Net [1] é, neste trabalho, a fonte do núcleo da lesão. Não se propõe uma nova arquitetura: usa-se a nnU-Net v2 2D treinada no Dataset002_INBreast16 (imagens de 16 *bits*; 107 imagens com massa; fold 0), cuja saída alimenta o pós-processamento.

Na mesma linha de arquitetura e perda situam-se a Attention U-Net [6] e as perdas Tversky, Focal Tversky e Boundary Loss [2]–[4], que atacam o viés de subsegmentação no treino. A restrição **(R1)** excluiu o retreino no horizonte deste estudo; o pós-processamento e o classificador de direção são a alternativa adotada. Em [3] combina-se Focal Tversky com Attention U-Net; em [4] introduz-se Boundary Loss para segmentação altamente desbalanceada.

B. Crescimento por contraste

A primeira abordagem considerada foi partir do núcleo e expandir a máscara na direção em que o contraste ainda parece pertencer à lesão, o princípio de contornos ativos e de *watershed*, com amplo uso em segmentação médica.

Em [7] formaliza-se a evolução de um contorno modulada por uma função de velocidade derivada do gradiente: o contorno desacelera perto de bordas fortes. Em [8] propõe-se variante baseada em homogeneidade estatística de intensidade dentro e fora da região, mais tolerante a bordas fracas desde que exista diferença de média entre lesão e fundo, precisamente o caso de halos difusos. Estudou-se ainda a implementação morfológica eficiente de [9] (*morphsnakes*), que não foi a instanciação operacional da V3.

Em paralelo, considerou-se *watershed* com dois marcadores, interno (núcleo, via transformada de distância) e externo (dilatação morfológica), seguindo [10], proposto para mamografia. O método é determinístico e de passo único, sem número de iterações nem força de balão para calibrar, e os dois marcadores forçam uma partição completa do espaço.

Essas famílias compartilham a mesma fonte de informação: contraste de intensidade ou magnitude de gradiente local. Na ponta de uma espícula real essa informação falta: a estrutura esmaece até intensidade e gradiente comparáveis aos de fibras do parênquima (ligamentos de Cooper, tecido fibroglandular). O problema deixa de ser qual método de crescimento usar e passa a ser que evidência, além de contraste, distingue espícula de fibra.

C. Geometria radial e espiculação

Espículas reais convergem radialmente para o centro da massa; fibras normais têm orientação aproximadamente aleatória, ou seguem padrão anatômico não centrado na lesão. A pista discriminativa está na geometria em relação ao núcleo.

Em [11] formaliza-se o *Radial Gradient Index* (RGI): mede-se o quanto o vetor gradiente em cada ponto candidato aponta para o centro da lesão. Estruturas que apontam para o núcleo são favorecidas; orientação aleatória é penalizada mesmo com magnitude de gradiente alta. A partir dessa ideia, a geometria radial passou a ser critério de primeira classe, ao lado do contraste.

Em [12] constrói-se um mapa de orientação de estruturas lineares e busca-se convergência, detector de padrão estelar, próximo de uma transformada de Hough para linhas. A característica ALOE (*Average Local Orientation Entropy*) ataca o mesmo problema de forma estatística: tecido normal tem orientação dominante; massa espiculada tem muitas orientações. O mapa E_{or} implementado não reproduz o histograma de entropia do ALOE: calcula-se uma resultante circular sobre o tensor de estrutura, no mesmo espírito. O mapa E_{rad} implementa o RGI.

Em [13] combinam-se probabilidade de massa, filtro de crista multiescala e *graph-cut*. No subconjunto mais difícil (54 massas mal definidas ou espiculadas), reportam *Area Overlap Ratio* de $0{,}689 \pm 0{,}160$ para a massa inteira e $0{,}540 \pm 0{,}164$ para a margem, com falhas até $0{,}361$ em baixo contraste. Esses

valores são teto de expectativa da literatura clássica, não meta empírica deste estudo; AOR não é reportado aqui.

D. Borda robusta a contraste e otimização combinatória

Halos e margens difusas têm contraste baixo por definição; um operador de Sobel responde pouco exatamente onde se precisa de um sinal de decisão. A congruência de fase [14] localiza estruturas pelo alinhamento de fase em múltiplas escalas, independentemente da amplitude absoluta. Formalizações via sinal monogênico [15] e refinamentos posteriores [16] confirmam a linha. A formulação original em [14] usa um banco de filtros *log-Gabor* orientados; a V3 adota versão isotrópica (um filtro radialmente simétrico por escala), preservando a robustez a contraste e delegando a discriminação por orientação aos mapas radial e de orientação.

Em [17] e em GrabCut [18] combinam-se termos de região e de fronteira por corte mínimo. Em [19] impõe-se *prior* de forma estelar no grafo; em [20] pondera-se o *prior* de forma de modo adaptativo. Em [21] formaliza-se competição de regiões com termo de curvatura, inspiração da penalidade da V3.3. Em [22] descreve-se *prior* de convexidade parcial para microlobulação, não implementado na V3. Nenhuma dessas referências, por si, resolve o diagnóstico de direção do erro (sub versus super), ponto de falha da família de crescimento.

III. Materiais e métodos

A. Base de dados e partição

As imagens provêm do INbreast [5], banco público de mamografias digitais de campo completo. A descrição original da base [5] indica 410 imagens e 115 casos. Na cópia utilizada neste estudo, a contagem de pacientes distintos a partir dos identificadores DICOM resultou em 108 (não 115); a divergência é de recontagem na base local, não de outro corpus. A aquisição original ocorreu no Breast Centre do Centro Hospitalar de S. João (Porto), entre abril de 2008 e julho de 2010, com autorização do Comitê de Ética do hospital e da Comissão Nacional de Proteção de Dados de Portugal; informações clínicas confidenciais foram removidas dos arquivos DICOM conforme o Supplement 55 do padrão DICOM [5].

O corpus deste estudo são as 107 imagens com massa do INbreast (50 pacientes; 16 *bits*), organizadas no Dataset002_INBreast16. A nnU-Net v2 2D foi treinada nesse conjunto sob o fold 0 (85 treino, 22 validação, por *imagem*); não há holdout externo de imagens com massa fora dessas 107. Das 22 de validação, 20 são de pacientes com outra imagem no treino; só 2 imagens de 1 paciente ficam fora do treino em nível de paciente. O treino foi interrompido na época 132 de 1 000 (melhor EMA pseudo Dice 0,8401; Tesla T4; PyTorch 2,11,0+cu128); a inferência usa o melhor *checkpoint*, sem TTA e sem pós-processamento da nnU-Net. O Dice do baseline nas 107 predições é 0,7226 (interno: 0,7313 nas 85 de treino e 0,6889 nas 22 de validação).

A avaliação principal do pós-processamento (triagem e baselines) é validação cruzada aninhada por *paciente*: 5 folds externos $\times$ 5 internos, sem paciente em dois folds. Os limiares da triagem são recalibrados só no treino de cada fold externo. A partição interna $75/32$ ($\approx 70/30$) do *pipeline* clássico de crescimento/triagem sobre as mesmas 107 imagens é a mesma referida como $\approx 70/30$, feita por imagem: 21 das 32 imagens do holdout têm outra imagem do mesmo paciente no ajuste.

B. Entrada e restrições

O *pipeline* recebe a ROI mamográfica e a máscara binária do núcleo gerada pela nnU-Net. Não há retreino da rede de segmentação **(R1)** no horizonte deste estudo de pós-processamento.

A família V3 opera por processamento clássico (gradiente, transformadas, busca gulosa) sobre a saída da nnU-Net. O *pipeline* final acrescenta um classificador probabilístico da direção do erro residual, sem alterar os pesos da rede.

