COMPARAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DE LESÕES CANCERÍGENAS EM IMAGENS DE ULTRASSOM MAMÁRIO

Leonardo Oenning Spilere[1], Marlon de Matos de Oliveira[2]

Resumo: O câncer de mama é a neoplasia de maior incidência entre mulheres no Brasil, com mais de 73.600 novos casos entre 2023-2025, e o diagnóstico precoce pode elevar as taxas de cura a até 90%. A ultrassonografia mamária, embora acessível e não invasiva, apresenta limitações de variabilidade interpretativa, motivando a investigação de ferramentas computacionais de apoio ao diagnóstico. Este trabalho desenvolveu e avaliou comparativamente modelos de inteligência artificial para detecção de lesões potencialmente cancerígenas em imagens de ultrassonografia mamária, em duas trilhas complementares: classificação binária, com redes neurais profundas, ResNet50, DenseNet121 e EfficientNet-B7, e métodos clássicos, SVM-RBF, LightGBM e Random Forest; e segmentação semântica, com U-Net e Attention U-Net. Os experimentos foram conduzidos sobre o conjunto público BUS-BRA, com 1.875 imagens de 1.064 pacientes, e o desempenho avaliado por acurácia, sensibilidade, especificidade, F1-score e AUC-ROC na classificação, e por coeficiente Dice e IoU na segmentação, com comparações por *bootstrap*, McNemar e DeLong. O ResNet50 obteve o melhor desempenho em classificação (AUC-ROC de 0,886; acurácia de 0,848), possivelmente equivalente ao DenseNet121 ($p > 0{,}05$) e superior aos demais modelos ($p < 0{,}05$). A Attention U-Net superou o U-Net clássico em segmentação (Dice de 0,914; IoU de 0,841). Os resultados evidenciam o potencial dessas arquiteturas como ferramentas de apoio ao diagnóstico precoce do câncer de mama.

Palavras-chave: câncer de mama; redes neurais convolucionais; aprendizado de máquina; segmentação de imagens médicas; diagnóstico assistido por computador.

[1]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. leonardo.spilere@unesc.net

[2]Orientador, Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. marlon.oliveira@unesc.net

Abstract: Breast cancer is the most prevalent neoplasia among women in Brazil, with more than 73,600 new cases between 2023-2025, and early diagnosis can raise cure rates to up to 90%. Breast ultrasonography, while accessible and non-invasive, presents interpretive variability limitations, motivating the investigation of computational tools to support diagnosis. This work developed and comparatively evaluated artificial intelligence models for the detection of potentially cancerous lesions in breast ultrasound images, across two complementary pipelines: binary classification, with deep neural networks, ResNet50, DenseNet121, and EfficientNet-B7, and classical methods, SVM-RBF, LightGBM, and Random Forest; and semantic segmentation, with U-Net and Attention U-Net. Experiments were conducted on the public BUS-BRA dataset, comprising 1,875 images from 1,064 patients, and performance was assessed by accuracy, sensitivity, specificity, F1-score, and AUC-ROC for classification, and by Dice coefficient and IoU for segmentation, with comparisons via bootstrap, McNemar, and DeLong tests. ResNet50 achieved the best classification performance (AUC-ROC of 0.886; accuracy of 0.848), being possibly equivalent to DenseNet121 ($p > 0.05$) and superior to all other models ($p < 0.05$). Attention U-Net outperformed classical U-Net in segmentation (Dice of 0.914; IoU of 0.841). The results demonstrate the potential of these architectures as tools to support early breast cancer diagnosis.

Keywords: breast cancer; convolutional neural networks; machine learning; medical image segmentation; computer-aided diagnosis.

1 INTRODUÇÃO

O câncer de mama representa a neoplasia de maior incidência entre mulheres no Brasil, com estimativa de mais de 73.600 novos casos para o triênio de 2023 a 2025, correspondendo a uma taxa de risco de 66,54 casos a cada 100 mil mulheres (INCA, 2024). O diagnóstico precoce é determinante para a redução da mortalidade, uma vez que, ao identificar lesões em estágio inicial, as taxas de cura podem atingir até 90% (FEBRASGO, 2025). Nesse contexto, a ultrassonografia mamária configura-se como ferramenta complementar relevante, especialmente em mulheres com mamas densas ou com achados inconclusivos na mamografia, por seu caráter não invasivo e sua ampla acessibilidade (Santos; Dias, 2015). Entretanto, a interpretação de imagens ultrassonográficas permanece fortemente dependente da experiência do profissional, o que pode introduzir variabilidade

diagnóstica e comprometer a acurácia dos resultados.

Diante desse cenário, técnicas de inteligência artificial, em especial o aprendizado profundo, têm sido investigadas como estratégia para apoiar e aprimorar o diagnóstico por imagem. Revisões recentes indicam que a integração de algoritmos de aprendizado profundo e sistemas de diagnóstico assistido por computador com a avaliação humana pode aumentar a acurácia, reduzir falsos positivos e otimizar o tempo de análise (Serrano et al., 2025). Latha et al. (2024) demonstraram que a arquitetura EfficientNet-B7, combinada a técnicas de IA explicável, pode alcançar acurácia de 99,14% no conjunto de dados BUSI para classificação de lesões mamárias em imagens de ultrassom, superando abordagens anteriores e evidenciando o potencial das arquiteturas modernas nessa tarefa.

Nastase et al. (2024) combinaram as arquiteturas MobileNetV2, VGG16 e EfficientNet-B7 com operações morfológicas de erosão e dilatação para extração de características intra e perilesionais, obtendo AUC de até 1,00 em alguns cenários e acurácia superior a 94% em conjunto de teste externo, demonstrando que informações do tecido ao redor da lesão contribuem de forma relevante para a classificação.

Foleis et al. (2025) avaliaram combinações de redes pré-treinadas, InceptionV3, EfficientNetB4, ResNet50 e VGG16, com descritores radiômicos classificados por SVM e integrados por comitês de fusão tardia, alcançando F1-score de 83,90% no conjunto BUS-BRA e 88,70% no BUSI, evidenciando o potencial da combinação entre aprendizado por transferência e características manuais para o diagnóstico assistido por computador.

