Universidade de Brasília – UnB
Faculdade de Ciências e Tecnologias em Engenharia – FCTE
Programa de Pós-Graduação em Engenharia Biomédica

Estudo comparativo entre modelos de redes neurais para classificação de tumores em ultrassonografia mamária

Tiago Rodrigues Pereira

Orientador: Dr. Cristiano Jacques Miosso

UnB – Universidade de Brasília

FCTE – Faculdade de Ciências e Tecnologias em Engenharia

Estudo comparativo entre modelos de redes neurais para classificação de tumores em ultrassonografia mamária

Tiago Rodrigues Pereira

Orientador: Dr. Cristiano Jacques Miosso

Dissertação de Mestrado em
Engenharia Biomédica

Publicação: 202A/2025

Brasília/DF, maio de 2025

UnB – Universidade de Brasília

FCTE – Faculdade de Ciências e Tecnologias em Engenharia

Programa de Pós-Graduação

Estudo comparativo entre modelos de redes neurais para classificação de tumores em ultrassonografia mamária

Tiago Rodrigues Pereira

Dissertação de mestrado submetida ao programa de pós-graduação em Engenharia Biomédica da Universidade de Brasília, como parte dos requisitos necessários para a obtenção do grau de mestre em Engenharia Biomédica

Aprovada por:

Dr. Cristiano Jacques Miosso

(Orientador)

Dr. Fabricio Ataides Braz

(Examinador interno)

Dr. Wagner Coelho de Albuquerque Pereira

(Examinador externo)

Brasília/DF, maio de 2025

Ficha Catalográfica

R. Pereira, Tiago
Estudo comparativo entre modelos de redes neurais para classificação de tumores em ultrassonografia mamária
[Distrito Federal], 2025.
91p., 210 × 297 mm (FCTE/UnB Gama, Mestrado em Engenharia Biomédica, 2025).
Dissertação de Mestrado em Engenharia Biomédica, Faculdade UnB Gama, Programa de Pós-Graduação em Engenharia Biomédica.

1. Deep learning
2. Visão computacional
3. Câncer de mama
4. Ultrassonografia

I. FCTE UnB/UnB.
II. Título (série)

Referência

R. Pereira, Tiago (2025). Estudo comparativo entre modelos de redes neurais para classificação de tumores em ultrassonografia mamária. Dissertação de mestrado em engenharia biomédica, Publicação 202A/2025, Programa de Pós-Graduação, Faculdade UnB Gama, Universidade de Brasília, Brasília, DF, 91p.

Cessão de Direitos

---

pereira.tiagorodri@gmail.com

Brasília, DF – Brasil

Resumo

O câncer de mama é o tipo que causa mais óbitos entre as mulheres no Brasil e o diagnóstico precoce eleva o prognóstico de pacientes com essa enfermidade. Dentre as principais técnicas não invasivas de detecção do câncer de mama, a ultrassonografia é frequentemente utilizada em mulheres jovens. No entanto, sua eficácia está amplamente vinculada à experiência do profissional de saúde responsável pela sua execução.

Assim sendo, a proposta desse estudo é a avaliação comparativa de múltiplas arquiteturas de redes neurais profundas para classificação de imagens de ultrassom da mama com lesões presentes. Nesta avaliação, são feitas seis classificações diferentes: uma quanto à malignidade e cinco variações das categorias BI-RADS. Dentre as arquiteturas implementadas, cinco são variantes de redes neurais convolucionais (CNNs) e uma se baseia no *vision transformer* (ViT).

Para treinamento, validação e teste, são utilizados bancos de dados públicos que contêm tanto rótulos quanto a malignidade como as categorias BI-RADS. Devido à baixa quantidade de dados disponíveis, foi necessário aplicar um aumento artificial por transformações aleatórias. Para aprimoramento dos resultados, são adotados modelos pré-treinados combinados com a técnica de *transfer learning* (TL). Adicionalmente, são avaliadas quatro diferentes estratégias de pré-processamentos: três baseadas na segmentação por região de interesse (ROI) e uma utilizando a imagem original.

A melhor arquitetura para classificação quanto à malignidade, MixNet-XL, alcançou acurácia de $88,3\,\%$, F1-*Score* de $82,0\,\%$, e área sob a curva ROC (AUC) de $91,5\,\%$. Enquanto o ViT, melhor para classificação BI-RADS 2 a 5 (multiclasse), apresentou acurácia balanceada, F1-*Score* médio por classe, e AUC de $87,4\,\%$, $71,0\,\%$, e $88,1\,\%$, respectivamente.

Os modelos implementados evidenciaram desempenho próximo ou superior ao estado da arte quando utilizados pré-processamentos com segmentação. O uso da imagem original como entrada apresentou resultados promissores nas arquiteturas CNNs propostas.

Devido às limitações de *hardware* para o treinamento dos modelos, as arquiteturas de alta complexidade não foram testadas, as quais apresentam desempenho de estado da arte em *benchmarks* gerais, como ImageNet1K. Para aplicação clínica, como na triagem de pacientes, os modelos com maior desempenho são prioritários em comparação àqueles cujo desenvolvimento considerou a eficiência computacional, como os avaliados neste estudo. Dessa forma, o uso de arquiteturas mais complexas, em conjunto com técnicas de IA explicável (xAI) para os rótulos definidos, configura-se como uma tendência atual para aplicação de redes neurais em sistemas de auxílio a radiologistas no diagnóstico de imagens de ultrassom mamário.

Palavras-chave: Câncer de mama, ultrassonografia, classificação de imagens mamárias, *deep learning*.

Abstract

Breast cancer is the type that causes the most deaths among women in Brazil and early diagnosis improves the prognosis for patients with this disease. Of the main non-invasive breast detection techniques, ultrasonography is frequently used in young women. However, its effectiveness is largely linked to the experience of the healthcare professional responsible for its execution.

In this study, I aim to perform a comparative evaluation of multiple deep neural network architectures for the classification of breast ultrasound images with lesions. In this assessment, six different classifications are performed: one regarding malignancy (benign or malignant) and five variations of the BI-RADS categories. Among the implemented architectures, five are variants of convolutional neural networks (CNNs) and one is based on the vision transformer (ViT).

For training, validation, and testing, public databases containing labels for both malignancy and the BI-RADS categories are used. Due to the limited amount of available data, we applied data augmentation using random transformations. To improve the results, pre-trained models combined with transfer learning (TL) are adopted. Additionally, four different preprocessing strategies are evaluated: three based on region of interest (ROI) segmentation and one using the original image.

The best architecture for malignancy classification, MixNet-XL, achieved an accuracy of 88.3 %, F1-Score of 82.0 %, and area under the ROC curve (AUC) of 91.5 %. Meanwhile, ViT, which performed best for BI-RADS 2 to 5 (multiclass) classification, reached balanced accuracy, average F1-Score per class, and AUC of 87.4 %, 71.0 %, and 88.1 %, respectively.

The implemented models performed close to or superior to the state-of-the-art when using preprocessing with segmentation. The use of the original image as input showed promising results in the proposed CNN architectures.

Hardware limitations for model training prevented testing of high-complexity architectures, which have demonstrated state-of-the-art performance on image classification benchmarks, such as ImageNet1K. For clinical application, such as patient screening, the highest-performing models are prioritized over those developed with computational efficiency in mind, like the ones evaluated in this study. Therefore, using more complex architectures, in conjunction with Explainable AI (xAI) techniques for the given labels, is a current trend in the application of neural networks to systems that aid radiologists in diagnosing breast ultrasound images.

Keywords: Breast cancer, ultrasound, BUS classification, Deep Learning.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Nomenclaturas e Abreviações

ACR Colégio Americano de Radiologia, do inglês *American College of Radiology*

Acurác. Acurácia

ADAM *Adaptive Moment Estimation*

Amost. Amostrador

ANS Agência Nacional de Saúde Complementar

AUC Área sob a curva ROC, do inglês *Area Under the Curve*

bin. binário

BI-RADS Sistema de Laudos e Registro de Dados de Imagem da Mama, do inglês *Breast Imaging Reporting and Data System*

CADx Sistemas de Auxílio ao Diagnóstico por Computador, do inglês *Computer-Aided Diagnosis*

CDIS Carcinoma Ductal *in situ*

CLI Carcinoma Lobular Invasivo

CNN Rede Neural Convolucional, do inglês *Convolutional Neural Network*

CPNM Câncer de Pele Não Melanoma

DNN Rede Neural Profunda, do inglês *Deep Neural Network*

Especif. Especificidade

FC *Fully Connected layer*

FLOPS *Floating-Points Operations Per Second*

FN Falso Negativo

FP Falso Positivo

GELU *Gaussian-Error Linear Unit*

GF gigaflop (GFLOP)

IA Inteligência Artificial

IDH Índice de Desenvolvimento Humano

INCA Instituto Nacional de Câncer

ML Aprendizado de Máquina, do inglês *Machine Learning*

MLP	Perceptron Multicamadas, do inglês *Multilayer Perceptron*
multicl.	multiclasse
NAS	Busca Automática de Redes Neurais Artificiais, do inglês *Neural Architecture Search*
Prec.	Precisão
Pré-proc.	Pré-processamento
ReLU	*Rectified Linear Unit*
RMSProp	*Root Mean Square Propagation*
ROC	Característica de Operação do Receptor, do inglês *Receiver Operating Characteristic*
ROI	Região de Interesse, do inglês *Region Of Interest*
Sensib.	Sensibilidade
SGD	Gradiente descendente estocástico, do inglês *Square Gradient Descent*
SUS	Sistema Único de Saúde
SVM	Máquinas de Vetores de Suporte, do inglês *Suport Vector Machine*
TL	*Transfer Learning*
VGG	*Visual Geometry Group*
ViT	*Vision Transformer*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

1 Introdução

O câncer é uma das principais causas de morte no mundo, sendo a causa de aproximadamente 10,1 milhões de mortes em 2019 (1). Segundo o Instituto Nacional de Câncer (INCA), a incidência estimada para o Brasil é de 704 mil novos casos para cada ano do triênio 2023 a 2025 (2).