C. Ambiente computacional

O pós-processamento e a estatística rodaram em CPU no Google Colab Pro; o treino e a reinferência com probabilidades da nnU-Net usaram uma Tesla T4. Na sessão da erosão em raio fixo (Intel Xeon 2,00 GHz, 2 núcleos): Python 3,13,16, NumPy 2,4,6, pandas 2,2,3, SimpleITK 2,5,7, OpenCV 5,0,0, SciPy 1,16,3, scikit-learn 1,6,1, scikit-image 0,25,2, Numba 0,64,0. Sessões anteriores diferem em até uma casa de versão (p.ex. OpenCV 4,14,0) e até 0,0067 de Dice no encolhimento. O classificador reportado usa `LogisticRegression(class_weight="balanced", max_iter=1000)` e `StandardScaler` do scikit-learn.

D. Aspectos éticos e de privacidade

Como a base é pública e foi liberada já anonimizada, este estudo não coleta novos dados de pacientes. A conformidade ética da aquisição original está documentada em [5] (Comitê de Ética hospitalar e Comissão Nacional de Proteção de Dados de Portugal; DICOM sem informações confidenciais). Não se apresenta número de protocolo CEP/CONEP brasileiro específico deste artigo; o uso restringe-se à base pública citada.

E. Pré-processamento

A ROI é normalizada pelos percentis 2 e 98 de intensidade (calculados apenas no interior da ROI), seguida de suavização gaussiana com $\sigma = 1{,}0$. A escolha de percentis, e não de mínimo/máximo absolutos, é deliberada: mamografias frequentemente têm *pixels* extremos isolados (ruído de aquisição, bordas de recorte, artefatos), e um único valor espúrio comprimiria a faixa dinâmica útil. Usar percentis descarta essas caudas antes de definir a escala.

Quando o maior lado da ROI excede 1400 *pixels* (`MAX_SIZE`), imagem e núcleo são reduzidos para o cálculo dos mapas e o resultado é reescalado à resolução original, salvaguarda de desempenho; a decisão de segmentação é sempre devolvida à grade original.

F. Mapas de evidência

Sete mapas são calculados sobre a imagem normalizada $\tilde{I}$. Cada um responde a uma falha observada do crescimento por contraste isolado.

a) Gradiente.: $g_x = \mathrm{Sobel}_x(\tilde{I})$, $g_y = \mathrm{Sobel}_y(\tilde{I})$, $E_{\mathrm{grad}} = \mathrm{norm}\big[\sqrt{g_x^2 + g_y^2}\big]$. Não se espera que resolva espiculação; em massas circunscritas de margem nítida, contudo, o gradiente forte e consistente ao redor do contorno é suficiente, e o termo evita que o algoritmo dependa só dos mapas especializados no caso "fácil".

b) Congruência de fase (versão isotrópica).: Em quatro escalas (comprimentos de onda 3, 5, 9 e 15 *pixels*), aplica-se filtro *log-Gabor* radialmente simétrico no domínio de Fourier:

$$\mathrm{PC}(x,y) = \frac{\sum_s \big|\mathrm{Re}(R_s(x,y))\big|}{\sum_s \big|R_s(x,y)\big| + \varepsilon}, \qquad s \in \{3,5,9,15\}. \quad (1)$$

Quatro escalas existem porque halos e espículas não têm largura de estrutura fixa: halos largos respondem melhor a 9 e 15 px, enquanto a borda fina de uma espícula responde melhor a 3 e 5 px; somar a energia evita escolher uma única escala por lesão. Trata-se de simplificação deliberada da formulação de congruência de fase [14] (sem banco de orientações); a discriminação por orientação é delegada aos mapas radial e de orientação.

c) Tensor de estrutura.: Este passo intermediário não é, em si, um mapa de evidência: é o alicerce de dois mapas diferentes. Com J_{xx}, J_{yy} e J_{xy} suavizados ($\sigma = 3$) a partir de g_x^2, g_y^2 e $g_x g_y$,

$$\theta = \tfrac{1}{2}\mathrm{atan2}(2J_{xy}, J_{xx} - J_{yy}),$$
$$\kappa = \frac{\sqrt{(J_{xx} - J_{yy})^2 + 4J_{xy}^2}}{J_{xx} + J_{yy} + \varepsilon}. \quad (2)$$

O índice radial usa o gradiente bruto; a evidência de orientação usa especificamente θ e a coerência κ extraídos daqui.

d) Índice radial.: Seja (c_y, c_x) o centróide do núcleo, $\theta_{\mathrm{radial}} = \mathrm{atan2}(y - c_y, x - c_x)$ e $\theta_{\mathrm{grad}} = \mathrm{atan2}(g_y, g_x)$. Define-se

$$E_{\mathrm{rad}} = \big|\cos(\theta_{\mathrm{grad}} - \theta_{\mathrm{radial}})\big| \cdot \mathrm{norm}\big[\sqrt{g_x^2 + g_y^2}\big], \quad (3)$$

implementação direta do RGI [11]. Quando os ângulos coincidem (gradiente apontando para o centro), o cosseno vale próximo de 1; quando são perpendiculares ou opostos, cai para próximo de 0, derrubando a evidência mesmo com magnitude de gradiente forte. Este é, dos sete mapas, o de maior poder teórico de discriminação entre espícula real e fibra de fundo: é o único que depende da *posição* da estrutura em relação ao núcleo, e não apenas da aparência isolada.

e) Evidência de orientação.: A evidência de orientação agrega estatística circular de ângulo duplo em vizinhança maior ($\sigma = 5$):

$$E_{\mathrm{or}} = \sqrt{\langle\cos 2\theta\rangle^2 + \langle\sin 2\theta\rangle^2} \cdot \kappa. \quad (4)$$

Ângulo duplo é forma padrão de medir consistência de orientação sem a descontinuidade de ângulos brutos em $0°/180°$. Não é implementação literal do histograma de entropia do ALOE: mede resultante circular, não dispersão de histograma, reutilizando o tensor já calculado para o índice radial. Chamar isso de "versão do ALOE" seria impreciso; é inspirado no mesmo princípio (consistência local de orientação como evidência).

f) Intensidade (mediana e MAD).: Com mediana e MAD das intensidades no núcleo, $\sigma = 1{,}4826 \cdot \mathrm{MAD}(\tilde{I} \mid \text{núcleo}) + 0{,}03$ e

$$E_{\mathrm{int}} = \exp\Big[-\frac{(\tilde{I} - \mathrm{mediana})^2}{2\sigma^2}\Big]. \quad (5)$$

O termo modela a intensidade dentro do núcleo como gaussiana e avalia, em cada *pixel* da ROI, a plausibilidade de pertencer à mesma distribuição. O núcleo, mesmo sendo a região "fácil", não é perfeitamente homogêneo: pode ter calcificações, vasos ou pequenas variações de textura. Média e desvio-padrão são sensíveis a esses extremos; mediana e MAD não se deixam distorcer por uma minoria de *pixels* atípicos. O fator 1,4826 converte MAD para escala comparável a desvio-padrão sob hipótese gaussiana. No espírito do termo de dados de GrabCut [18], mas sem mistura de gaussianas completa e **sem** modelar a distribuição do fundo, só a do núcleo.

g) Distância.: A transformada de distância euclidiana ao núcleo é invertida e normalizada pelo 95º percentil das distâncias positivas, produzindo um mapa que decai suavemente conforme o *pixel* se afasta do núcleo. Sem esse termo, um *pixel* de textura parecida com a lesão, mas fisicamente distante (por exemplo, do outro lado da mamografia), poderia acumular escore alto nos outros seis mapas e ser puxado para a máscara. A distância é restrição de plausibilidade espacial: não impede crescimento, apenas o torna progressivamente mais caro.

h) Continuidade.: Densidade local de vizinhos já pertencentes à máscara em curso, numa janela 3×3. A motivação é um caso adversarial: um *pixel* isolado, cercado de tecido normal, mas que por acaso tem intensidade e gradiente parecidos com os da lesão, teria escore alto nos outros mapas sem nada que penalizasse a desconexão da frente de crescimento. A continuidade resolve isso localmente; a garantia rígida de conexão ao núcleo só é imposta no pós-processamento.

i) Circularidade adaptativa do termo radial.: Em núcleos já bastante circulares (massas circunscritas), o índice radial tem pouco a acrescentar, o gradiente sozinho delimita bem o contorno. Em núcleos alongados ou irregulares, já um indício visual de morfologia não redonda, vale a pena confiar mais no sinal geométrico radial. Com $C = 4\pi \cdot \text{Área}/\text{Perímetro}^2$ ($C = 1$ no círculo),

$$E'_{\mathrm{rad}} = \mathrm{clip}\big[E_{\mathrm{rad}} \cdot \big(1 + (1 - C) \cdot 0{,}35\big),\, 0,\, 1\big]. \quad (6)$$

Núcleos menos circulares reforçam o sinal radial. Foi só depois, com [20], que se reconheceu o mesmo princípio, dar mais peso a um *prior* de forma exatamente onde a evidência local sugere que ele é necessário. O que se implementou é versão de um único parâmetro escalar, não a formulação contínua completa; é o único ponto do *pipeline* de crescimento em que o peso de um termo muda de imagem para imagem.