Diniz et al. (2024) propuseram o EfficientEnsemble, um ensemble de EfficientNets aplicado à ultrassonografia mamária com processamento de imagens, atingindo acurácia de 96,7% e AUC de aproximadamente 0,96, reforçando que estratégias de combinação de modelos produzem resultados mais robustos do que arquiteturas isoladas.

Sulaiman et al. (2024) aplicaram a arquitetura Attention U-Net ao conjunto BUSI para segmentação de tumores, obtendo acurácia de 98% e coeficiente Dice de 0,92, demonstrando que mecanismos de atenção integrados às arquiteturas de segmentação permitem identificar com precisão as regiões de interesse em imagens de ultrassom mamário.

Varshney, Verma e Kaur (2025) propuseram o modelo AG-MRUNet, uma variante da arquitetura MultiResUNet com mecanismos de atenção espacial e por canais, alcançando acurácia de 98,28% no BUSI e 99,25% no BUS-BRA com coeficiente Dice de até 99,80%, além de tempo de trei-

namento reduzido, evidenciando o potencial de arquiteturas com atenção como ferramentas eficientes para apoio ao diagnóstico precoce.

No contexto brasileiro, Filho et al. (2024) desenvolveram um classificador híbrido integrando LightGBM, MLP e SVM com otimização por enxame de partículas, atingindo acurácia de 98% em imagens de ultrassom mamário, demonstrando que abordagens híbridas com métodos clássicos de aprendizado de máquina seguem competitivas, especialmente quando a base de dados é limitada.

Vedova e Ruaro (2024) compararam CNN, SVM e SVM combinada ao MobileNetV2 para análise de imagens citológicas de punção aspirativa por agulha fina da tireoide, com a combinação SVM e MobileNetV2 alcançando sensibilidade de 99% na detecção de casos malignos, demonstrando o potencial da IA como ferramenta de apoio em análises citológicas em contexto institucional local, publicado no Repositório Institucional da UNESC (Universidade do Extremo Sul Catarinense).

Em contexto histórico e relevante, Parise (2003) desenvolveu o sistema BreastSystem, que integrou processamento digital de imagens e redes neurais do tipo Feedforward treinadas com o algoritmo Backpropagation para identificação de microcalcificações em mamografias digitais, sinalizando precocemente o potencial das redes neurais no apoio ao diagnóstico oncológico.

Apesar dos avanços observados na literatura, ainda são escassos os estudos brasileiros que realizam comparações sistemáticas e controladas entre múltiplas abordagens de IA aplicadas especificamente à ultrassonografia mamária (Dallagassa; Oliveira, 2024). A maioria dos trabalhos nacionais concentra-se em modalidades isoladas ou em revisões de literatura, sem conduzir experimentos comparativos com métricas padronizadas e testes entre modelos distintos. Adicionalmente, questões como interpretabilidade dos modelos, robustez frente ao desequilíbrio de classes e validação em bases de dados de origem brasileira ainda carecem de investigação aprofundada.

Para contribuir com o preenchimento dessas lacunas, este trabalho propõe o desenvolvimento e a avaliação comparativa de modelos de inteligência artificial aplicados à detecção de células potencialmente cancerígenas em imagens de ultrassonografia mamária, integrando arquiteturas de classificação como EfficientNet-B7, ResNet e DenseNet, modelos de segmentação baseados em U-Net com mecanismos de atenção, e métodos clássicos de aprendizado de máquina como SVM, LightGBM e florestas aleatórias. O desempenho dos modelos é avaliado por meio de acurácia,

sensibilidade, especificidade, precisão, F1-score e AUC-ROC, com comparações entre as abordagens.

A relevância desta pesquisa se justifica, portanto, tanto pelo impacto clínico quanto pelo caráter técnico-científico que proporciona. Do ponto de vista social, o desenvolvimento de modelos computacionais confiáveis para apoio ao diagnóstico pode contribuir para a democratização do acesso a exames de maior qualidade, especialmente em regiões com escassez de especialistas em radiologia mamária. Do ponto de vista científico, a comparação rigorosa entre diferentes arquiteturas e abordagens, com métricas padronizadas e análise, oferece subsídios concretos para orientar escolhas metodológicas em pesquisas futuras na área de computação aplicada à saúde. Adicionalmente, ao explorar aspectos de interpretabilidade e robustez dos modelos, o estudo contribui para uma abordagem mais segura, contextualizada e eticamente fundamentada da inteligência artificial no cuidado oncológico.