Dentre as mulheres no Brasil, a taxa de incidência do câncer de mama para o triênio (2023 a 2025) será de $30,1\,\%$ em relação a todos os casos, o maior com exclusão dos casos de câncer de pele não melanoma (3). Além disso, o câncer de mama é o tipo com maior taxa de mortalidade nas mulheres, sendo responsável por 20.165 óbitos de mulheres no Brasil em 2023 ($16,55\,\%$ do total de óbitos causado por câncer e $3,05\,\%$ do total de óbitos nessa população) (4, 3).

O câncer de mama tem melhor prognóstico caso detectado precocemente e tratado de forma oportuna (5, 6). A existência de programas de saúde pública para exames periódicos de grupos de risco é essencial para diminuir a mortalidade e aumentar a capacidade de sobrevida devido detecção e diagnóstico precoce (6, 7).

Dentre diversas técnicas de detecção precoce, as técnicas baseadas em imagem são comumente usadas para identificar o câncer de mama devido à efetividade comprovada e recomendada por órgãos reguladores (5, 8). Nessa modalidade, a mamografia é o padrão internacional para rastreamento e detecção (9). Contudo, ela não atende a todos os possíveis casos devido a limitações de sensibilidade, como nos casos de mamas densas, e a preocupação com níveis de ionização (5, 10). Outra técnica usada de forma complementar ou exclusiva, como na detecção precoce em mulheres abaixo dos 30 anos, é a ultrassonografia mamária (5, 8).

A ultrassonografia mamária bilateral é amplamente usada para detecção de câncer de mama. Com base em parâmetros técnicos calculados por (8), estimou-se mais de mais de 22 mil procedimentos anuais para detecção precoce no Sistema Único de Saúde (SUS) em uma população de 1.118.418 mulheres em uma determinada localidade do Brasil. Esse total é $39\,\%$ superior ao número de mamografias estimado para a mesma população.

Normalmente, as imagens de exames de detecção no câncer de mama, como a ultrassonografia mamária bilateral, são analisadas visualmente por um radiologista ou profissional

qualificado. Isso faz com que a eficácia da avaliação fique sujeita à experiência, subjetividade e tempo disponível desse profissional (10). Em vista disso, o desenvolvimento de modelos para análise e classificação de imagens de ultrassom é essencial para melhorar a eficiência e confiabilidade dos exames de detecção precoce do câncer de mama (11).

Sistemas de Auxílio ao Diagnóstico por Computador, do inglês *Computer-Aided Diagnosis* (CADx), são estudados há décadas na literatura para aplicação no câncer de mama. Muitos desses sistemas realizam um processo de três etapas: a detecção de lesões, seguida pela extração convencional de características e pela classificação (10).

O uso da aprendizagem de máquina para desenvolvimento de melhores sistemas de detecção precoce cresceu nos últimos anos. Sociedades representantes de radiologistas, como a do Reino Unido, implementaram recentemente diretrizes para uso dessa tecnologia clinicamente. Isso ilustra a rápida adoção dessa técnica em ambientes clínicos (12, 13).

Entre os diferentes subsistemas que compõe um CADx, a classificação de lesões é um problema que modelos que utilizam redes neurais profundas, do inglês *Deep Neural Networks* (DNNs), buscam solucionar (14). Principalmente quando são consideradas diferentes categorias definidas pelo Sistema de Laudos e Registro de Dados de Imagem da Mama (BI-RADS). Este sistema padroniza a interpretação e categorização de exames por imagem da mama (15). Um dos principais pontos que o BI-RADS busca solucionar é a subjetividade da interpretação das imagens de exames da mama, pois os radiologistas, geralmente, realizam uma avaliação qualitativa dessas imagens.

Contudo, o desenvolvimento de DNNs para classificação de imagens implica desafios para essa pesquisa. A principal é a necessidade de uma grande quantidade de amostras para treinamento de modelos Rede Neural Profunda, do inglês *Deep Neural Network* (DNN) que se aproximem ao desempenho de estado da arte (16). Uma das formas para alcançar melhor desempenho é utilizar a transferência de conhecimento, do inglês *transfer learning* (TL), por meio de modelos treinados para outra aplicação (17, 18).

Nesse sentido, essa dissertação visa a análise comparativa de múltiplas arquiteturas de redes neurais focada na classificação de lesões mamárias baseada tanto nas categorias BI-RADS como quanto à malignidade em imagens de ultrassonografia bilateral. Na abordagem proposta, os bancos de imagens e os modelos propostos utilizando TL estão disponíveis publicamente na literatura para reprodutibilidade dos resultados.

As limitações desta dissertação são a baixa quantidade de amostras nas bases públicas existentes e a capacidade computacional disponível para treinamento das DNNs. Esta capacidade computacional limitada restringe as arquiteturas possíveis para ajuste fino por TL, devido às altas demandas computacionais de modelos de estado da arte para classificação de imagens gerais (19).

1.1 Objetivos

1.1.1 Objetivo Geral

A análise de múltiplos modelos de redes neurais profundas para classificação de lesões em imagens de ultrassom da mama em classes definidas tanto pelo BI-RADS quanto à malignidade ou não da lesão.

1.1.2 Objetivos Específicos

Os objetivos específicos são:

- Avaliação do desempenho de diferentes pré-processamentos aplicados a imagens de ultrassom da mama
- Comparação de diferentes arquiteturas de redes neurais profundas (DNNs) para mesma classificação (CNNs e *Transformer*)
- Comparação de modelos para diferentes grupos de classificações derivadas das categorias no BI-RADS.
- Comparação de diferentes tipos de *transfer learning* na arquitetura ViT.

Essa dissertação foi estruturada em sete capítulos: o capítulo dois descreve a fundamentação teórica sobre o câncer de mama, descrevendo a anatomia patológica e classificação das lesões mamárias, a detecção precoce por imagem, e o estado da arte para classificação de lesões mamárias em imagens de ultrassom utilizando redes neurais profundas. O capítulo três apresenta os materiais e métodos utilizados, com a descrição dos bancos de dados utilizados e seu o pré-processamento, das arquiteturas dos modelos propostos, treinamento, e métricas utilizadas para avaliação. O capítulo quatro descreve os modelos proposto, apresentando os diferentes experimentos feitos e respectiva implementação. No capítulo cinco são apresentados e discutidos os resultados para os modelos de arquitetura CNN para cada classificação, enquanto o capítulo seis aborda a arquitetura do *vision transformer* (ViT). No capítulo sete, a conclusão é apresentada, seguida dos trabalhos futuros.

2 Fundamentação Teórica sobre câncer de mama e estado da arte de modelos de redes neurais para classificação de lesões

Esse capítulo apresenta informações relacionadas ao câncer de mama, descrevendo estatísticas de incidência e mortalidade, anatomia patológica e detecção precoce por imagem. Também são apresentados a classificação BI-RADS e o estado da arte de redes neurais profundas para imagens de ultrassom.

2.1 Câncer de mama

O termo câncer de mama é uma denominação médica referente a uma lesão ou neoplasma de tipo maligno em tecido mamário (20). Uma lesão mamária também pode ser classificada como benigno, dependendo das características histológicas e anatômicas (20). Ambas as classificações têm em comum o crescimento recente e não ordenado de tecido, que pode ou não formar uma massa física (tumor) (20).

Nos casos de lesão maligna, a detecção precoce é essencial para bom prognóstico do indivíduo, este que em grande parte se refere a população feminina (6). Nesse sentido, duas estratégias podem ser aplicadas: o diagnóstico precoce, destinado a mulheres que apresentam sintomas de suspeita de câncer, ou rastreamento, destinado ao grupo de mulheres assintomáticas com maior risco (8).

2.1.1 Incidência e mortalidade no Brasil e no mundo

Dentre diversos tipos de neoplasias que ocorrem no corpo humano, as que ocorrem na mama recebem atenção especial por ser o tipo que causa mais óbitos nas mulheres (2). No Brasil a incidência do câncer de mama se mostrou maior em regiões com maior Índice de Desenvolvimento Humano (IDH) (2). Esse fato se espelha ao que é observado em países com maior IDH e se relaciona com envelhecimento populacional, mudanças de hábitos e comportamentos, e aumento dos diagnósticos devido melhor difusão do rastreamento mamográfico (2).

Em números, o câncer de mama representa 11,5 % do total de casos previstos, com incidência em 2022 de 23,8 % para sexo feminino conforme observado na Figura 2.1. No Brasil, a estimativa de incidência percentual entre neoplasias em 2023 do câncer de mama feminina em cada região, do maior para o menor com exclusão de pele não melanoma, é (2): de 32,9 % na Região Sudeste; 28,1 % na região Centro-Oeste; 28,1 % no Nordeste; e de 22,4 % na Região Norte.