G. Fusão linear

A linha de otimização combinatória estudada [13], [17], [19] apontaria para uma energia multi-termo resolvida por corte

mínimo, com garantia de ótimo global. Optou-se, neste estudo, por soma ponderada de mapas: *graph-cut* exige infraestrutura de min-cut/max-flow substancialmente mais pesada, e era necessário validar primeiro o conjunto de evidências.

Os mapas são combinados com pesos fixos calibrados manualmente (Seção III-K):

$$S = \frac{\sum_k w_k E_k}{\sum_k w_k}, \qquad S \in [0,1], \tag{7}$$

em que k percorre intensidade, gradiente, fase, radial reponderado, orientação, distância e continuidade. A divisão por $\sum w$ mantém S em $[0,1]$ quando os pesos individuais são alterados; sem ela, cada mudança de peso deslocaria a escala do escore e exigiria recalibrar os limiares.

H. Crescimento guloso best-first

Com o mapa S em mãos, a opção mais simples, um limiar único aplicado a S, foi descartada: um limiar fixo trata igualmente um *pixel* colado no núcleo e um *pixel* distante, e não usa a estrutura espacial do problema (a lesão é uma região conexa que cresce a partir de uma semente).

A solução implementada é busca gulosa *best-first*: uma fila de prioridade começa com os vizinhos imediatos do núcleo; a cada passo, remove-se o *pixel* de maior escore ainda não incorporado, aceito se passar em três testes: escore acima do limiar corrente, distância ao centróide dentro do raio máximo, e área acumulada abaixo do teto. Vizinhos do *pixel* aceito entram na fila. Na V3.1, os parâmetros foram `MAX_GROWTH_FACTOR` $= 1{,}70$ (fixo), `INITIAL_THRESHOLD` $= 0{,}62$, `MIN_THRESHOLD` $= 0{,}50$, `DECAY` $= 8 \times 10^{-7}$, `MAX_ITERATIONS` $= 100\,000$ e distância máxima igual a 3,0 vezes o raio do núcleo.

O algoritmo (i) define $\text{max_pixels} = \lfloor |\text{núcleo}| \cdot 1{,}70 \rfloor$ e $\text{max_distance} = 3{,}0 \cdot r_{\text{nucleo}}$; (ii) inicializa a fila com os vizinhos do núcleo; (iii) enquanto a fila não é vazia e $t <$ `MAX_ITERATIONS`, extrai o *pixel* de maior escore; (iv) descarta o candidato se $S < \tau(t)$, se a distância ao centróide excede max_distance ou se a área atingiu o teto; (v) caso contrário, incorpora o *pixel* e empilha os vizinhos ainda não visitados. O teto de área é a última linha de defesa: na calibração inicial, os mapas sozinhos não garantiram parada em todos os casos.

a) Limiar decrescente (histerese).: O limiar de aceitação decai com a iteração t:

$$\tau(t) = \max\big(\texttt{MIN_THRESHOLD},\ \texttt{INITIAL_THRESHOLD} - t \cdot \texttt{DECAY}\big). \tag{8}$$

Com limiar fixo, ou o algoritmo era conservador demais no início (perdendo *pixels* de fronteira genuína, mas com escore moderado) ou generoso demais no fim (aceitando ruído de fundo depois de já ter crescido bastante). A solução é o mesmo princípio da limiarização por histerese de Canny, aplicado não a *pixels* isolados, mas a uma frente de propagação: nas primeiras iterações só passam os de evidência muito forte; à medida que a fronteira avança, evidência mais moderada (tipicamente mais distante, parte mais incerta da lesão) passa a ser aceitável.

b) Teto de distância radial.: Sem essa restrição, a busca gulosa, sendo gulosa, eventualmente encontra um caminho de *pixels* de escore moderado que a leva para bem longe do núcleo, especialmente em textura fibroglandular densa. Limitar a distância euclidiana ao centróide a três vezes o raio médio do núcleo é uma aproximação heurística, e deliberadamente fraca, do *prior* de forma estelar [19]: em vez de impor visibilidade em linha reta de cada ponto ao centro (o que exigiria a maquinaria de grafo), simplesmente corta-se o crescimento além de um raio fixo. Não é a mesma coisa matematicamente; persegue o mesmo objetivo prático.

c) Teto de área.: Independentemente do que os mapas de evidência digam, a máscara final nunca pode exceder `MAX_GROWTH_FACTOR` vezes a área do núcleo. Na calibração inicial esse fator estava em $4{,}0\times$, e observou-se casos de baixo contraste em que o crescimento consumia praticamente toda a ROI antes de qualquer outro critério de parada, sinal de que os mapas, sozinhos, não garantem parada em todos os casos. O fator foi reduzido para $1{,}70\times$ (Seção III-K). Nenhum termo do escore penalizava fronteiras irregulares nem vazamento espicular pelo tecido vizinho. Na V3.1, o *overgrowth* sistemático produziu Dice $0{,}6292$ ($\Delta = -0{,}0934$; $90{,}7\%$ das 107 imagens pioraram).

I. Pós-processamento

A máscara que sai da busca gulosa não é a máscara final. Três operações corrigem inconsistências observadas nos experimentos, mais uma garantia de união com o núcleo: (i) fechamento morfológico com disco de raio 2, a busca gulosa, discreta e *pixel* a *pixel*, tende a deixar a fronteira com pequenas reentrâncias e buracos de um *pixel*; o fechamento suaviza isso sem alterar significativamente a forma geral; (ii) remoção de componentes conexos menores que 50 *pixels*, grupos que passaram nos critérios de aceitação mas não formam estrutura espacialmente significativa (ruído); (iii) retenção apenas de componentes que intersectam o núcleo original, garantia rígida de que a máscara final seja extensão contígua do núcleo (o termo de continuidade empurra nessa direção durante o crescimento; só aqui a propriedade é restrição absoluta); (iv) união explícita do núcleo à máscara final, a região de que a rede já tinha certeza nunca desaparece do resultado.

J. Confiança de fronteira

Faltava responder como o algoritmo sinaliza que não tem certeza do próprio resultado. Calcular confiança sobre toda a máscara, média do escore em todos os *pixels* incluídos, foi descartado: a maior parte da área é o próprio núcleo, onde a rede e os mapas já têm altíssima certeza. Uma média sobre toda a máscara ficaria artificialmente alta, mascarada pela região fácil, exatamente a que não é a fonte de incerteza.

A confiança é a média, nos *pixels* de fronteira externa (erosão binária de um *pixel* da máscara final), de

$$0{,}30\, E_{\text{phase}} + 0{,}25\, E'_{\text{rad}} + 0{,}20\, E_{\text{grad}} + 0{,}25\, E_{\text{int}}. \tag{9}$$

Os pesos são diferentes dos da fusão de crescimento: deixou-se de fora orientação, distância e continuidade, porque essas três são termos de apoio à decisão de crescimento (plausibilidade espacial, forma), não evidências diretas de que a fronteira encontrada é uma borda real de lesão. O que interessa aqui é especificamente "essa fronteira parece uma borda real?", e as quatro evidências mantidas respondem diretamente a essa pergunta.