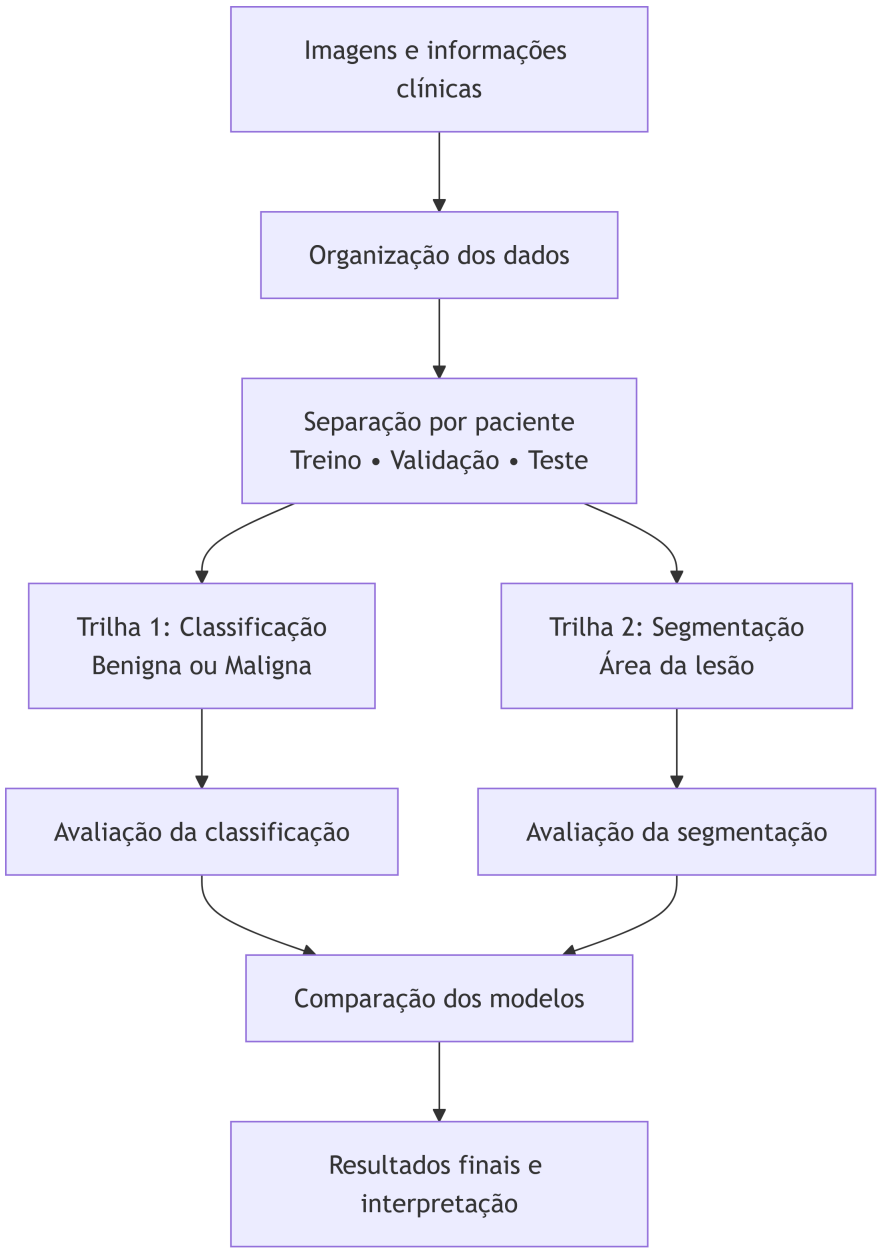
O trabalho está organizado da seguinte forma: a Seção 2 reúne os materiais e métodos utilizados no desenvolvimento deste trabalho; a Seção 3 apresenta os resultados obtidos; e a Seção 4 apresenta as considerações finais.

2 MATERIAIS E MÉTODOS

Trata-se de uma pesquisa de natureza aplicada, com abordagem quantitativa, objetivos descritivos e comparativos, e procedimentos de caráter experimental. O estudo desenvolveu e avaliou comparativamente modelos de inteligência artificial voltados à detecção de lesões potencialmente cancerígenas em imagens de ultrassonografia mamária, por meio de experimentos computacionais controlados, nos quais variáveis, protocolos e métricas de desempenho foram previamente definidos para assegurar comparabilidade e reprodutibilidade. O fluxo experimental foi organizado em duas trilhas complementares: (i) classificação binária de lesões (*benign* = 0; *malignant* = 1), que estima a categoria diagnóstica da imagem; e (ii) segmentação binária, que delimita pixel a pixel a região da lesão, conforme a Figura 1.

Em termos práticos, a classificação responde *qual é a classe da lesão*, enquanto a segmentação responde *onde está a lesão*. As duas trilhas compartilham a mesma base de imagens e o mesmo protocolo de divisão por paciente, para garantir comparabilidade metodológica e reduzir risco de viés por vazamento de informações entre conjuntos.

Figura 1 - Diagrama de estrutura do projeto



Fonte: Elaborado pelo autor.

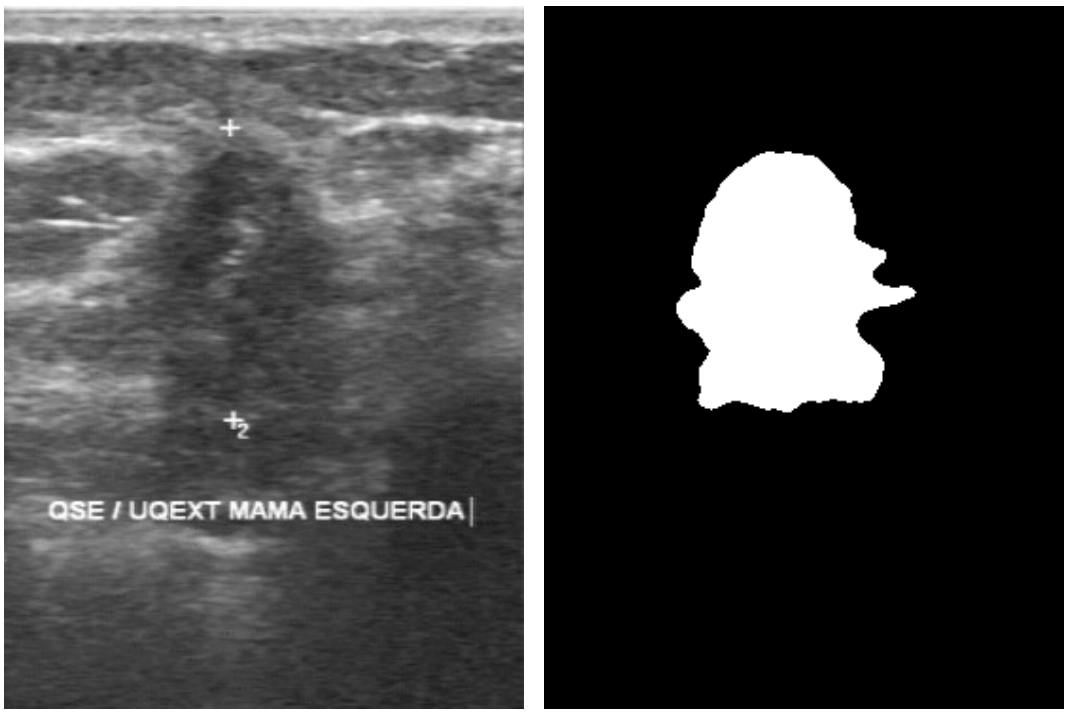
2.1 BASE DE DADOS

Os dados utilizados neste estudo foram obtidos do conjunto público *Breast Ultrasound Dataset for Assessing Computer-aided Diagnosis Systems* (BUS-BRA)[3], desenvolvido por Gómez-Flores, Gregorio-Calas e Pereira (2024)no Programa de Engenharia Biomédica da Universidade Federal do Rio de Janeiro (PEB/COPPE-UFRJ) em colaboração com o Instituto

[3]Conjunto de dados de acesso aberto disponível em: <https://doi.org/10.5281/zenodo.8231412>. Os códigos-fonte para replicação dos experimentos originais estão disponíveis em: <https://github.com/wgomezf/BUS-BRA>.