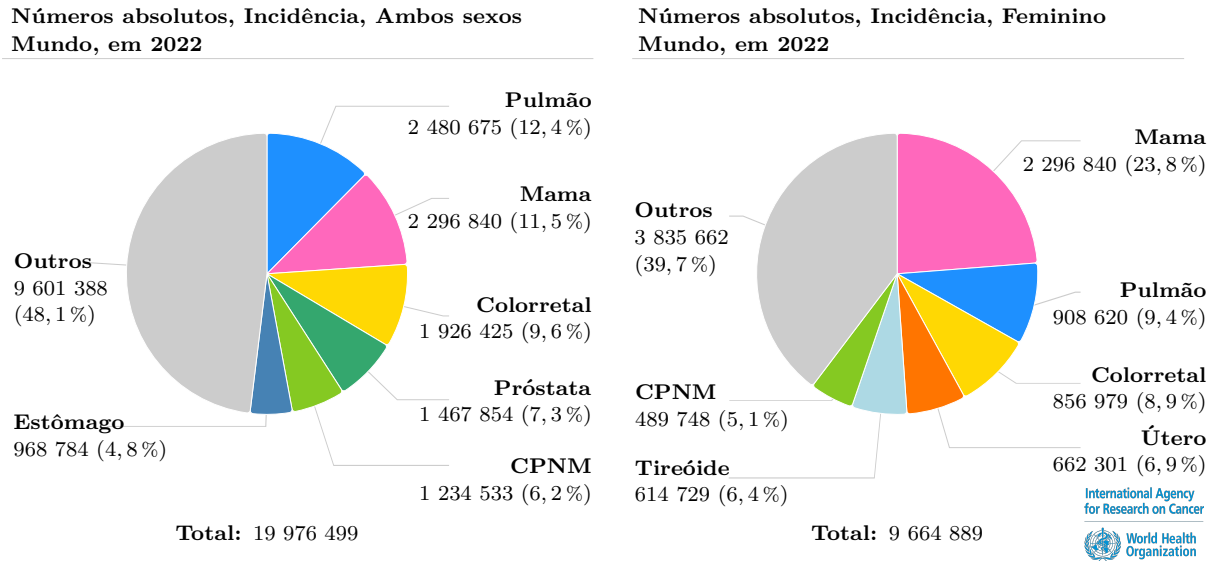


Figura 2.1. Incidência de câncer no mundo para ambos os sexos e incidência de câncer para sexo feminino em números absolutos e porcentagem em 2022. Adaptado de: (21).

Dentre diferentes aplicações, a incidência pode ser utilizada para estabelecer a população sintomática e assintomática, e assim estimar a quantidade de procedimentos necessários para rastreamento e diagnóstico precoce (8).

Conforme estudo do INCA (8), foram definidos parâmetros técnicos para estimar a quantidade de procedimentos necessários no SUS com base na população sintomática e assintomática de mulheres no Brasil. Para ultrassonografia mamária bilateral, o parâmetro técnico, definido pelo INCA no triênio 2023–2025, para detecção precoce em mulheres sintomáticas abaixo de 30 anos é de 278,35 % e, para mulheres acima de 30 anos, de 207,03 % da população estimada. Para mulheres assintomáticas de 50 a 69 anos, faixa recomendada para rastreamento, foi calculado um parâmetro de 3,50 % dessa população a ser rastreada no SUS (8). A memória do cálculo dos parâmetros técnicos é detalhada no Anexo A.

Aplicando a metodologia definida em (8), em conjunto com a população feminina total do Brasil do Censo 2022 e a população feminina de 50 a 59 a ser rastreada no SUS, dados de cobertura da Agência Nacional de Saúde Complementar (ANS) em dez/2024, foram estimados os procedimentos da Tabela 2.1.

Tabela 2.1. Total de procedimentos anuais estimados para detecção precoce do câncer de mama, em 2025, no Brasil, conforme metodologia de (8). População do sexo feminino no Brasil em 2022: 104.548.325; População do sexo feminino de 50 a 69 anos em 2022: 22.360.526; Cobertura pela ANS em Dez/2024 para população alvo entre 50 a 69 anos: 23,9 % (22, 23).

Procedimentos	População feminina						Total de procedimentos
	Sintomática (< 30 anos)		Sintomática (> 30 anos ou mais)		Rastreamento (50 a 69 anos)		
	N (valor de A)=>	53.320	N (valor de B)=>	588.607	N (valor de C)=>	17.016.360	
	Parâmetros	Número de procedimentos	Parâmetros	Número de procedimentos	Parâmetros	Número de procedimentos	
Mamografia bilateral para rastreamento	-	-	-	-	50,00 %	8.508.180	8.508.180
Mamografia	28,15 %	15.009	147,10 %	865.841	2,90 %	493.474	1.374.325
Ultrassonografia mamária bilateral	278,37 %	148.426	207,03 %	1.218.593	3,50 %	595.573	1.962.592
Biópsia por agulha grossa	18,07 %	9.635	17,74 %	104.419	0,73 %	124.219	238.273
Exérese de nódulo de mama (nodulectomia)	13,03 %	6.948	6,16 %	36.258	0,11 %	18.718	61.924
Exame anatomopatológico de mama – Biópsia	31,10 %	16.582	23,90 %	140.677	0,84 %	142.937	300.197
Exame Citopatológico de Mama	3,44 %	1.834	2,50 %	14.715	-	-	16.549

2.1.2 Anatomia patológica da mama

As mamas são glândulas de tamanho variável localizadas na parede torácica anterior. Sua variabilidade de volume está relacionada a gordura subcutânea, variando com a gordura corporal do indivíduo. Com função de sustentação, diversas fibras de tecido conjuntivo permeiam entre os lóbulos de gordura e os glândula mamaria (repouso e lactante) (20).

Segundo (20), lesões mamárias podem se manifestar de duas maneiras: assintomáticas ou apresentando sintomas clínicos, como derrame papilar, lesões sólidas bem e mal delimitadas e microcalcificações.

2.1.2.1 Lesões neoplásicas papilares

As lesões papilares estão associadas ao sintoma de derrame papilar, este que se refere a saída de secreção do mamilo fora de períodos normais. As lesões relacionadas a esse tipo de sintoma podem ser benignos, como papiloma intraductal, ou maligno, como o carcinoma ductal *in situ* (CDIS) papilar, carcinoma papilar encapsulado, e o carcinoma papilar sólido (20).

O papiloma intraductal é uma lesão benigna situado nas paredes dos ductos lactíferos e pode causar obstrução localizada (20). Ele se caracteriza como uma projeção papilar revestida de células e enraizada de um ducto, conforme exemplo de ultrassom na Figura 2.2a.

No caso das neoplasias papilares malignas citadas, o CDIS papilar é um subtipo do CDIS e se caracteriza como uma proliferação de células maligna restrita aos ductos, em pode ou não existir obstrução localizada dos mesmos (20). Na Figura 2.2b se observa um exemplo imagem de ultrassom de um caso confirmado por histologia desse tipo de câncer.

As outras neoplasias papilares malignas têm suas particularidades em relação a caracterização, sintomas clínicos e histologia. Como exemplo de imagem de ultrassom, temos o carcinoma papilar encapsulado na Figura 2.2c.

2.1.2.2 Lesões sólidas bem delimitadas

As lesões sólidas bem delimitadas se caracterizam como massas ou nódulos com região limite bem definida. Elas são classificadas como: lesões císticas, lesões sólido-císticas, e lesões sólidas (20).

As lesões císticas são não tumorais e benignas que se apresentam como nódulos ou cistos. Já as lesões sólido-císticas, como o papiloma intraductal e os carcinomas papilares

(a) **(b)** **(c)**

Figura 2.2. Exemplos de lesões neoplásicas papilares em exame de ultrassonografia bilateral. **(a)** é um caso de papiloma intraductal; **(b)** um caso de CDIS papilar; e **(c)** um caso de carcinoma papilar encapsulado. Fonte: (a) Stanislavsky(24) (2014); (b) Ashraf(25) (2023); (c) Knipe e Radswiki(26) (2011).

que foram descritos anteriormente, são aquelas que se caracterizam pela presença de massas ou nódulos associados ao derrame papilar. Podendo assim ser neoplasia benigna ou maligna, dependendo da sua caracterização (20).

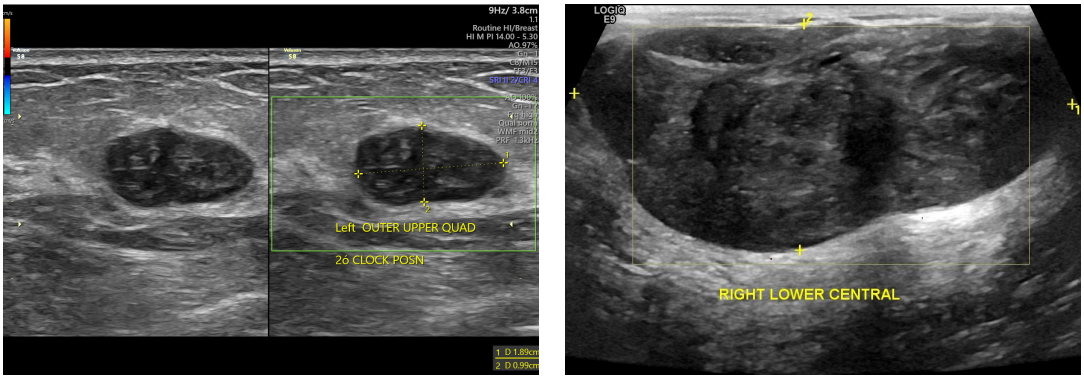
As lesões sólidas se caracterizam pela presença de algum tipo de massa ou nódulo palpável. Como exemplos citados temos: Adenose simples, adenomas, fibroadenomas, tumor filoide, e lesões mesenquimais (20).

A adenose simples é um processo benigno que ocorre na mama, muito comum na gestação, apresentando aumento da unidade funcional produtora de leite em um lóbulo mamário. A adenose tumoral, benigna, ocorre quando a adenose em diferentes lóbulos se converge em uma área (20).

O adenoma é a proliferação da quantidade de unidades funcionais produtoras de leite em um tumor bem delimitado de característica benigna e que não apresenta risco de se tornar maligno (20).