Imagens abaixo de `CONFIDENCE_THRESHOLD` eram marcadas para revisão, alinhado ao uso de CAD como apoio à decisão. Na V3.1, contudo, o sinal de baixa confiança não alterava a máscara devolvida: o crescimento era aplicado mesmo quando o próprio algoritmo não confiava no resultado.

K. Calibração qualitativa dos parâmetros

Os pesos dos mapas e os parâmetros de crescimento não vieram de derivação teórica fechada: foram calibrados observando o comportamento em imagens concretas e ajustando na direção que corrigia o problema mais visível a cada rodada. Essa trajetória é informação sobre o comportamento do algoritmo, não alegação de otimalidade.

A primeira configuração permitia crescimento de até $4{,}0\times$ a área do núcleo, com limiar inicial $0{,}56$, piso $0{,}40$ e limiar de confiança $0{,}55$. O padrão de erro mais visível era *overgrowth* em baixo contraste, a máscara final ocupando fração grande demais da ROI. O ajuste foi tornar o algoritmo mais conservador em três frentes: reduzir o teto de $4{,}0\times$ para $1{,}70\times$; elevar o limiar inicial de $0{,}56$ para $0{,}62$ e o piso de $0{,}40$ para $0{,}50$ (exigir evidência mais forte no começo e no fim); elevar o limiar de confiança de $0{,}55$ para $0{,}65$, deslocando mais imagens de fronteira para revisão.

Em paralelo, os pesos de fusão foram reequilibrados para dar mais peso conjunto às evidências mais robustas a contraste, intensidade ($0{,}22 \rightarrow 0{,}25$), congruência de fase ($0{,}20 \rightarrow 0{,}25$) e gradiente ($0{,}15 \rightarrow 0{,}20$), enquanto distância caiu ($0{,}10 \rightarrow 0{,}03$) e continuidade também ($0{,}05 \rightarrow 0{,}02$). A leitura é que os termos de contenção espacial faziam menos trabalho útil do que os de estrutura e, com o teto de área já apertado ($1{,}70\times$), tornaram-se parcialmente redundantes; não se trata de ótimo obtido por busca. O *grid search* posterior (V3.3) não reabilitou o crescimento.

L. Correções incrementais do crescimento (V3.2 e V3.3)

A V3.2 introduziu ajustes manuais pontuais, sem abandonar a fila de prioridade: (1) *Fallback de confiança:* se a confiança da fronteira fica abaixo do limiar, devolve-se o núcleo original; (2) `MAX_GROWTH_FACTOR` reduzido de 1,70 para 1,30; (3) `MAX_DISTANCE_MULTIPLIER` reduzido de 3,0 para 1,5; (4) *fator de crescimento adaptativo* ao tamanho da lesão (lesões menores recebem teto intermediário entre um fator reduzido e o fator base), motivado pela relação perímetro/área $\propto 1/\sqrt{\text{área}}$ em formas aproximadamente circulares.

Na V3.2.1, identificou-se que, com fallback ativo, o núcleo ainda passava por redução e reexpansão (`MAX_SIZE`$= 1400$), gerando diferenças de arredondamento: em 16 das 27 imagens analisadas sem crescimento, o número de *pixels* final permanecia distinto do núcleo. A correção foi devolver o núcleo original diretamente, sem o redimensionamento.

A V3.3 acrescentou penalidade de curvatura, inspirada em *Region Competition* [21]. Em seguida, um *grid search* formal de $5 \times 5 \times 4 = 100$ combinações (curvatura, confiança e fator de crescimento) testou se alguma calibração do crescimento guloso superava o baseline. Foi esse resultado negativo, $0/100$ configurações positivas, que motivou a reorientação metodológica (Seção IV).

M. Distância em relação ao corte mínimo

A V3 é uma instanciação operacional de fusão multi-termo, não a arquitetura de *graph-cut* [17], [19], [23]. A busca gulosa não garante ótimo global e é sensível à ordem da fila. A ponderação adaptativa resume-se à circularidade reescalando o termo radial, sem pesos contínuos por posição. O *prior* estelar [19] é aproximado por um raio de contenção, sem visibilidade em linha reta no grafo. O termo de região é uma gaussiana robusta só do interior; o fundo não é modelado. Não há termo de convexidade parcial para microlobulação [22]. A triagem não completa essa arquitetura: muda a pergunta de pesquisa.

N. Pipeline de triagem final

A reorientação metodológica substitui a premissa "toda máscara precisa ser expandida" por um diagnóstico de direção. O classificador final é uma regressão logística (`LogisticRegression` do scikit-learn, `class_weight="balanced"`, `max_iter=1000`), com atributos padronizados por `StandardScaler` antes do ajuste. Um *Random Forest* raso foi avaliado na fase exploratória; o *pipeline* reportado usa exclusivamente a regressão logística.

Empregam-se 18 atributos calculáveis sem o contorno de referência na inferência: *forma*, `circularity`, `solidity`, `aspect_ratio`, `core_pixels_log`; *intensidade*, `core_intensity_mean`, `core_intensity_std`, `ring_intensity_mean`, `intensity_contrast_core_ring`; *confiança de fronteira*, `boundary_confidence_mean`, `boundary_confidence_std`, `ring_confidence_mean`, `core_confidence_mean`, `confidence_gradient_ring_minus_boundary`; *gradiente*, `boundary_gradient_mean`, `ring_gradient_mean`; *fase/radial*, `boundary_phase_mean`, `boundary_radial_mean`, `ring_phase_mean`. A referência entra apenas no alvo de treinamento, $\log(\text{area}_{\text{GT}}/\text{area}_{\text{baseline}})$, nunca como entrada em produção.

O alvo único do classificador é a *proxy* de área $\text{area}_{\text{baseline}}/\text{area}_{\text{GT}} > 1{,}30$ (50 positivos em 107; 46,7%). Essa proxy não coincide com a taxonomia diagnóstica por *precision/recall* (Seção IV): os 83% e as categorias BEM/SUPER/SUB/FALHA são descrições à parte.

O classificador foi treinado e avaliado por validação cruzada por paciente: a probabilidade p_{super} de cada imagem é sempre *out-of-fold*. No protocolo aninhado, o limiar HIGH é recalibrado no treino de cada fold externo (grade 0,30–0,90, passo 0,05; LOW = HIGH −0,20); os HIGH escolhidos foram 0,35; 0,35; 0,65; 0,40; 0,70. Com limiares fixos de referência 0,55/0,35:

- se $p_{\text{super}} \geq$ `HIGH`: `AUTO_ENCOLHER`, aplica-se encolhimento guiado por confiança;
- se $p_{\text{super}} <$ `LOW`: `MANTIDO`, devolve-se o núcleo sem alteração;
- caso contrário: `REVISAO_MANUAL`, preserva-se o núcleo e sinaliza-se a imagem a especialista.

O encolhimento separa quanto reduzir (pela confiança) de quais *pixels* manter (pela distância à borda). Seja N_{nuc} o número de *pixels* do núcleo na ROI. A confiança por *pixel* é

$$\begin{aligned}\text{conf} = {} & 0{,}30\,\text{fase} + 0{,}25\,\text{radial} \\ & + 0{,}20\,\text{gradiente} + 0{,}25\,\text{intensidade},\end{aligned} \tag{10}$$

cada mapa em $[0, 1]$. Define-se n_{hc} como o número de *pixels* do núcleo com conf $\geq 0{,}45$, $n_{\min} = \lfloor 0{,}60\, N_{\text{nuc}} \rfloor$ e

$$n_{\text{keep}} = \min\big(\max(n_{\min},\, n_{\text{hc}}),\, N_{\text{nuc}}\big). \tag{11}$$

Mantêm-se os n_{keep} *pixels* mais profundos do núcleo pela transformada de distância euclidiana (`scipy.ndimage.distance_transform_edt`).
Se a ROI excede 1 400 *px* no maior lado, a contagem ocorre na imagem reduzida (74/107 imagens). Exemplo (caso 22580244 do INbreast): $N_{\text{nuc}} = 20\,712$, $n_{\min} = 12\,427$, $n_{\text{hc}} = 16\,577$, $n_{\text{keep}} = 16\,577$. Nas 107 imagens, o piso $n_{\min}$ foi o ativo em 11 e a confiança em 96; a fração $n_{\text{keep}}/N_{\text{nuc}}$ teve média 0,802. O código original usava $n_{\text{keep}} = \max(n_{\min}, n_{\text{hc}})$ sem teto N_{nuc}; a execução atual acrescenta o teto. O fluxo completo está na Figura 2.