Nacional de Câncer (Rio de Janeiro, Brasil). O BUS-BRA é um repositório de acesso aberto que reúne imagens de ultrassonografia mamária anonimizadas, com anotações de segmentação de lesões, resultados de biópsia comprovada e categorias BI-RADS (2, 3, 4 e 5), além de partições padronizadas de treino, validação e teste para garantir reprodutibilidade e comparações justas entre diferentes abordagens de diagnóstico assistido por computador conforme o par de imagens na Figura 2.

Figura 2 - Imagem de ultrassom do *dataset* e respectiva segmentação



(a) Ultrassom: *bus_0003-l.png* (b) Máscara: *mask_0003-l.png*

Fonte: Gómez-Flores, Gregorio-Calas e Pereira (2024).

O conjunto totalizou 1.875 imagens provenientes de 1.064 pacientes do sexo feminino, adquiridas por meio de quatro equipamentos de ultrassom distintos, com 1.268 lesões classificadas como benignas e 607 como malignas. As imagens estão disponíveis no formato PNG, organizadas em *data/raw*, com máscaras de segmentação correspondentes em *data/masks* e referência clínica consolidada em *reference/bus_data.csv*. Essa organização permite rastrear, para cada imagem, o paciente de origem, o rótulo clínico e a máscara da lesão associada.

Para estruturar os dados de entrada, foi gerado o arquivo *data/metadata.csv* a partir da varredura dos arquivos reais. O *patient_id* foi extraído do padrão de nomenclatura do BUS-BRA (*bus_XXXX-Y.png*), em que *XXXX* representa o identificador do paciente e *Y* o índice da imagem. A

coluna *Pathology* foi convertida para rótulo binário (0 = benigno; 1 = maligno), e a máscara correspondente foi associada pela convenção *mask_XXXX-Y.png*. Em seguida, foram gerados os arquivos *train.csv*, *val.csv* e *test.csv*.

A divisão dos dados foi estratificada por paciente na proporção 70/15/15 (treino/validação/teste), respeitando as partições padronizadas do BUS-BRA. O conjunto de treino reuniu 1.323 imagens de 744 pacientes, o de validação 270 imagens de 160 pacientes e o de teste 282 imagens de 160 pacientes. A estratificação por paciente evita *data leakage*, impedindo que imagens do mesmo paciente apareçam em partições distintas e evitando superestimação de desempenho.

No pré-processamento, as imagens em escala de cinza foram convertidas para RGB em memória, etapa necessária para compatibilidade com *backbones* pré-treinados em *ImageNet*. Em seguida, foram redimensionadas para 224×224 *pixels* e normalizadas pelos valores de média e desvio-padrão do *ImageNet*. Durante o treino de classificação, foram aplicadas transformações de aumento de dados (*data augmentation*), incluindo espelhamento horizontal ($p = 0,5$) e rotação aleatória (±10°); os conjuntos de validação e teste permaneceram determinísticos.

2.2 MODELAGEM

As tecnologias foram selecionadas por maturidade, reprodutibilidade e aderência ao problema: *PyTorch* e *timm* para redes profundas de classificação, *segmentation-models-pytorch* para arquiteturas de segmentação, e *scikit-learn*/*LightGBM* para modelos clássicos.

Na trilha de classificação profunda, os quatro modelos foram treinados separadamente, em execuções independentes, conforme a Tabela 1. Para manter comparabilidade, todos seguiram o mesmo protocolo: otimizador *Adam* (método de ajuste dos pesos por gradiente adaptativo), *learning rate* de 1×10^{-4}, *weight decay* de 1×10^{-4} (regularização para reduzir sobreajuste), *scheduler CosineAnnealingLR* (redução suave da taxa de aprendizado), *batch size* de 16 e até 10 épocas. Foi aplicado *early stopping* com paciência 10, selecionando o melhor modelo pela maior *AUC-ROC* em validação. A função de perda foi *CrossEntropyLoss* com ponderação de classes.

Tabela 1 - Treinamento separado dos modelos de classificação profunda.

Modelo	**Descrição do treino**
baseline_stub	Classificador mínimo (pooling global + camada linear), treinado do zero como linha de base.
ResNet50	Rede residual profunda treinada para duas classes na configuração padrão da arquitetura.
DenseNet121	Rede com conectividade densa entre camadas, ajustada para duas classes.
EfficientNet-B7	Rede com escalonamento composto de profundidade/largura/resolução,treinada com a arquitetura completa.

Fonte: Elaborado pelo autor.

Na trilha clássica, a extração de características utilizou um *encoder ResNet50* pré-treinado(CNN profunda) apenas como extrator, com a saída global *pooled* (vetor resumido da imagem). Sobre esses vetores, foram treinados classificadores clássicos *SVM-RBF*, *LightGBM* e *Random Forest*. A busca de hiperparâmetros foi feita por *GridSearchCV* com *StratifiedKFold* (k=5), usando *scoring = roc_auc*.

Na trilha de segmentação, foram treinadas *U-Net* e *Attention U-Net* (com atenção *SCSE*), ambas com *encoder ResNet34*. A função de perda combinou *BCEWithLogits* e *Dice loss*, unindo estabilidade probabilística com sobreposição espacial. A binarização das máscaras preditas utilizou limiar de 0,5; a seleção do melhor modelo foi feita pelo maior *Dice* em validação, com avaliação final no conjunto de teste.

O ambiente computacional de treino e avaliação utilizou AMD Ryzen 9 5900X, 32 GB RAM e Nvidia GeForce RTX 5070 Ti (16 GB VRAM), executando Arch Linux em WSL2 sobre Windows 11, com aceleração por CUDA via suporte de GPU no WSL2. O desenvolvimento e versionamento foram conduzidos em notebook Dell i5-1235U, 16 GB RAM, Intel Iris Xe, sob Omarchy Linux 3.4.