Os fibroadenomas são lesões fibroepiteliais e o tipo mais comum de neoplasia benigna entre mulheres jovens (entre 20 e 30 anos). Como ocorrem devido à proliferação de células constituintes dos lóbulos mamários, eles são mais sensíveis ao estímulo hormonal (como aumento de volume no fim do ciclo menstrual) (20). Na Figura 2.3a é apresentado o ultrassom de um caso confirmado de fibroadenoma em uma mulher de 40 anos.

O tumor filoide é semelhante ao fibroadenoma; porém, pode apresentar comportamento de lesão benigna, localmente agressiva ou maligna. A principal característica é crescimento exagerado e muita celularidade, em que os casos com mais de 4cm geralmente são malignos (20). Conforme observado no ultrassom da Figura 2.3b, esse tumor pode chegar a tamanhos consideráveis, nesse caso sendo de $6{\times}3$ cm^2.



(a) **(b)**

Figura 2.3. Ultrassom de dois casos confirmado por biópsia. **(a)** fibroadenoma na mama esquerda em uma mulher de 40 anos, e **(b)** tumor filoide de tamanho 6×3 cm^2 na mama direita em uma mulher de 45 anos. Fonte: (a) Awal(27) (2020); (b) Adaptado de Hacking e Farouk(28) 2017.

As lesões mesenquimais se originam no estroma interlobular, ou seja, no tecido conjuntivo com vascularização entre os lóbulos mamários. Dentre os diferentes tipos temos os tumores que podem variar de benigno, como lipomas, ou malignos, como os angiossarcomas. Em específico, os sarcomas são tumores malignos que não contém tecido epitelial, componente presente na maioria dos tumores mamários. A ocorrência desse tipo é rara, portanto, o diagnóstico geralmente ocorre no método de exclusão (20).

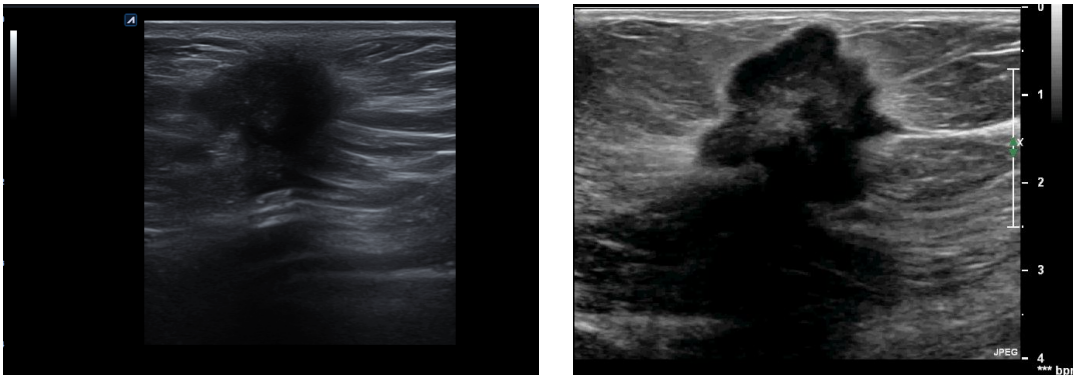
2.1.2.3 Lesões sólidas mal delimitadas

As lesões sólidas mal delimitadas, também denominadas como infiltrativas, se caracterizam pela penetração entre tecidos da mama, em que podem ser tanto do tipo benigna como maligna.

Dentre diversos subtipos, pode-se destacar os carcinomas. Em específico, o carcinoma invasivo tipo não especial (ou ductal invasor) é o câncer mais incidente na mama e muito agressivo, representando 70 % a 75 % do casos, conforme Siegel et al. (2021 apud Lucena(20), 2023, p. 40).

Segundo Siegel et al. (2021 apud Lucena(20), 2023, p. 40), outro câncer de mama frequente são os carcinomas do tipo especial, representando 25 % dos casos, como o Carcinoma Lobular Invasivo (CLI) sendo o mais frequentes nessa classe.

Na Figura 2.4a é apresentada uma imagem de ultrassom na mama esquerda de um caso de carcinoma ductal invasivo em uma mulher de 25 anos. Já na Figura 2.4b se apresenta um caso de carcinoma lobular de uma mulher de 70 anos na mama direita.



(a) **(b)**

Figura 2.4. Ultrassom de dois casos confirmados por biópsia: **(a)** caso de carcinoma ductal invasivo na mama esquerda em uma mulher de 25 anos, e **(b)** carcinoma lobular invasivo de tamanho 2,7×2,5 cm^2 na mama direita em uma mulher de 70 anos. Fonte: (a) Obaid(30) (2025); (b) Jin e Babu(31) (2012).

2.1.2.4 Microcalcificações e outras lesões neoplásicas

As microcalcificações são estruturas que podem estar presentes em grande parte das lesões mamárias. Em conjunto com outras características, elas fornecem indícios para definição do tipo de lesão durante o diagnóstico diferencial. Elas aparecem durante exames de diagnóstico precoce por imagem, porém são mais bem caracterizadas quando é feita a análise histológica da biópsia (20).

Outras lesões neoplásicas citadas por (20) são os linfomas, cânceres primários que se desenvolvem nos linfonodos das mamas; doença de Paget, um raro tipo que ocorre no mamilo, ilustrado por um exemplo de ultrassom na Figura 2.5; e o carcinoma inflamatório, um tipo de lesão que provoca inflamação da mama.

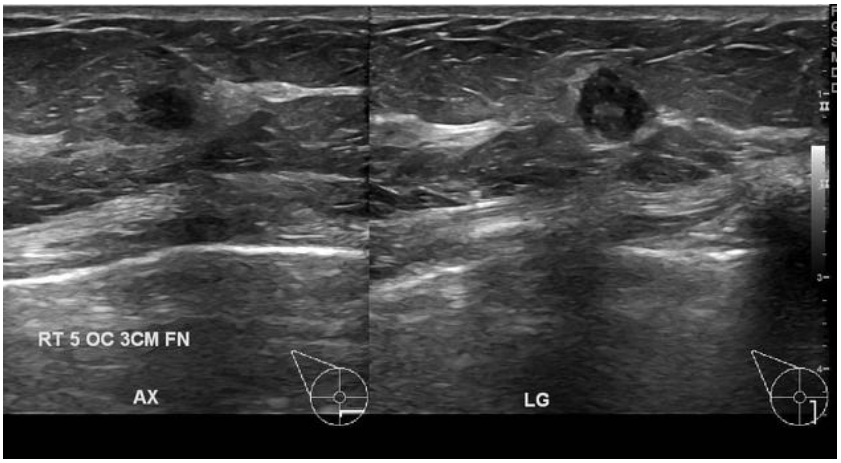


Figura 2.5. Ultrassom de caso confirmado por histologia para doença de Paget. Esse caso ocorreu em uma mulher de 55 anos na mama direita com sintoma de mamilo avermelhado. Fonte: (32).

2.2 Detecção precoce por imagem

Conforme apresentado anteriormente, as lesões mamárias são bem variadas e os casos de câncer em estágio inicial têm bom prognóstico de recuperação para o paciente. Atualmente, a recomendação do Ministério da Saúde do Brasil é o rastreamento de mulheres entre 50 e 69 anos por meio da mamografia e, nos casos abaixo dessa faixa etária, indica-se o uso de ultrassom (5, 6).

No caso das mulheres abaixo dos 50 anos, o uso do ultrassom é mais indicado devido a, estatisticamente, esse grupo conter mamas mais densas. A denominação de mama densa está relacionada àquelas onde a concentração de fibras é maior do que gordura. A maior concentração de fibras tende a dificultar a visualização de lesões em estágio inicial na mamografia pelo fato de que essas estruturas estão na mesma faixa de intensidade nesses tipos de imagens (20).

Mesmo sendo um tipo amplamente adotado, a ultrassonografia mamária é vista pelos especialistas como método complementar a mamografia. Algumas das dificuldades encontradas são características que distinguem lesões benignas e malignas, tornando-as semelhantes, além da dificuldade de análise do exame devido a artefatos. Isso torna a acurácia do exame dependente da experiência do avaliador e faz com que a quantidade de falso-positivos seja maior do que em outros exames (10, 20).

Segundo estudo sistemático de Rood (33), a ultrassonografia para detecção e rastreamento primário tem potencial de alta sensibilidade e especificidade, alcançando percentuais de 80, 1 % e 88, 4 % respectivamente. Em mamas densas, a sensibilidade se mantém alta comparado à queda observada nas mamografias, de 85 % para 47, 8 % a 65, 8 % (34, 35).

2.3 Sistema de Laudos e Registro de Dados de Imagem da Mama (BI-RADS)

Como observado previamente (Figuras 2.2, 2.3, 2.4 e 2.5), existe uma grande variedade de lesões e, no ultrassom, sua avaliação é padronizada pelo Sistema de Laudos e Registro de Dados de Imagem da Mama, do inglês *Breast Imaging Reporting and Data System* (BI-RADS).

O BI-RADS padroniza a avaliação das imagens em exames realizados com mamografia, ultrassom e ressonância magnética. Esse sistema foi desenvolvido pelo Colégio Americano de Radiologia, do inglês *American College of Radiology* (ACR), e é o padrão ouro usado no mundo para avaliação e classificação de imagens da mama (15).

Ressalta-se que o padrão ouro de diagnóstico de câncer é a realização da biópsia com avaliação histológica. Ou seja, como a confirmação de câncer é apenas feita após biópsia,

o objetivo dos exames por imagem é fornecer evidências suficientes que lesões identificadas tenham probabilidade de serem malignas para haver necessidade de biópsia. A biopsia é um procedimento invasivo, custoso e inviável de ser realizado em todos os casos suspeitos de câncer mamário devido à falta de infraestrutura.