IV. Diagnóstico e reorientação metodológica

A investigação progrediu por etapas: diagnóstico estatístico da falha da V3.1, correções incrementais, busca em 100 combinações de hiperparâmetros e, por fim, a mudança de pergunta, de consertar o crescimento para diagnosticar sub versus supersegmentação.

Três evidências motivaram a reorientação metodológica: (i) 90,7% das 107 imagens pioraram sob crescimento sem *gate* ($p \approx 3{,}6 \times 10^{-17}$; $r \approx -0{,}95$); (ii) em análise de casos selecionados para investigar o comportamento do crescimento na V3.2, houve três aplicações efetivas, todas com redução do Dice; (iii) 0 de 100 combinações do *grid search* superaram o baseline nesta formulação de evidência.

O diagnóstico caso a caso mostrou a premissa implícita das versões de crescimento, toda máscara da nnU-Net precisa ser expandida, inadequada na maior parte das imagens (Figura 1, centro).

Em 89 de 107 imagens (83,2%) o baseline já está bem segmentado (BEM, 50) ou supersegmentado (SUPER, 39), de

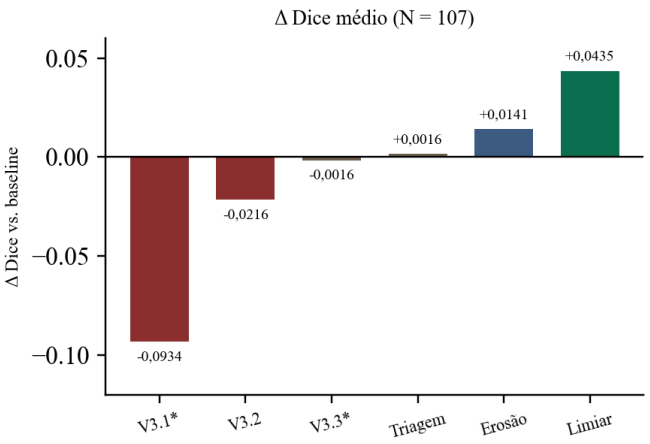

Fig. 3. Δ Dice médio vs. baseline ($N = 107$): V3.x, triagem, erosão e limiar (valores da revisão experimental).

modo que um crescimento universal é inadequado; apenas 9 (8,4%) são subsegmentadas (SUB) e 9 (8,4%) são falhas de detecção (Figura 1, centro). A regra operacional é: falha se Dice $< 0{,}30$; candidato a SUB se recall $< 0{,}85$; candidato a SUPER se precision $< 0{,}70$; se ambos estão baixos, prevalece o maior desvio ao limiar (SUPER se $(0{,}70 - \text{precision}) > (0{,}85 - \text{recall})$, senão SUB); caso contrário, BEM. Os totais 50/39/9/9 cobrem as 107 imagens. O alvo único do classificador permanece a *proxy* $\text{area}_{\text{baseline}}/\text{area}_{\text{GT}} > 1{,}30$ (50 positivos), distinta do 83%. Com a proxy de área, area $< 0{,}80$ seleciona 3/107 imagens (corte usado para conciliar a menção às três subsegmentadas por área). Nenhuma calibração de hiperparâmetro compensaria resolver o problema errado na maioria das instâncias. A triagem decide primeiro se deve corrigir e em que direção; os baselines simples testam se essa decisão supera correções triviais.

As versões V3.1–V3.3 tentavam, numa única passada, decidir se alterar, quanto alterar e como alterar espacialmente. O *pipeline* final separa essas responsabilidades (classificador, contagem de *pixels* de alta confiança, transformada de distância) e introduz o estado `REVISAO_MANUAL`. O classificador não refina a margem espiculada: estima a direção do erro residual.

V. Experimentos e resultados

A. Protocolo

A métrica principal é o Dice médio nas 107 imagens (50 pacientes), comparado ao baseline nnU-Net (Dice 0,7226). O protocolo principal é a CV aninhada por paciente; testes são Wilcoxon pareado bicaudal com `zero_method="wilcox"` (pares com $\Delta = 0$ descartados; n efetivo $= 51$ na triagem). IC 95% bootstrap do Δ: 10 000 reamostragens, semente 42. Tamanhos de efeito: rank-biserial pareado e d de Cohen pareado. Também se reportam Hausdorff (HD e HD95, borda a borda, em *pixels*; `pixel_spacing` ausente) para baseline, triagem, limiar e erosão, e estratificação espiculada versus

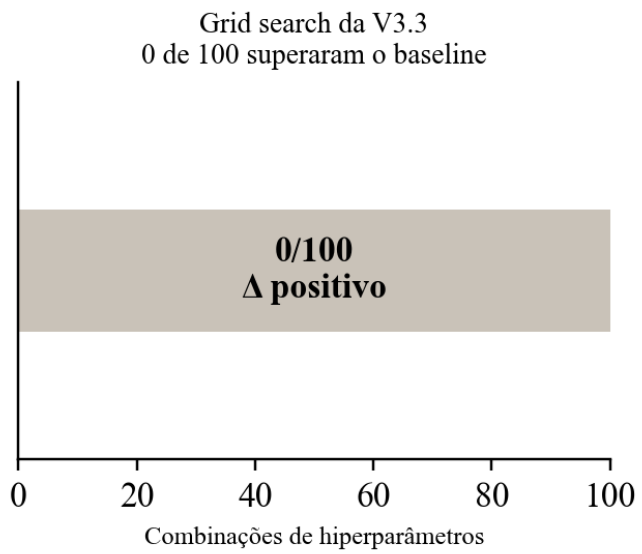

Fig. 4. *Grid search* da V3.3: $0/100$ combinações com Δ positivo. Os 100 deltas individuais não foram arquivados; a figura registra apenas o resultado agregado.

demais (XML do INbreast: 16 espiculadas / 91 não). *Falha catastrófica*: Dice pós-processamento zero quando o baseline era positivo; três imagens têm Dice 0 em qualquer limiar (50994408, 50997515, 51048738).

A Figura 5 e a Tabela I resumem baseline, limiar, erosão, V3.x e triagem.

B. V3.1: crescimento sem gate de confiança

Dice médio $0{,}6292$ ($\Delta = -0{,}0934$). Pioraram $90{,}7\%$ das 107 imagens ($p \approx 3{,}6 \times 10^{-17}$; $r \approx -0{,}95$). O efeito não é ruído: o algoritmo é estruturalmente enviesado para piorar. O viés está no desenho: cresce até 70% além do núcleo (`MAX_GROWTH_FACTOR` $= 1{,}70$, fixo, independente do tamanho da lesão), num raio de até $3\times$ o raio do núcleo, não penaliza fronteira irregular nem vazamento espicular, e ignora o próprio sinal de baixa confiança na hora de decidir o que devolver. A Figura 1 (esquerda) condensa a fração $90{,}7\%$; a Figura 3 compara os Δ médios disponíveis.

C. V3.2 / V3.2.1: fallback e crescimento adaptativo

Na configuração padrão (crescimento $1{,}70$; alcance $2{,}0\times$; confiança $0{,}65$ com fallback), sem recalibrar: Dice $0{,}7010$ (mediana $0{,}8160$); $\Delta = -0{,}0216$, IC $[-0{,}0337; -0{,}0111]$; $p = 2{,}0 \times 10^{-11}$; 8 melhoraram / 66 pioraram / 33 empataram (rank-biserial $-0{,}897$; $d = -0{,}358$). O fallback devolveu o núcleo em $96/107$ imagens; nas 11 aceitas o crescimento reduziu o Dice em todas (média $-0{,}1911$). A faixa $\sim$ 0,63–0,72 do manuscrito anterior misturava uma amostra de 30 casos (0,63) com a V3.3 (0,72) e é substituída pelo valor pontual. A correção V3.2.1 eliminou inconsistências de redimensionamento quando o fallback devolvia o núcleo. A confiança heurística da fronteira não se mostrou preditor confiável, o que levou ao classificador. A Figura 6 resume o fallback ($96/107$) e as 11 imagens em que o crescimento efetivo piorou o Dice.