As dependências principais foram *Python* 3.12.9 e, conforme o arquivo *requirements.txt*, as bibliotecas *torch* (>=2.11.0), *torchvision* (>=0.21.0), *timm* (>=0.9.0), *segmentation-models-pytorch* (>=0.3.3), *scikit-learn* (>=1.3.0), *lightgbm* (>=4.1.0), *numpy* (>=1.24.0), *pandas* (>=2.0.0), *scipy* (>=1.11.0), *opencv-python* (>=4.8.0), *Pillow* (>=12.2.0), *matplotlib* (>=3.7.0), *seaborn* (>=0.12.0), *PyYAML* (>=6.0) e *tqdm* (>=4.65.0).

2.3 AVALIAÇÃO E MÉTRICAS

Na classificação, a avaliação no conjunto de teste incluiu acurácia (proporção global de acertos), precisão (confiabilidade dos positivos preditos), *recall*/sensibilidade (capacidade de identificar casos positivos), especificidade (capacidade de identificar casos negativos), *F1-score* (média harmônica entre precisão e *recall*), *AUC-ROC* (capacidade de discriminação entre classes ao longo de diferentes limiares) e matriz de confusão (distribuição de acertos e erros por classe).

Na segmentação, foram medidos *Dice* (sobreposição entre máscara predita e máscara real), *IoU* (*Intersection over Union*), sensibilidade por pixel e especificidade por pixel. Essas métricas quantificam tanto a qualidade da sobreposição espacial quanto o comportamento do modelo em termos de falsos positivos e falsos negativos.

A comparação inferencial entre classificadores foi realizada com *bootstrap* não paramétrico ($n = 1000; \alpha = 0{,}05$) para estimar intervalos de confiança, teste de *McNemar* para diferença pareada de acerto/erro entre modelos, e teste de *DeLong* para diferença pareada de *AUC*. A interpretação dos resultados considerou variação significante ($p < 0,05$), magnitude do efeito (diferença entre métricas) e incerteza (largura dos intervalos), buscando evidência comparativa robusta entre abordagens profundas, clássicas e de segmentação.

3 RESULTADOS E DISCUSSÃO

Os resultados de classificação e segmentação no conjunto de teste são apresentados nas Tabelas 2 e 3. Em classificação, o melhor desempenho global foi obtido pelo ResNet50 (AUC-ROC de 0,886; acurácia de 0,848), seguido por DenseNet121 (AUC-ROC de 0,884). Entre os modelos clássicos, o LightGBM apresentou o melhor equilíbrio (AUC-ROC de 0,823; F1 de 0,573). Em segmentação, o Attention U-Net superou o U-Net clássico em todas as métricas reportadas.

As curvas ROC consolidadas estão apresentadas na Figura 3. As matrizes de confusão em formato de figura estão apresentadas na Figura 4, com um painel para cada modelo de classificação. As matrizes numéricas também foram extraídas dos arquivos de predição (por exemplo, ResNet50: TN=173, FP=21, FN=22, TP=66).

Tabela 2 - Comparação de métricas dos modelos de classificação.

Modelo	Acurácia	Sensib.	Especif.	Precisão	F1	AUC-ROC
Modelos profundos						
baseline_stub	0,312	**1,000**	0,000	0,312	0,476	0,535
ResNet50	**0,848**	0,750	0,892	**0,759**	**0,754**	**0,886**
DenseNet121	0,812	0,682	0,871	0,706	0,694	0,884
EfficientNet-B7	0,670	0,034	**0,959**	0,273	0,061	0,552
Modelos clássicos (features ResNet50)						
SVM-RBF	0,727	0,318	0,912	0,622	0,421	0,799
LightGBM	**0,784**	**0,466**	0,928	**0,745**	**0,573**	**0,823**
Random Forest	0,716	0,148	**0,974**	0,722	0,245	0,789

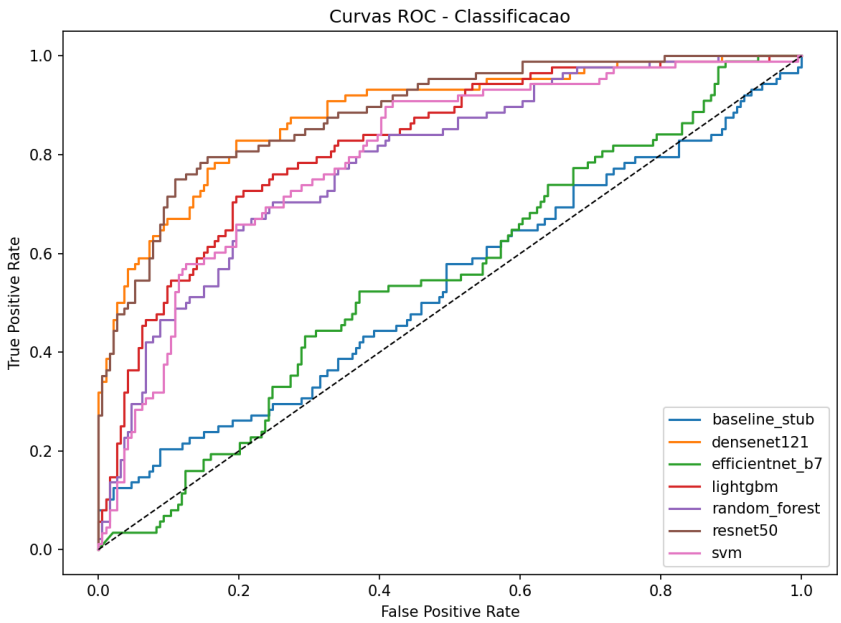
Fonte: Elaborado pelo autor.

Tabela 3 - Comparação de métricas dos modelos de segmentação no teste.