Em uma avaliação usando BI-RADS para exame de ultrassom é realizada a caracterização da lesão usando componentes morfológicos de forma, margem, tamanho e orientação. Também consta na descrição dos efeitos que a lesão causa no tecido adjacente. As imagens de ultrassom geram características específicas que podem ser citadas, como capacidade distinta de reflexão das ondas de ultrassom entre tecidos (ecogenicidade) e artefatos presentes após atravessar uma lesão (reforço acústico ou aparecimento de sombras).

Em equipamentos de ultrassom mais caros, podem existir recursos adicionais que ajudam na avaliação, como a elastografia, que mede rigidez do tecido, e o doppler de potência. Alguns casos especiais contêm estruturas específicas que podem ser avaliadas pelo BI-RADS no ultrassom, como cistos simples e microcalcificações.

Após avaliação utilizando BI-RADS, a imagem e as lesões contidas nela são classificadas em sete categorias, definidas na Tabela 2.2.

Tabela 2.2. Classificações no BI-RADS quanto à probabilidade de ser câncer. Adaptado de (36, 20).

Classificação	Prob. de câncer	Controle
Categoria 0: Inconclusivo	-	Necessário novos exames por imagem
Categoria 1: Negativo	Aprox. 0 % de chance	Rastreamento de rotina
Categoria 2: Benigno	Aprox. 0 % de chance	Rastreamento de rotina
Categoria 3: Provavelmente benigno	$> 0\,\%$ a $\leq 2\,\%$ de chance	Acompanhamento em curto prazo (6 meses)
Categoria 4: Suspeito	$> 2\,\%$ a $< 95\,\%$ de chance	Biópsia
Categoria 4a: Baixa suspeita de ser maligno	$> 2\,\%$ a $\leq 10\,\%$ de chance	Biópsia
Categoria 4b: Média suspeita de ser maligno	$> 10\,\%$ a $\leq 50\,\%$ de chance	Biópsia
Categoria 4c: Alta suspeita de ser maligno	$> 50\,\%$ a $< 95\,\%$ de chance	Biópsia
Categoria 5: Altamente sugestivo de ser maligno	$\geq 95\,\%$	Biópsia
Categoria 6: Maligno confirmado por biópsia	-	Cirurgia de remoção (excisão cirúrgica) caso aplicável

2.4 Sistemas de auxílio ao diagnóstico por computador (CADx)

O desenvolvimento de sistemas de auxílio ao diagnóstico por computador (CADx) para exames por imagem busca melhorar a eficiência e precisão no diagnóstico, diminuindo custos operacionais e quantidade de falso-positivos.

Inicialmente, esses sistemas eram formados por algoritmos determinísticos que utilizam técnicas tradicionais de processamento de imagens para identificar as lesões nas imagens, realizar a segmentação e/ou auto contorno das bordas, e classificação geralmente binária (ser benigno ou maligno) (14).

Atualmente, esses sistemas utilizam modelos de Aprendizado de Máquina, do inglês *Machine Learning* (ML), um dos subgrupos da área de Inteligência Artificial (IA). Conforme compilação histórica feita por (14), com adições mais recentes, os principais marcos na área são:

- 1950: Introdução do teste de Turing para avaliar se computadores têm capacidade de pensar equivalente à de humanos (37).
- 1956: Fundação da área de pesquisa IA no *Dartmouth Summer Research Project on Artificial Intelligence*.
- 1958: Rosenblatt simula o perceptron, o algoritmo base do aprendizado supervisionado para classificação binária (38).
- 1966: Criação do primeiro *chatbot* que utiliza processamento de linguagem natural, codinome ELIZA, por Weinzenbaum (39).
- 1969: Minsky e Papert demonstraram vantagens e limitações do perceptron. Iniciando o que muitos denominam como inverno da IA, período de 10 anos com pouco investimento e incertezas na área (40).
- 1976: O sistema MYCIN foi apresentado, um dos primeiros sistemas aplicados na área médica com reconhecimento de padrões (41).
- 1986: Publicação do algoritmo de *backpropagation* por Rumelhart *et al.* (42).
- 1998: Introdução, por LeCun, da arquitetura de Rede Neural Convolucional, do inglês *Convolutional Neural Network* (CNN) para análise de documentos (43).
- 2006: Laboratório de Hilton apresenta melhoras substanciais no treinamento de DNNs (44, 45).
- 2012: Alex *et al.* ganham a competição ImageNet 2012 com uso de CNNs com várias camadas e demonstram seu uso em reconhecimento de objetos em imagens (46).

- 2017: Introdução da arquitetura dos *transformers* por pesquisadores da Google (47).
- 2021: Adaptação dos *transformers* para imagens, *Vision Transformers* (48).

2.5 Redes Neurais Profundas (DNN)

Por décadas, os melhores modelos de ML necessitavam de um pré-processamento cuidadoso para extrair características dos dados brutos que eram eficientes para detecção ou classificação de padrões da entrada. Uma das soluções encontradas para automatizar esse processo e melhorar o desempenho foi a DNN. Ela se baseia no princípio de automaticamente descobrir características úteis com uso de múltiplas camadas (49).

No caso de imagens, uma DNN é formada de múltiplas camadas em que, como exemplo, a camada um transforma a entrada e gera características treináveis sobre bordas em direções específicas; outra camada detecta grupamentos específicos de bordas, e as camadas seguintes buscariam variações mais específicas para detectar caraterísticas relevantes para o objetivo escolhido (49).

Em relação as formas de ML, a mais comum é o aprendizado supervisionado. Nessa modalidade, o banco de dados de treinamento contém, em conjunto com os dados de entrada, as categorias da característica em estudo. Essas categorias são usadas como âncora para o ajuste dos parâmetros durante o treinamento de um modelo de classificação. O objetivo é que a categoria desejada esteja na melhor pontuação comparada às outras (49).

Para classificação quanto à malignidade de lesões mamárias, diversos estudos recentes mostraram eficácia de DNNs em imagens de ultrassom (50, 51). Porém, os estudos na área se mantêm um tanto quanto restritos quanto à disponibilização dos modelos treinados para testes e, caso aplicável, dos dados utilizados quando não disponíveis online. Esses problemas dificultam a comparação entre DNNs e reprodutibilidade dos resultados. Principalmente quando estudos com modelos de estado da arte utilizam bases privadas de tamanho muito maior que os públicos (10, 52).

2.5.1 Estado da arte

O estudo usando DNNs para imagens de câncer de mama aumentou drasticamente nos últimos anos. Segundo Luo *et al.* (51), foram encontrados 366 artigos entre 2012 e 2022 nessa aplicação, em que mais da metade foi publicada nos últimos 3 anos. Para imagens de ultrassom, as arquiteturas propostas foram na maioria variações de CNNs para classificação (51). Contudo, arquiteturas mais recentes, como o *vision transformer* (ViT), evidenciaram um melhor desempenho na classificação de imagens no domínios naturais (48).

Em um estudo de 2019, Qi *et al.* (10) propuseram uma arquitetura de treinamento em cascata de duas CNNs que alcançou acurácia de 89, 39 % e sensibilidade 95, 29 % para classificação binária (maligno e benigno). Em (52) e (53) foram apresentadas classificações multiclasse quanto às categorias BI-RADS, resultando em uma Área sob a curva ROC, do inglês *Area Under the Curve* (AUC) de 87, 3 % e acurácia balanceada de 77, 45 %, respectivamente. No primeiro, é proposta uma arquitetura híbrida com mecanismo de atenção espacial para imagem de entrada, como principal, e atenção "BI-RADS" aplicada a cada canal. Nesta, são utilizadas características de um relatório BI-RADS (como formato e ecogenicidade), em conjunto com características espaciais calculadas da arquitetura principal, para classificação nas categorias BI-RADS. No segundo é utilizada uma arquitetura focada na extração de características espaciais em duas etapas, não necessitando de informações adicionais para seu treinamento.

Utilizando arquiteturas recentes, como *transformers*, Mo *et al.* (54) propõem uma modelagem mais anatômica ao se basear na extração de características da segmentação de tecidos distintos, alcançando uma acurácia de 85, 50 % no banco de dados BUSI (55). Semelhante ao indicado por (16) com um modelo multimodal, o uso dessa tecnologia em ambiente clínico precisa, além de acurácia melhor que especialistas da área, fornecer alguma evidência que suporte a decisão tomada e ser completamente automatizado. Ou seja, sem segmentação manual ou com Região de Interesse, do inglês *Region Of Interest* (ROI), predefinidas.

2.6 Métricas para avaliação de modelos de classificação

As arquiteturas e metodologia de treinamento de DNNs são complexas e tornam a sua avaliação difícil, caso não seja utilizado um padrão de avaliação. No caso dos modelos de classificação treinados de forma supervisionada, é pressuposto que o banco de dados em que foram treinados contém as classes corretas. Assim, é possível avaliar a predição de um modelo quando se considera a classe com maior valor de saída, ou maior probabilidade dependendo da arquitetura, como a correta.

Quando comparamos a classe correta com a prevista na classificação binária, existem quatro possíveis resultados que podem ser visualizados em uma matriz de confusão: Verdadeiro Positivo (VP), caso a classe correta e prevista sejam positivas (0); Falso Negativo (FN), caso a classe correta seja positiva (0) e a prevista negativa (1); Falso Positivo (FP), caso classe correta seja negativa (1) e a prevista positiva (0); e Verdadeiro Negativo (VN), quando a classe correta e prevista são negativas (1) (56). Na Figura 2.6 é ilustrado como esses resultados são dispostos na matriz de confusão.