D. V3.3 e grid search

Com penalidade de curvatura: Dice $0{,}7210$ ($\Delta = -0{,}0016$; $p = 0{,}018$). Falhas catastróficas foram neutralizadas (0 imagens), porém 0% de sucesso entre as imagens em que houve crescimento: o algoritmo deixou de destruir, mas também não melhorou quando ousava expandir. O *grid search* de 100 combinações ($5 \times 5 \times 4$ nos eixos de curvatura, limiar de confiança e fator de crescimento) produziu $0/100$ configurações positivas em relação ao baseline (Tabela I; Figura 4): nesta formulação de evidência, varrer o espaço de hiperparâmetros não superou o núcleo bruto.

E. Baselines simples e triagem

No mesmo esquema aninhado por paciente, um único limiar sobre a probabilidade da nnU-Net (escolhido no treino de cada fold; 0,97 ou 0,98) obtém Dice $0{,}7661$, $\Delta = +0{,}0435$, IC $[+0{,}0241; +0{,}0637]$, $p = 7{,}5 \times 10^{-6}$ (72/32/3). A erosão uniforme em raio fixo (disco; $r = 6$ *px* nos cinco folds) obtém Dice $0{,}7367$, $\Delta = +0{,}0141$, IC $[+0{,}0081; +0{,}0202]$, $p = 2{,}1 \times 10^{-5}$. A erosão por fração da área (0,90) rende $\Delta = +0{,}0097$. A triagem com HIGH recalibrado fica em Dice $0{,}7242$, $\Delta = +0{,}0016$, IC $[-0{,}0041; +0{,}0071]$, $p = 0{,}141$ (32 melhoraram / 19 pioraram / 56 empataram; rank-biserial $+0{,}237$; $d = +0{,}056$), efeito pequeno, IC que inclui zero. Comparada à erosão em raio fixo, a triagem perde $\Delta = -0{,}0124$ ($p = 0{,}0005$); ao limiar, $\Delta = -0{,}0418$ ($p = 1{,}9 \times 10^{-6}$). O número antigo da triagem ($\Delta = +0{,}0046$; $p = 0{,}0095$; limiares fixos nas mesmas 107) permanece só como resultado interno, com folds por imagem. Escolha metodológica: limiar e erosão entram como *baselines* de referência; a triagem é o resultado negativo central (Tabela I).

O Δ do limiar **não** tem validade externa demonstrada: as probabilidades vêm de uma rede avaliada nas próprias imagens de treino. O ótimo (0,97–0,98) está numa crista estreita (acima de 0,99 o Dice despenca), é específico deste modelo (probabilidades muito confiantes) e é plausível que reflita sobreconfiança *in-sample*, não um método generalizável. Por taxonomia, o ganho interno concentra-se em SUPER ($+0{,}1212$) e FALHA ($+0{,}0566$); em BEM é $-0{,}0035$ e em SUB, $-0{,}0458$.

F. Desempenho do classificador

AUC *out-of-fold* contra a proxy de área: $0{,}7147$ (em 10 partições de pacientes: 0,6923–0,7453). ECE $0{,}1118$; Brier $0{,}2266$, discriminação moderada, calibração fraca (`class_weight="balanced"` desloca as probabilidades). `AUTO_ENCOLHER` acerta quando o alvo $= 1$ (TP 32, FP 21, FN 18, TN 36; sensibilidade $0{,}640$; especificidade $0{,}632$; VPP $0{,}604$). `MANTIDO` acerta quando o alvo $= 0$ (20 de 26). `REVISAO_MANUAL` é faixa de incerteza, sem acerto definido.

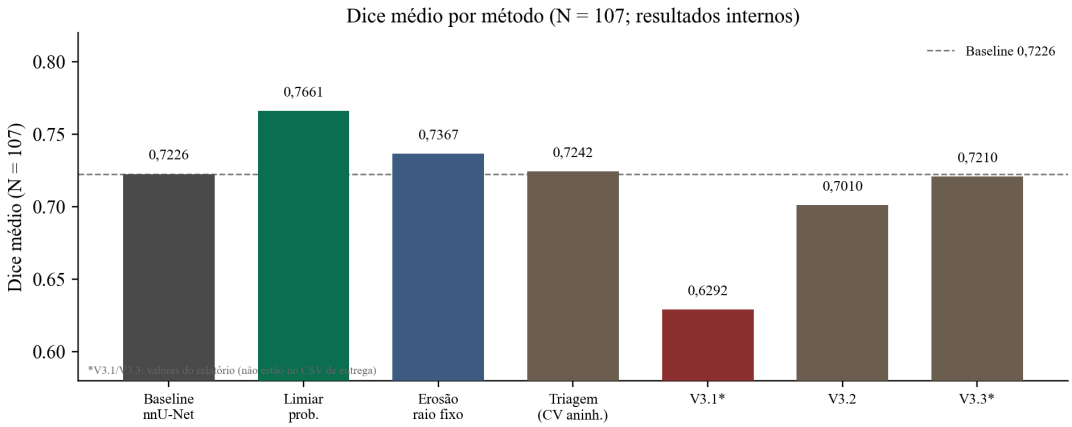

Fig. 5. Dice médio por método ($N = 107$): baseline, limiar, erosão, triagem e V3.x (Tabela I).

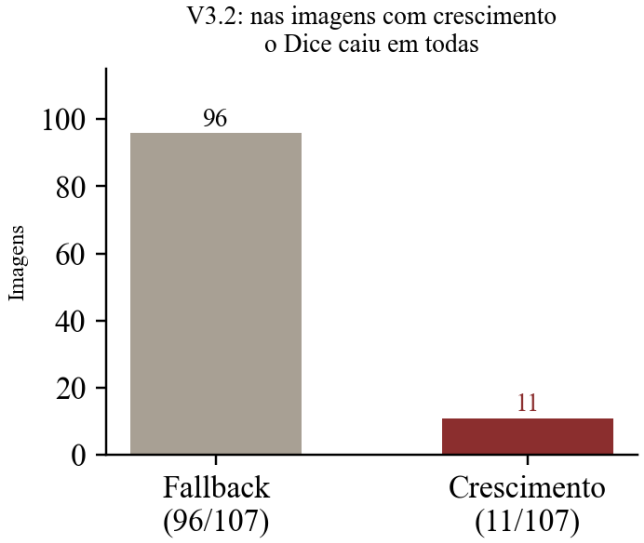

Fig. 6. V3.2 (configuração padrão): fallback em 96/107; nas 11 com crescimento o Dice caiu em todas.

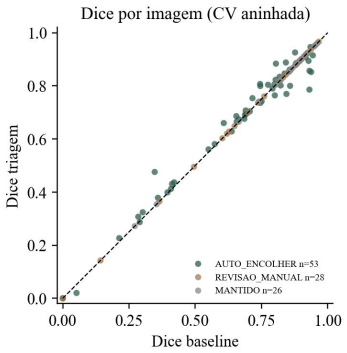

Fig. 7. Dice por imagem: baseline versus triagem (CV aninhada), colorido por destino (`AUTO_ENCOLHER`, `REVISAO_MANUAL`, `MANTIDO`).