Modelo	Dice	IoU	Sensib. por pixel	Especif. por pixel
Attention U-Net	**0,914**	**0,841**	**0,900**	**0,993**
U-Net	0,909	0,833	0,893	0,992

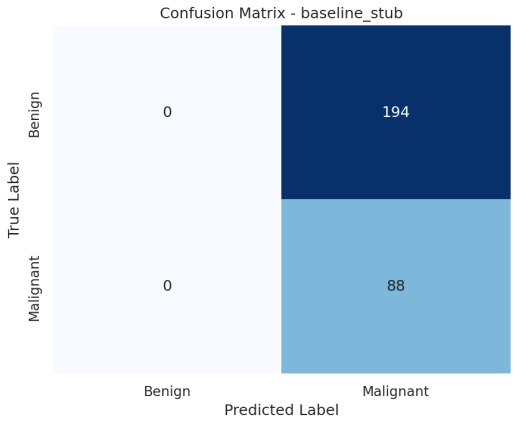
Fonte: Elaborado pelo autor.

Figura 3 - Curvas ROC dos modelos de classificação

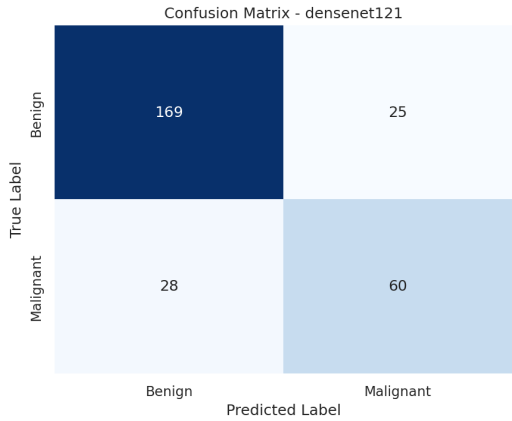


Fonte: Elaborado pelo autor.

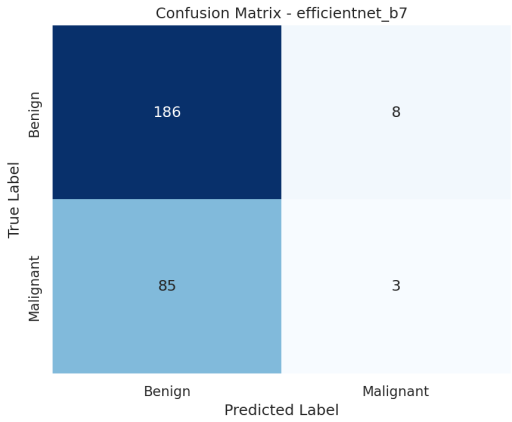
Figura 4 - Matrizes de confusão dos modelos de classificação



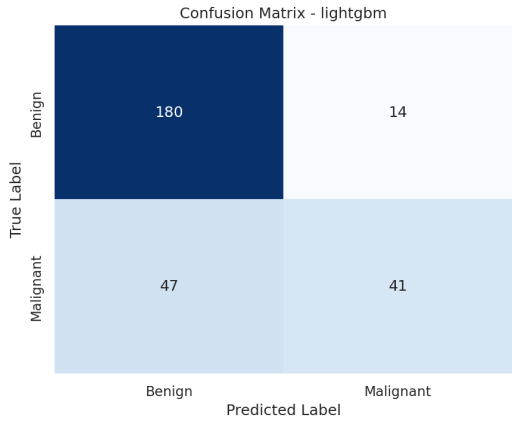
(a) Baseline



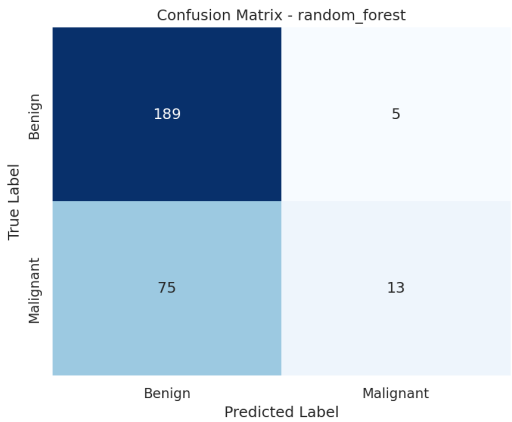
(b) DenseNet121



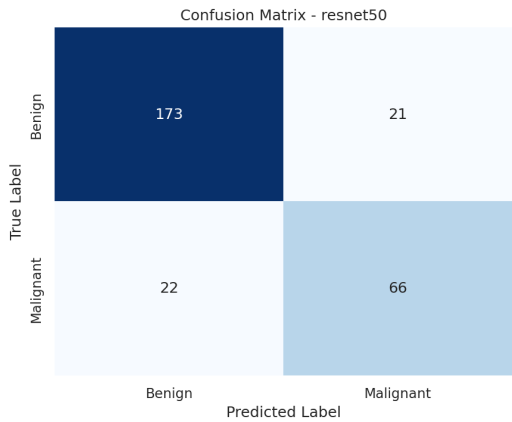
(c) EfficientNet-B7



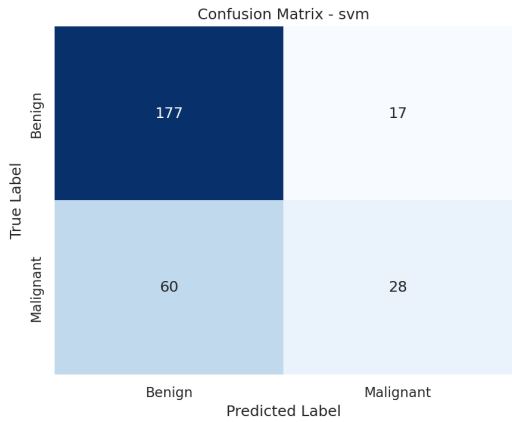
(d) LightGBM



(e) Random Forest



(f) ResNet50



(g) SVM-RBF

Fonte: Elaborado pelo autor.

Foram executados 1.000 reamostragens bootstrap ($\alpha = 0{,}05$), conforme Tabela 4. Em termos de acurácia, os intervalos de confiança de ResNet50 e DenseNet121 apresentaram ampla sobreposição. Os testes pareados (McNemar e DeLong) estão resumidos na Tabela 5; observou-se diferença significativa entre ResNet50 e LightGBM ($p < 0{,}05$), porém sem diferença significativa entre ResNet50 e DenseNet121

Tabela 4 - Intervalos de confiança (95%) por bootstrap para classificação.