Os valores contidos na matriz de confusão podem ser usados para calcular outras

		Classe inferida	
		Maligno	Benigno
Classe verdadeira	Maligno	VP	FN
	Benigno	FP	VN

Figura 2.6. Matriz de confusão para modelos de classificação binária de lesões. A classe maligna representa o resultado positivo, enquanto a benigna representa o negativo.

métricas comuns em estudos de redes neurais. Como acurácia, precisão, sensibilidade, especificidade, F1-*Score* e a curva Característica de Operação do Receptor, do inglês *Receiver Operating Characteristic* (ROC) (56). As quatro primeiras podem ser calculadas de três maneiras: para classificação binária, pela média das amostras (média micro) e pela média das classes (média macro) (57).

As métricas de média macro representam melhor aplicações multiclasse ou binárias em que os dados de teste são desbalanceados entre o número das classes e se avalia o impacto para ambos os casos (57).

2.6.1 Acurácia

A acurácia representa a probabilidade que a predição do modelo esteja correta. Assim, a acurácia binária (*bin*), a acurácia média micro (μ) e a acurácia média macro (M) são, respectivamente, calculadas conforme (56, 57):

$$Acc_{bin} = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FN} + \mathrm{FP} + \mathrm{VN}}, \tag{2.1}$$

$$Acc_{\mu} = \frac{\sum_{k=1}^{N} \mathrm{VP}_k + \mathrm{VN}_k}{\sum_{k=1}^{N} \mathrm{VP}_k + \mathrm{FN}_k + \mathrm{FP}_k + \mathrm{VN}_k}, \tag{2.2}$$

$$Acc_{M} = \frac{\sum_{k=1}^{N} \frac{\mathrm{VP}_k + \mathrm{VN}_k}{\mathrm{VP}_k + \mathrm{FN}_k + \mathrm{FP}_k + \mathrm{VN}_k}}{N}, \tag{2.3}$$

onde k é a classe em análise e N o número total de classes.

2.6.2 Precisão

A precisão representa a probabilidade que as predições positivas estão corretas, ou seja, o nível de confiança para predição de câncer. Assim, a precisão binária (*bin*), a

precisão média micro (μ) e a precisão média macro (M) são, respectivamente, calculadas conforme (56, 57):

$$P_{bin} = \frac{\text{VP}}{\text{VP} + \text{FP}}, \tag{2.4}$$

$$P_{\mu} = \frac{\sum_{k=1}^{N} \text{VP}_k}{\sum_{k=1}^{N} \text{VP}_k + \text{FP}_k}, \tag{2.5}$$

$$P_{M} = \frac{\sum_{k=1}^{N} \frac{\text{VP}_k}{\text{VP}_k + \text{FP}_k}}{N}, \tag{2.6}$$

sendo k a classe analisada e N o número total de classes.

2.6.3 Sensibilidade

A sensibilidade, também chamada *recall*, é a probabilidade da classe positiva ser corretamente predita. Assim, a sensibilidade binária (bin), a sensibilidade média micro (μ) e a sensibilidade média macro (M) são, respectivamente, calculadas conforme (56, 57):

$$Sens_{bin} = \frac{\text{VP}}{\text{VP} + \text{FN}}, \tag{2.7}$$

$$Sens_{\mu} = \frac{\sum_{k=1}^{N} \text{VP}_k}{\sum_{k=1}^{N} \text{VP}_k + \text{FN}_k}, \tag{2.8}$$

$$Sens_{M} = \frac{\sum_{k=1}^{N} \frac{\text{VP}_k}{\text{VP}_k + \text{FN}_k}}{N}, \tag{2.9}$$

onde k representa a classe em análise e N o número total de classes.

2.6.4 Especificidade

A especificidade é a sensibilidade aplicado para a classe negativa, ou seja, a probabilidade de a classe negativa ser corretamente predita pelo modelo. Assim, a especificidade binária (bin), a especificidade média micro (μ) e a especificidade média macro (M) são, respectivamente, calculadas conforme (56, 57):

$$Espec_{bin} = \frac{\text{VN}}{\text{VN} + \text{FP}}, \tag{2.10}$$

$$Espec_{\mu} = \frac{\sum_{k=1}^{N} \text{VN}_k}{\sum_{k=1}^{N} \text{VN}_k + \text{FP}_k}, \tag{2.11}$$

$$Espec_{M} = \frac{\sum_{k=1}^{N} \frac{\text{VN}_k}{\text{VN}_k + \text{FP}_k}}{N}, \tag{2.12}$$

onde k é a classe em análise e N o número total de classes.

2.6.5 F1-*Score*

O F1-*Score*, também chamado de F-*Score* ou F-*Measure*, é uma variante do $F_\beta - Measure$ em que a precisão e a sensibilidade têm igual importância e podem ser representadas por essa métrica ($\beta = 1$) (56, 58). O F1-*Score* binário (*bin*), o F1-*Score* por média micro (μ) e o F1-*Score* por média macro (M) são, respectivamente, calculados conforme (56, 57):

$$F_{1-bin} = 2 \cdot \left(\frac{P_{bin} \cdot Sens_{bin}}{P_{bin} + Sens_{bin}} \right), \tag{2.13}$$

$$F_{1-\mu} = 2 \cdot \left(\frac{P_{\mu} \cdot Sens_{\mu}}{P_{\mu} + Sens_{\mu}} \right), \tag{2.14}$$

$$F_{1-M} = 2 \cdot \left(\frac{P_{M} \cdot Sens_{M}}{P_{M} + Sens_{M}} \right), \tag{2.15}$$

sendo k a classe em análise e N o número total de classes.

2.6.6 Curva ROC

A ROC, é um gráfico usado para visualizar o desempenho de um modelo preditivo. Em que, a área sob a curva ROC (AUC) é uma métrica muito usada para ranqueamento de modelos classificadores. Uma propriedade da curva ROC é a invariabilidade para a classificação desbalanceada. Ou seja, mesmo uma quantidade maior de uma classe para outra, a curva ROC permanece a mesma (56).

Ela é calculada a partir da relação entre a taxa de verdadeiro positivo (Taxa VP) e a de falso positivo (Taxa FP), conforme (56):

$$\text{Taxa VP} \approx \frac{\text{VP}}{\text{VP} + \text{FP}} = P_{\mu}, \tag{2.16}$$

$$\text{Taxa FP} \approx \frac{\text{FP}}{\text{VN} + \text{FN}}. \tag{2.17}$$

3 Métodos implementados para classificação em imagens de ultrassom

3.1 Banco de Imagens de Ultrassonografia Mamária

Uma quantidade, variedade e generalidade suficiente de dados é um dos principais requisitos para uso de redes neurais como técnica para classificação de padrões (59). Porém, os bancos de dados públicos de imagens de ultrassonografia da mama ainda são comparativamente pequenos em quantidade de amostras. Nesse trabalho foram definidos três requisitos principais para escolha do banco de dados:

- Publicado em periódicos ou repositórios com revisão por pares e comitê de ética incluso.
- As classes anotadas precisam incluir o BI-RADS: pelo menos 3–5 (conforme Tabela 2.2).
- As classes foram totais ou parcialmente confirmadas por biópsia (estudo histológico).
- Conter máscara de segmentação manual das lesões.

A necessidade de confirmar por biópsia os rótulos “maligno”e “benigno”origina-se da hipótese de que a classificação puramente visual introduziria o erro humano dos especialistas.

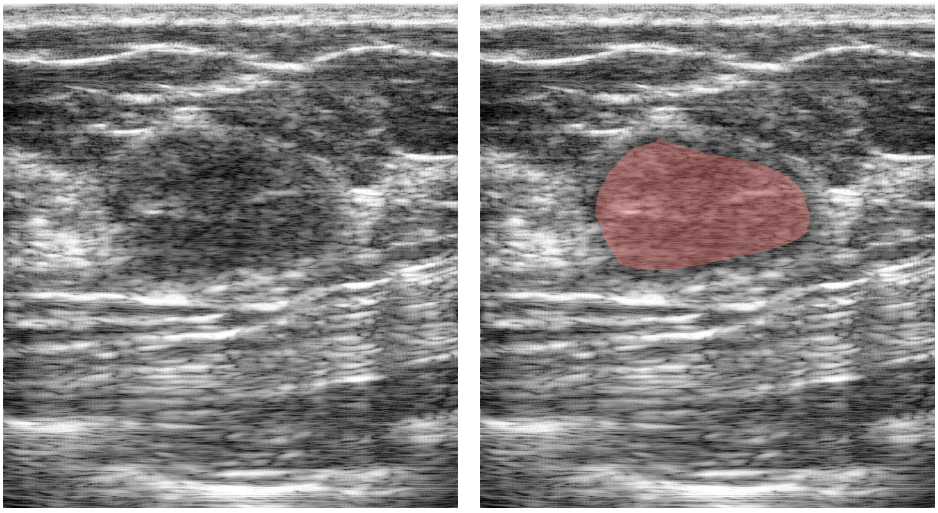
No total, foram encontrados três bancos de dados públicos que atendem a esses requisitos: OASBUD (60), BrEaST (61), e BUS-BRA (62).

3.1.1 Open Access Series of Breast Ultrasonic Data (OASBUD)

Nesse banco de dados foram extraídas 200 imagens de 100 pacientes do sexo feminino entre 24 e 75 anos (média 49,5 anos). As imagens são provenientes do Instituto de Oncologia na Varsóvia, Polônia. Cada paciente apresenta uma lesão que foi escaneada em dois planos diferentes, perpendiculares entre si: um é longitudinal e outro transversal. Isso totaliza 100 lesões distintas, sendo 52 malignas e 48 benignas (60).