G. Hausdorff e morfologia

Hausdorff (borda a borda, simétrico, em *pixels*; HD=máximo, HD95=percentil 95) foi calculado para baseline, triagem, limiar e erosão ($r = 6$ *px*) nas 107 imagens. Com *todas* as componentes, o HD é dominado por fragmentos distantes (baseline: média 1177,6, mediana 1269,1). Limiar e erosão o reduzem porque eliminam esses fragmentos: medianas 780,0 e 1021,3 ($p = 8{,}8 \times 10^{-12}$ e $8{,}4 \times 10^{-17}$); HD95 mediana 630,9 e 900,6. A triagem fica em 1217,1 (todas) e 68,1 (maior componente). Na *maior* componente (borda da massa principal), o ganho é pequeno ou nulo: limiar mediana HD 67,2 vs. 70,6 do baseline ($p = 0{,}565$); erosão 65,3 ($p = 0{,}020$ no HD, $p = 0{,}101$ no HD95). Por morfologia (XML: "Spiculated Region", 16 vs 91): o limiar ganha $+0{,}0868$ nas espiculadas e $+0{,}0359$ nas demais; a triagem não muda o Dice em nenhum dos dois grupos (espiculadas $\Delta = +0{,}0091$, $p = 0{,}139$; demais $+0{,}0003$, $p = 0{,}433$). Análise descritiva, n pequeno nas espiculadas (9 pacientes) e sem correção para múltiplas comparações (Figura 8).

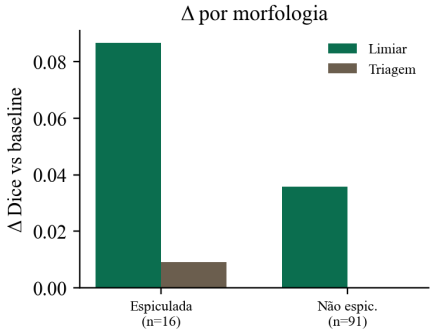

Fig. 8. Δ Dice por morfologia: limiar vs. triagem (espiculada $n = 16$; demais $n = 91$).

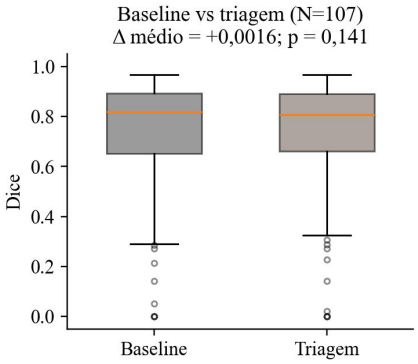

Fig. 9. Distribuição do Dice: baseline versus triagem aninhada ($N = 107$). Δ médio $= +0{,}0016$ ($p = 0{,}141$).

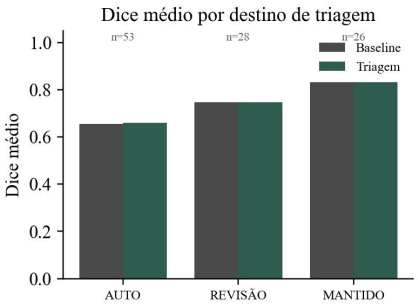

Fig. 10. Dice médio por destino de triagem (baseline vs. pós-triagem; $N = 107$).

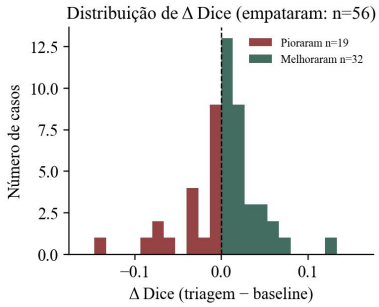

Fig. 11. Distribuição de Δ Dice (triagem $-$ baseline) na CV aninhada: melhoraram 32; pioraram 19; empataram 56.

a) Exemplos clínicos por destino.: As Figuras 12–14 ilustram referência, baseline (azul) e máscara pós-destino (verde). Em `REVISAO_MANUAL`/`MANTIDO` a máscara não muda; em `AUTO_ENCOLHER` o encolhimento atua. Os painéis de extremo ilustram a maior melhora e a maior piora observadas em `AUTO_ENCOLHER`; a estatística oficial permanece a da Tabela I.

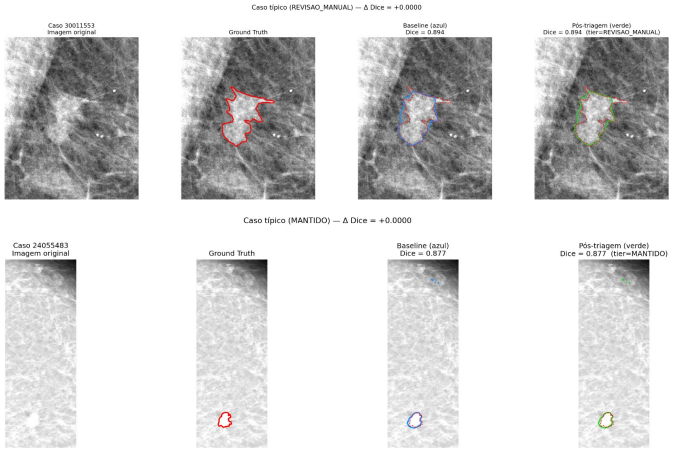

Fig. 12. Exemplos clínicos sem correção geométrica automática: `REVISAO_MANUAL` (acima, $\Delta = 0$) e `MANTIDO` (abaixo, $\Delta = 0$).

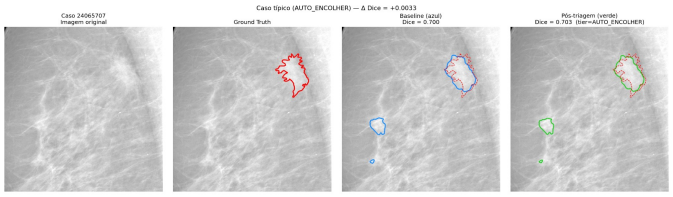

Fig. 13. Exemplo clínico (`AUTO_ENCOLHER`): referência, baseline (azul) e pós-triagem (verde).

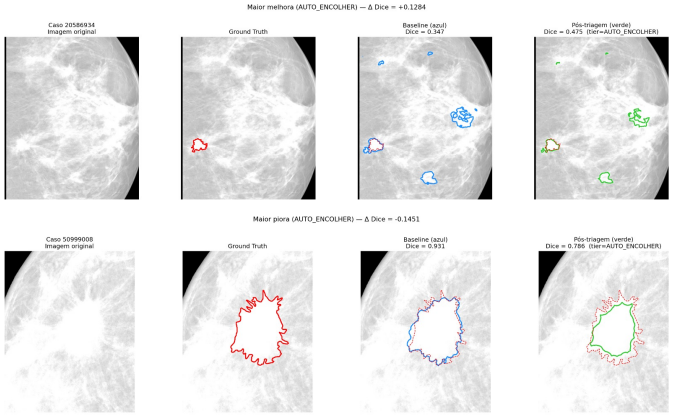

Fig. 14. Extremos de Δ Dice em `AUTO_ENCOLHER`: maior melhora ($+0{,}1287$; caso 20586934) e maior piora ($-0{,}1452$; caso 50999008).

VI. Discussão

Três leituras dos resultados não exigem números além dos medidos.

A. A pergunta precede o algoritmo de correção

A V3.1–V3.3 compartilhavam a premissa de que toda máscara da nnU-Net deveria ser expandida. Essa premissa

era falsa na maior parte das imagens (Seção IV). O *grid search* (0/100) é a evidência empírica de que calibração do crescimento guloso não corrige o problema errado. A triagem decide primeiro se deve corrigir e em que direção, mas, nesta base e neste modelo, não supera baselines triviais.

B. Três responsabilidades

As versões de crescimento misturavam, numa única passada, se alterar, quanto alterar e como alterar. A triagem separa: o classificador decide se e quando (com zona de incerteza), n_{hc} decide quanto, e a transformada de distância decide a geometria. O piso de 60% do núcleo impede zerar a máscara por encolhimento.

C. Baselines simples vencem a triagem

O resultado central é negativo para a triagem: $\Delta = +0{,}0016$, IC que inclui zero (Tabela I). Limiar ($\Delta = +0{,}0435$) e erosão ($+0{,}0141$) superam a triagem *neste* corpus interno, o que basta para invalidar a reivindicação de ganho do classificador de direção de erro; não basta para validar o limiar como método. O classificador (AUC 0,71, ECE 0,11) não calibra bem a decisão. A distribuição de Δ (Figura 11) mostra 32/19/56.