Modelo	**IC Acurácia**	**IC F1-score**	**IC AUC-ROC**
baseline_stub	[0,259; 0,369]	[0,411; 0,539]	[0,459; 0,609]
ResNet50	**[0,805; 0,887]**	**[0,679; 0,822]**	**[0,842; 0,923]**
DenseNet121	[0,762; 0,855]	[0,608; 0,767]	[0,834; 0,923]
EfficientNet-B7	[0,617; 0,723]	[0,000; 0,130]	[0,481; 0,627]
SVM-RBF	[0,674; 0,780]	[0,312; 0,525]	[0,742; 0,848]
LightGBM	[0,734; 0,830]	[0,465; 0,667]	[0,770; 0,872]
Random Forest	[0,663; 0,766]	[0,133; 0,354]	[0,732; 0,841]

Fonte: Elaborado pelo autor.

Tabela 5 - Resumo de testes pareados entre modelos.

Comparação	**McNemar**	**DeLong AUC**
ResNet50 vs DenseNet121	0,1939	0,8869
ResNet50 vs LightGBM	0,0336	0,0129
ResNet50 vs SVM-RBF	$3{,}71 \times 10^{-5}$	0,0019
ResNet50 vs Random Forest	$5{,}11 \times 10^{-5}$	0,0005
ResNet50 vs EfficientNet-B7	$\mathbf{4{,}33 \times 10^{-7}}$	$< \mathbf{10^{-15}}$
DenseNet121 vs LightGBM	0,4158	0,0457

Fonte: Elaborado pelo autor.

Em relação à literatura, observou-se convergência parcial e divergências esperadas. Os resultados obtidos neste estudo não alcança-

ram os níveis reportados por Latha et al. (2024) (acurácia de 99,14% com EfficientNet-B7 em BUSI) e por Nastase et al. (2024) (AUC até 1,00), o que contradiz a expectativa de desempenho absoluto semelhante quando se considera apenas a arquitetura. No presente trabalho, o EfficientNet-B7 apresentou desempenho inferior (AUC de 0,552), enquanto ResNet50 e DenseNet121 foram superiores no cenário BUS-BRA com divisão estratificada por paciente.

Comparando com Foleis et al. (2025) e Diniz et al. (2024), os resultados confirmam a relevância de estratégias robustas de modelagem e avaliação, mas indicam valores absolutos mais modestos no protocolo adotado. Em segmentação, os resultados de Dice (0,914 para Attention U-Net) se aproximam dos achados de Sulaiman et al. (2024) (Dice de 0,92), mas permanecem abaixo dos valores extremos relatados por Varshney, Verma e Kaur (2025). Dessa forma, os achados expandem a discussão ao evidenciar desempenho competitivo em BUS-BRA com protocolo orientado por paciente, cenário metodologicamente mais restritivo que divisões por imagem.

Também se destaca que comparações diretas de métricas entre estudos devem ser interpretadas com cautela, pois diferenças de base (BUS-BRA versus BUSI), critérios de particionamento, pré-processamento e balanceamento alteram substancialmente a dificuldade experimental. Nesse sentido, os resultados são coerentes com o contexto metodológico deste trabalho e com o escopo de generalização interna do experimento.

Do ponto de vista clínico, a combinação de alta especificidade (por exemplo, 0,892 no ResNet50 e 0,928 no LightGBM) com sensibilidade moderada sugere utilidade potencial como ferramenta de apoio, mas ainda com risco de perda de casos malignos quando se privilegia excessivamente a redução de falso-positivo. Em termos computacionais, os resultados reforçam que arquiteturas profundas bem ajustadas tendem a capturar melhor os padrões discriminativos do ultrassom mamário do que pipelines clássicos baseados em atributos extraídos.

O desbalanceamento de classes (1.268 casos benignos versus 607 malignos) impactou diretamente as métricas, com tendência de alguns modelos a favorecer a classe majoritária. Esse comportamento foi evidente em modelos com alta especificidade e baixa sensibilidade (por exemplo, Random Forest), indicando que métricas agregadas como acurácia devem ser analisadas juntamente com recall, F1 e AUC-ROC.

Os resultados indicam que o ResNet50 apresentou possívelmente

o melhor desempenho geral em classificação (maior AUC-ROC e maior acurácia), com DenseNet121 estatisticamente equivalente nos testes ResNet50 vs DenseNet121 (McNemar e DeLong sem significância, $p > 0{,}05$). Entre os modelos clássicos, o LightGBM foi o mais consistente, porém inferior aos melhores modelos profundos em AUC e F1. Em segmentação, o Attention U-Net foi superior ao U-Net clássico em Dice, IoU e métricas por pixel.

As principais limitações do estudo incluem: uso de um único dataset (BUS-BRA), limitação no número de épocas de treinamento, ausência de validação externa em dados de outras instituições brasileiras e possibilidade de viés associado à variação de equipamentos/protocolos de ultrassom.

4 CONCLUSÃO

Este trabalho desenvolveu e avaliou comparativamente modelos de inteligência artificial aplicados à detecção de lesões potencialmente cancerígenas em imagens de ultrassonografia mamária, integrando abordagens de classificação profunda, métodos clássicos de aprendizado de máquina e arquiteturas de segmentação semântica.

Os objetivos propostos foram integralmente alcançados. Os experimentos foram conduzidos sobre o conjunto público BUS-BRA, com 1.875 imagens de 1.064 pacientes e divisão estratificada por paciente na proporção 70/15/15 (treino/validação/teste) para mitigar o risco de vazamento de dados. Foram avaliados quatro modelos de classificação profunda, ResNet50, DenseNet121, EfficientNet-B7 e *baseline_stub*, três classificadores clássicos sobre características extraídas por *encoder* pré-treinado, SVM-RBF, LightGBM e Random Forest, e dois modelos de segmentação, U-Net e Attention U-Net.