No conjunto com lesões malignas, todas foram confirmadas histologicamente por biópsia de agulha grossa. Enquanto isso, no conjunto de lesões benignas, 37 dos 48 casos foram confirmados histologicamente e o restante por acompanhamento (60).

As imagens originais variam de tamanho entre 2608×510 pixels e 1040×510 pixels, em que todos os planos possuem a respectiva imagem da ROI para a lesão conforme exemplo na Figura 3.1. Para cada lesão existe anotação de cinco classes BI-RADS (3–5): 3 com 31 amostras, 4a com 11 amostras, 4b com 16 amostras, 4c com 17 amostras, e 5 com 25 amostras. Os resultados histológicos não foram incluídos nos dados disponibilizados (60). Devido à pouca quantidade de amostras, esses dados foram usados para teste dos modelos treinados.



(a) (b)

Figura 3.1. Exemplo de imagem de ultrassom reconstruída dos dados de ultrassom brutos (a) e respectivo contorno do tumor benigno (b), em vermelho. Fonte: (60).

3.1.2 BrEaST

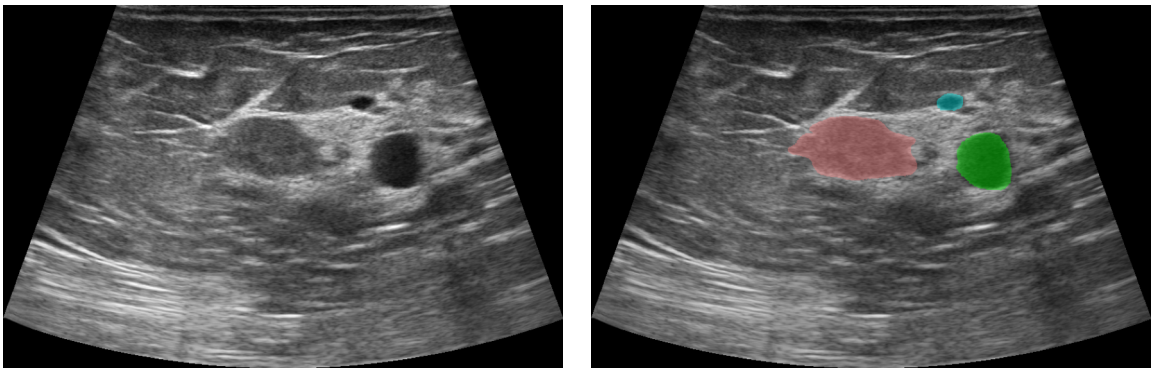
O banco de imagens BrEaST, Piotrzkowska *et al.* (61), contém 256 casos de pacientes distintas do sexo feminino entre 18 e 87 anos e sua coleta foi realizada em cinco centros médicos na Polônia entre 2019–2022.

Os dados são divididos em 154 lesões benignas, 98 malignas e 4 normais, em que 77 % deles passaram por confirmação histológica por biópsia. Também são incluídas as classes BI-RADS, de 1 a 5 com subtipo 4 (4a, 4b e 4c), e características descritivas da avaliação pela metodologia BI-RADS e aspectos gerais. Esses descritores são: composição do tecido, sinais patológicos observados pelo médico avaliador, formato da lesão, margem, ecogenicidade, artefatos posteriores, calcificações, presença de halo e espessamento da pele. No geral, o banco de dados BrEaST contém uma boa quantidade para cada

descritor, porém existem casos sem definição ou não disponível (61).

Também estão inclusas a interpretação do radiologista, o tipo de verificação realizada para confirmação da classificação (biópsia, exames de acompanhamento, não aplicável), e o diagnóstico baseado na histologia (61).

Um detalhe importante nas imagens do BrEaST é que elas não contêm artefatos artificiais que são geradas durante a aquisição e comuns nesse tipo de dado, como marcadores de medida, pictogramas, e anotações textuais. Para as máscaras de segmentação das lesões, manualmente delimitadas por profissionais, também podem existir máscaras adicionais que sinalizam anormalidades identificadas ao redor das lesões (61). Um exemplo de amostra do BrEaST e das respectivas máscaras pode ser visualizado na Figura 3.2.



(a) (b)

Figura 3.2. Um exemplo de imagem de ultrassom do BrEaST (a) e os múltiplos tumores identificados (b). Onde vermelho é o principal e azul e verde os secundários. Fonte: (61).

O uso desse banco de dados foi para teste dos modelos propostos. Devido à grande quantidade de informações fornecidas de cada caso é de se esperar que esse banco de dados se tornará um padrão para validação de outros modelos. Porém, como o banco foi disponibilizado recentemente, a quantidade de modelos publicados que utilizaram esses dados é baixa.

Durante a implementação para avaliação da classificação BI-RADS completa foi considerado que os subtipos do 4 (4a, 4b e 4c), são da categoria 4.

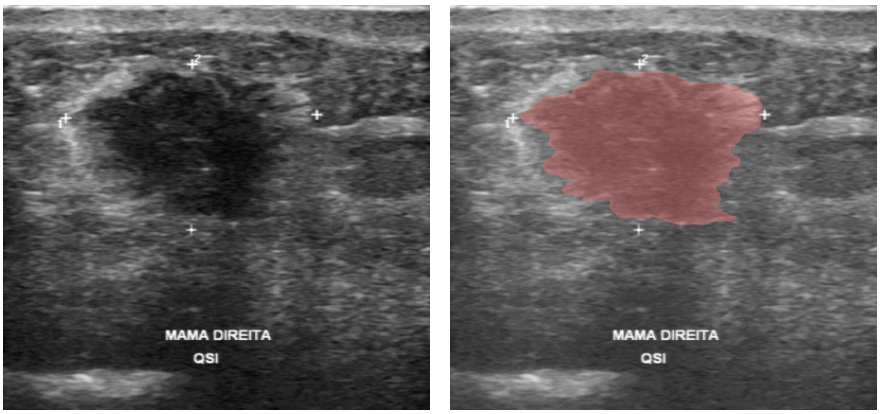
3.1.3 BUS-BRA: *Breast Ultrasound Dataset for Assessing CADx Systems*

O banco de imagens de Gómez-Flores *et al.* (62), denominado BUS-BRA, é o maior de seu tipo que atende os requisitos definidos. Nele estão presentes 1875 imagens de 1064 pacientes do sexo feminino entre 16 e 89 anos (média 47 anos).

As amostras são subdivididas em 722 casos benignos e 342 casos malignas, em que

todos foram confirmados por biópsia de agulha grossa. Para 76 % dos casos, foram adquiridas duas vistas ortogonais entre si, resultando em 1268 imagens de lesões benignas e 607 com lesões malignas (62).

Em relação à distribuição das imagens pelas categorias BI-RADS, temos: 562 na categoria 2, 279 na categoria 3, 393 na categoria 4, e 88 na categoria 5. Também estão disponíveis os resultados histopatológicos das biópsias nas anotações de cada caso. Na Figura 3.3 é apresentado um exemplo de caso de tumor maligno e respectiva ROI do BUS-BRA.



(a) (b)

Figura 3.3. Exemplo de imagem de ultrassom do BUS-BRA com tumor maligno (a) e respectivo contorno do tumor (ROI) (b), em vermelho. Fonte: (62).

Existem na literatura outras bases de dados com boa quantidade de amostras e bastante citados, como o de Al-Dhabyani *et al.* (55) que contém 697 lesões classificadas como benignas, malignas ou normais. Porém, a falta de confirmação por biópsia, a duplicação de amostras, a diferença entre máscaras ROI, e outras inconsistências podem sobrestimar o desempenho dos modelos treinados (63).

Em comparação com a base de dados BrEaST, o BUS-BRA contém mais de quatro vezes a quantidade de casos distintos e todos foram confirmados por biópsia. Esse fato favorece uma estimativa de desempenho mais realista dos modelos treinados para o diagnóstico de malignidade, por meio do cálculo das métricas indicadas. Contudo, as imagens do BUS-BRA contém artefatos artificiais que podem adicionar um viés ao modelo durante o treinamento. Assim, é necessário uso de outras bases, como BrEaST, para validação dos modelos.

Devido à quantidade superior de amostras, o BUS-BRA foi escolhido como base de dados principal para treinamento e teste dos modelos implementados nessa pesquisa. Em que, BrEaST e o OASBUD são aplicados como teste complementar aos modelos para comparação. Também foram utilizadas as imagens de (55) para testes, porém apenas para classificação quanto à malignidade (#1).

3.2 Pré-processamento

O pré-processamento realizado nesta pesquisa parte da necessidade de aplicar diferentes técnicas de processamento às imagens originais para garantir um melhor desempenho do modelo. Ao total, foram implementados quatro tipos distintos de pré-processamento.

O primeiro se baseia na hipótese de manter o máximo de informação espacial e de formato da lesão. Para isso, é calculado o centro de massa da ROI, definindo um ponto central nela conforme Figura 3.4a. Essa coordenada é usada como centro do retângulo de contorno, que sobrepõe os limites da ROI original acrescido de um fator de aumento linear de 20 % para cada lado (Figura 3.4b).

O fator de aumento linear, com valor de 20 %, foi definido após testes no banco de dados BUS-BRA, garantindo que, no mínimo, a ROI original estivesse contida no novo retângulo para maioria dos casos. Isso se deve ao fato de que o centro do retângulo, sendo o centro de massa da ROI original, está ligeiramente deslocado e pode não conter os extremos do contorno inicial.

Em seguida, é computado o tamanho do maior retângulo (N×M) e reaplicado a cada contorno previamente feito com esse novo tamanho. As coordenadas do centro de massa anterior coincidem com o centro do novo retângulo de contorno maior, conforme observado na Figura 3.4c. O novo contorno é preenchido com 1 e se torna a nova ROI de cada amostra (Figura 3.4d).