D. Relação com a literatura

O fracasso da V3.1 é coerente com o teto relatado em [13] para margens mal definidas e com o fato de que contraste local não distingue espícula de fibra [11], [12]. O núcleo da nnU-Net já é, na maior parte das imagens, uma solução que não pede expansão (Dice 0,7226 interno). O Δ interno do limiar nas espiculadas ($+0{,}0868$) é coerente com superssegmentação nessas margens, sob a mesma ressalva *in-sample*.

E. Implicações para saúde digital

O trabalho ilustra avaliação honesta de pós-processamento: a nnU-Net fornece o núcleo; baselines simples devem ser reportados antes de reivindicar ganho de um classificador de direção de erro; a zona `REVISAO_MANUAL` preserva supervisão humana. O valor principal é o deslocamento da pergunta, de "como expandir sempre" para "em que direção corrigir, se for o caso", e a evidência negativa de que a triagem não agrega, nestas 107 imagens, além de correções triviais internas.

VII. Limitações

Todos os números sobre as predições são *internos*: a nnU-Net viu as 107 imagens no treino (fold 0 por imagem; 20/22 da validação têm paciente no treino). A CV aninhada por paciente corrige a calibração da triagem, não a dependência do baseline. Não há conjunto de teste externo nem retreino da rede. O limiar ótimo (0,97–0,98) está numa crista estreita, é específico deste modelo e pode ser artefato de sobreconfiança *in-sample*; não se reivindica validade externa desse ganho. A contagem de 108 pacientes (versus 115 casos em [5]) refere-se à recontagem por identificador DICOM na base local. A estratificação morfológica é descritiva ($n = 16$ espiculadas; 9 pacientes) e o rótulo indica região espiculada na imagem, não necessariamente na massa da máscara. Pesos da V3 vêm de calibração manual; a busca gulosa não garante ótimo global. Versões de biblioteca diferem levemente entre sessões Colab. O uso da base pública INbreast apoia-se nas autorizações documentadas em [5]; não há protocolo CEP brasileiro específico. A generalização a outros equipamentos permanece aberta. O corte original das "3 subsegmentadas por área" não foi localizado no código; $\mathrm{area} < 0{,}80$ reproduz a contagem e é tratado como inferido. Os p-valores de Hausdorff não têm correção para múltiplas comparações.

a) Uso de inteligência artificial (Edital 001/2026, item 11.3).: A nnU-Net v2 2D (Dataset002_INBreast16, fold 0) produz o núcleo; a regressão logística estima p_{super} a partir de 18 atributos sem a referência na inferência. A avaliação principal é CV aninhada por paciente ($N = 107$ imagens, 50 pacientes), com zona de revisão humana (`REVISAO_MANUAL`). O sistema é proposto como apoio à segmentação, não como substituto de leitura radiológica.

VIII. Trabalhos futuros

Caminhos imediatos: retreinar a nnU-Net com partição por paciente e conjunto de teste externo; avaliar se o limiar de probabilidade permanece superior fora deste modelo; alinhar o alvo do classificador à taxonomia por *precision/recall*; calibrar p_{super} (ECE atual 0,11). Quando houver janela de retreino, cabe *fine-tuning* com Tversky, Focal Tversky e Boundary Loss [2]–[4]. O resultado 0/100 e o Δ nulo da triagem desencorajam insistir no crescimento guloso e no encolhimento seletivo nesta formulação, frente a baselines triviais.

IX. Conclusão

Sob a restrição de não retreinar a nnU-Net, o crescimento guloso multi-evidência não recuperou a margem (V3.1: 0,7226 $\rightarrow$ 0,6292; V3.2: 0,7010, $\Delta = -0{,}0216$), e 100 combinações de hiperparâmetros não reverteram esse quadro. A premissa de expansão universal era inadequada em 83,2% das imagens (BEM+SUPER = 89/107). Em CV aninhada por paciente, a triagem não supera o baseline ($\Delta = +0{,}0016$; $p = 0{,}141$) e fica abaixo de limiar e erosão usados como baselines internos. O Δ do limiar ($+0{,}0435$) não tem validade externa demonstrada. Todos esses números são resultados internos.

A contribuição é a evidência negativa do crescimento guloso e da triagem por direção de erro, com formalização de n_{hc} e baselines triviais que delimitam o que um pós-processamento precisa superar antes de ser reivindicado.

TABELA I

COMPARAÇÃO ($N = 107$ IMAGENS, 50 PACIENTES). LIMIAR, EROSÃO E TRIAGEM: CV ANINHADA POR PACIENTE. p: WILCOXON PAREADO. RESULTADOS INTERNOS (NNU-NET TREINADA NESTAS IMAGENS).

Método	**Dice**	Δ	p	**Notas**
Baseline nnU-Net	0,7226	—	—	interno
Limiar probabilidade	0,7661	+0,0435	$7{,}5 \times 10^{-6}$	0,97/0,98
Erosão raio fixo	0,7367	+0,0141	$2{,}1 \times 10^{-5}$	$r = 6$ *px*
Triagem (HIGH rec.)	0,7242	+0,0016	0,141	32/19/56
V3.1	0,6292	−0,0934	$3{,}6 \times 10^{-17}$	90,7% pioraram
V3.2	0,7010	−0,0216	$2{,}0 \times 10^{-11}$	pontual
V3.3	0,7210	−0,0016	0,018	0% sucesso
Grid (V3.3)	—	0/100 pos.	—	—

REFERÊNCIAS

[1] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.

[2] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *MLMI 2017 (MICCAI Workshop)*, ser. LNCS, vol. 10541. Springer, 2017, pp. 379–387.

[3] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *IEEE ISBI*, 2019, pp. 683–687.

[4] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed, "Boundary loss for highly unbalanced segmentation," in *MIDL*, ser. PMLR, vol. 102, 2019, pp. 285–296.

[5] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[6] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," arXiv:1804.03999, 2018.

[7] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[8] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[9] P. Marquez-Neila, L. Baumela, and L. Alvarez, "A morphological approach to curvature-based evolution of curves and surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 2–17, 2014.

[10] S. Xu, H. Liu, and E. Song, "Marker-controlled watershed for lesion segmentation in mammograms," *Journal of Digital Imaging*, vol. 24, no. 5, pp. 754–763, 2011.

[11] M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Transactions on Medical Imaging*, vol. 17, no. 4, pp. 510–517, 1998.

[12] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammograms," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 611–619, 1996.

[13] Y. Tao, S.-C. B. Lo, M. T. Freedman, E. Makariou, and J. Xuan, "Multilevel learning-based segmentation of ill-defined and spiculated masses in mammograms," *Medical Physics*, vol. 37, no. 11, pp. 5993–6002, 2010.

[14] P. Kovesi, "Image features from phase congruency," *Videre: Journal of Computer Vision Research*, vol. 1, no. 3, pp. 1–26, 1999.

[15] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001.

[16] X. Yang, D. Zhang *et al.*, "Edge detection methods based on differential phase congruency of monogenic image," arXiv:1607.00097, 2016.

[17] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *IEEE ICCV*, vol. 1, 2001, pp. 105–112.

[18] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[19] O. Veksler, "Star shape prior for graph-cut image segmentation," in *ECCV*, ser. LNCS, vol. 5304, 2008, pp. 454–467.

[20] H. Wang, H. Zhang, and N. Ray, "Adaptive shape prior in graph cut image segmentation," *Pattern Recognition*, vol. 46, no. 5, pp. 1409–1414, 2013.

[21] S. C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 884–900, 1996.

[22] D. C. Siu, T. Chan, and R. Lui, "Image segmentation with partial convexity shape prior using discrete conformality structures," *SIAM Journal on Imaging Sciences*, vol. 13, no. 4, pp. 2105–2139, 2020.

[23] L. R. Ford and D. R. Fulkerson, *Flows in Networks*. Princeton University Press, 1962.