O desempenho foi medido por acurácia, sensibilidade, especificidade, F1-score e AUC-ROC na classificação, e por coeficiente Dice e IoU na segmentação, com comparações inferenciais por *bootstrap* não paramétrico, McNemar e DeLong. O ResNet50 obteve o melhor desempenho em classificação (AUC-ROC de 0,886; acurácia de 0,848), sendo estatisticamente equivalente ao DenseNet121 ($p > 0{,}05$) e superior a todos os demais modelos ($p < 0{,}05$). A Attention U-Net superou o U-Net clássico em segmentação (Dice de 0,914; IoU de 0,841), confirmando o ganho dos mecanismos de atenção na delimitação das regiões de interesse.

As principais contribuições situam-se nos planos científico e prático. Cientificamente, o estudo conduz uma fundamentada entre abordagens profundas, clássicas e de segmentação sobre o BUS-BRA com protocolo de

divisão por paciente. O achado de que o EfficientNet-B7, frequentemente citado como referência, apresentou desempenho inferior ao ResNet50 e ao DenseNet121 neste cenário evidencia que métricas da literatura devem ser interpretadas em função do protocolo experimental e do conjunto de dados adotados, oferecendo subsídios concretos para escolhas metodológicas em pesquisas futuras. Praticamente, os modelos desenvolvidos demonstram potencial como ferramentas de apoio ao diagnóstico assistido por computador, especialmente em regiões com escassez de especialistas em radiologia mamária, e o protocolo reprodutível disponibilizado constitui base metodológica reaproveitável por investigações futuras na área de computação aplicada à saúde oncológica.

Como encaminhamentos para trabalhos futuros, destacam-se a validação externa em conjuntos de dados de outras instituições brasileiras, a incorporação de técnicas de interpretabilidade como *Grad-CAM* e *SHAP* para aproximar as decisões computacionais do raciocínio radiológico, a exploração de estratégias de *ensemble* entre os modelos mais promissores e a integração de dados multimodais que combinem imagens de ultrassonografia com informações clínicas estruturadas, como categoria BI-RADS e dados demográficos, visando aproximar as ferramentas computacionais das condições reais da prática clínica.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Os autores declaram que ferramentas de Inteligência Artificial generativa foram utilizadas exclusivamente como apoio auxiliar à elaboração deste trabalho, incluindo assistência na revisão linguística, estruturação textual e organização de ideias. Nenhuma ferramenta de IA foi utilizada para substituição da autoria intelectual, análise científica, interpretação dos resultados ou tomada de decisões acadêmicas. A responsabilidade integral pelo conteúdo, originalidade, veracidade das informações e conformidade com os princípios de integridade científica permanece integralmente atribuída aos autores.

REFERÊNCIAS

DALLAGASSA, M. R.; OLIVEIRA, A. J. C. de. **Uso de machine learning no diagnóstico de câncer de mama através de ultrassonografia**. Journal of Health Informatics, v. 16. 2024, Disponível em: <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1289>. Acesso em: 25 ago. 2025.

DINIZ, J. O. B. et al. **EfficientEnsemble: Diagnóstico De Câncer De Mama Em Imagens De Ultrassom Utilizando Processamento De Imagens E Ensemble De EfficientNets**. 2024.

FEBRASGO. **Diagnóstico precoce pode garantir taxa de cura de até 90% no câncer de mama**. 2025. Disponível em: <https://www.febrasgo.org.br/noticia/diagnostico-precoce-pode-garantir-taxa-de-cura-de-ate-90-no-cancer-de-mama/>. Acesso em: 28 ago. 2025.

FILHO, J. O. F. M. et al. **Detecção de câncer de mama por imagem com classificador híbrido**. Journal of Health Informatics, Sociedade Brasileira de Informática em Saúde, v. 16. 2024.

FOLEIS, V. K. et al. **Transfer learning and handcrafted features ensembles for ultrasound breast cancer image classification**. Revista de Informatica Teorica e Aplicada, Federal University of Rio Grande do Sul, Institute of Informatics, v. 32, p. 11–17. ISSN 21752745. 2025.

GóMEZ-FLORES, W.; GREGORIO-CALAS, M. J.; PEREIRA, W. Coelho de A. **Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems**. Medical Physics, v. 51, n. 4, p. 3110–3123. 2024, Disponível em: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.16812>.

INCA. **Estimativa 2023–2025: 73 610 novos casos de câncer de mama no Brasil, com risco estimado de 66,54 casos por 100 000 mulheres**. 2024. Disponível em: <https://sogimig.org.br/cancer-de-mama-incidencia-fatores-de-risco-e-formas-de-diagnostico/>. Acesso em: 30 ago. 2025.

LATHA, M. et al. **Revolutionizing breast ultrasound diagnostics with efficientnet-b7 and explainable ai**. BMC Medical Imaging, BioMed Central Ltd, v. 24. ISSN 14712342. 2024.

NASTASE, I. N. A. et al. **Role of inter- and extra-lesion tissue, transfer learning, and fine-tuning in the robust classification of breast lesions**. Scientific Reports, Nature Research, v. 14. ISSN 20452322. 2024.

PARISE, V. C. **Sistema De Auxílio À Identificação E Classificação De Microcalcificações Em Mamogramas Digitais Via Processamento Digital De Imagens E Redes**. 2003.

SANTOS, A. M. R. dos.; DIAS, M. B. K. **Diretrizes para a deteccao precoce do cancer de mama no Brasil**. [S.l.]: INCA. 166 p. ISBN 9788573182736. 2015.

SERRANO, A. R. B. et al. **Aplicações de algoritmos de inteligência artificial na detecção precoce do câncer de mama: Uma revisão integrativa**. Revista Educação em Saúde, v. 13, p. 271–281. 2025.

SULAIMAN, A. et al. **Attention based unet model for breast cancer segmentation using busi dataset**. Scientific Reports, Nature Research, v. 14. ISSN 20452322. 2024.

VARSHNEY, T.; VERMA, K.; KAUR, A. **A modified multiresunet model with attention focus for breast cancer detection**. Computers and Electrical Engineering, Elsevier Ltd, v. 124. ISSN 00457906. 2025.

VEDOVA, F. B. D.; RUARO, A. F. **Teste De Modelos De Inteligência Artificial Para Análise De Exames De Punção Aspirativa Com Agulha Fina**. 2024. 15 p.