Com a nova ROI gerada, é realizado o processamento da máscara com a imagem original, gerando uma imagem segmentada igual exemplo da Figura 3.4e. O resultado dessa segmentação coloca no centro uma nova imagem, inicializada com zeros, de tamanho (N×M) igual observado na Figura 3.4f.

Esse processamento faz com que, mesmo quando a imagem de entrada sofre corte e redimensionamento antes de ser processada pelo modelo, seja mantida a informação espacial original para nova transformação.

A imagem resultante (Figura 3.4f) contém apenas um canal e serve como entrada para parte dos modelos propostos. Porém também é feito um processo de permutação desse canal para os outros 2, criando uma imagem de 3 canais (RGB) para os experimentos que necessitam de entrada RGB. Esse processamento totaliza dois dos quatro tipos propostos.

O pré-processamento #3 se baseia no proposto por Gómez-Flores *et al.* (62) e serve como base de comparação do desempenho. As técnicas propostas foram implementadas invertendo os canais vermelho e azul, levando à incompatibilidade de comparação dos modelos treinados nessa pesquisa com outras que seguiram exatamente a implementação original.

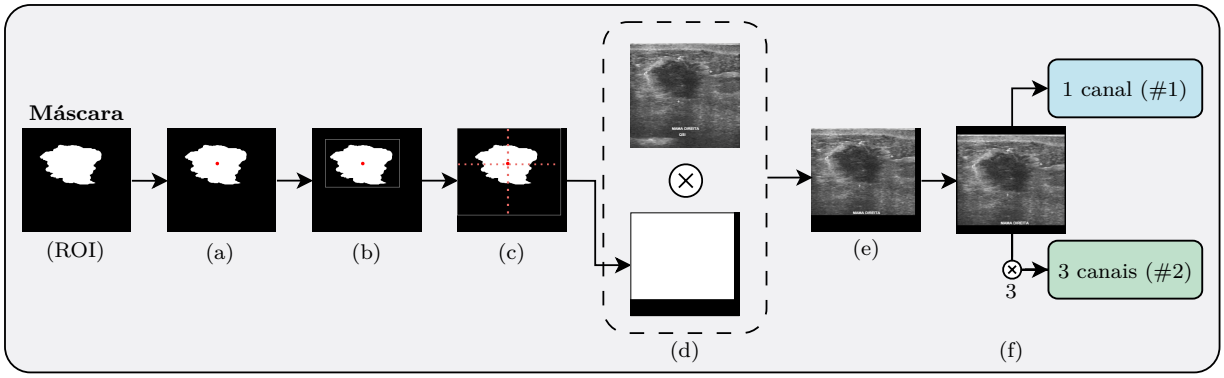


Figura 3.4. Pré-processamento proposto para as imagens de entrada e resulta em duas variações pela quantidade de canais (#1 e #2). (a) Na ROI original, calcula-se o centro de massa do contorno da lesão. (b) A área de segmentação é calculada com um fator de aumento linear de 20 % para cada lado. (c) Área de segmentação é aumentada, a partir do centro do contorno, para a maior dimensão existente entre as amostras (N×M). (d) Nova ROI é criada com preenchimento da área aumentada. Imagem de ultrassom é segmentada pela ROI, resultando em (e). A imagem de ultrassom segmentada é copiada para nova imagem zerada de dimensão N×M.

O fluxograma do pré-processamento #3 está presente na Figura 3.5. O processo começa com as mesmas etapas do primeiro pré-processamento até a etapa de cálculo do retângulo de contorno com fator de aumento de 20 % (Figura 3.4b). Essas coordenadas são aplicadas a imagem de ultrassom e respectiva máscara, realizando operação de corte (Figura 3.5c).

Em seguida, é aplicada a operação de alongamento de contraste na imagem de ultrassom. Nela é realizada uma normalização de máximo-mínimo, em que os níveis de intensidades são padronizados por terem sido coletados com diferentes configurações de ganho em equipamentos distintos (62). As imagens de ultrassom normalizadas são assim segmentadas para RGB (Figura 3.5d), em: canal vermelho (R) é uma cópia da ROI onde todas as posições maiores que zero viram 255; verde (G) é uma cópia do ultrassom segmentado pela ROI; e azul uma cópia do ultrassom.

A Figura 3.5e ilustra o resultado do pré-processamento #3 para uma amostra do BUS-BRA. A principal diferença do proposto por (62) é a modificação do retângulo de contorno, em que no original era aplicada uma tolerância de 10 pixels enquanto esse é de 20 % da área do retângulo.

O pré-processamento #4 vem da hipótese que DNNs não precisam de uma etapa anterior para extrair a ROI quando ela não existe, como outra DNN, para classificar as imagens de ultrassom. Portanto o único pré-processamento feito é a segunda etapa, ou seja, redimensionamento. Ela também é aplicada em todos os outros tipos e inclui a geração de um arquivo único com todos os dados necessários para treinamento, como as classes.

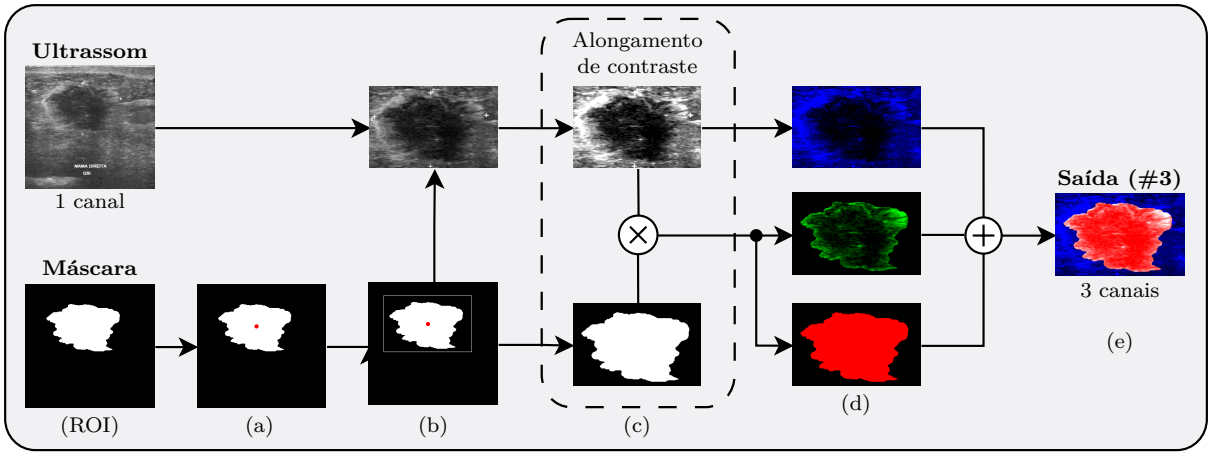


Figura 3.5. Pré-processamento #3 proposto, modificado de (62). (a) Na ROI original, calcula-se o centro de massa do contorno da lesão. (b) São calculados a área de segmentação com fator de aumento 20 %. (c) A imagem de ultrassom e ROI são segmentadas, em que, na original, é aplicado alongamento de contraste; (d) São definidos os canais RGB após segmentação; (e) Os canais RGB são mesclados para gerar imagem de saída.

Na segunda parte são gerados arquivos HDF5 para os diferentes tipos de classificação e pré-processamento e tamanho de entrada dos modelos escolhidos. Em relação à classificação foram feitas as variações apresentadas no Quadro 3.1.

Quadro 3.1. Classificações propostas para treinamento e validação dos modelos.

#	Categoria	Classe Positiva (A)	Classe Negativa (B)	Classes finais
1	Malignidade	maligno	benigno	maligno & benigno
2	BI-RADS	-	-	2, 3, 4, 5
3	BI-RADS	5	2	2, 5
4	BI-RADS	4,5	2,3	A & B
5	BI-RADS	5	2,3,4	A & B
6	BI-RADS	3,4,5	2	A & B

Durante o treinamento de um modelo existem transformações específicas que são aplicadas dependendo da arquitetura do modelo. No geral, após carregamento do arquivo HDF5, é feito um redimensionamento para um tamanho (N×M) e um corte centralizado com dimensão um pouco menor. Essa transformação é herdada das recomendações propostas quando os modelos foram treinados para a aplicação original, como para classificação do ImageNet-1k (64, 65, 66).

Após o pré-processamento, as imagens são separadas aleatoriamente em *subsets* de treinamento, validação, e teste, em respectivos percentuais de 70 %, 15 %, 15 %. Isso resulta, considerando apenas os dados do BUS-BRA, em 1312 imagens de treinamento, 282 para validação e 281 para teste para os experimentos sem exclusão de classes.

3.3 *Transfer Learning* (TL)

Para aplicação de DNNs no reconhecimento de padrões em imagens, é necessária, em relação a técnicas anteriores, grande quantidade de dados para melhorar o desempenho, principalmente em técnicas mais recentes como os *vision transformers* (48). Porém, os bancos de dados públicos utilizados contêm quantidade de amostras baixa, em comparação a aplicações de DNNs na literatura, conforme resumo observado na Tabela 3.1.

Tabela 3.1. Bancos de dados públicos de exames de ultrassom da mama com categorias BI-RADS até 2024. Em termos de malignidade, excluem-se os casos sem lesão (rótulo "normal"), restando apenas aqueles com rótulos "benigno" e "maligno"

Nome	Malignidade (benigno/maligno)	BI-RADS	Número total de casos
OASBUD(60)	48/52	3–5	100
BrEaST(61)	154/98	1–5	256
BUS-BRA(62)	722/342	2–5	1064

Diante das limitações observadas nas bases de dados, uma forma de melhorar o desempenho de DNNs é utilizando a técnica de transferência de conhecimento, conhecida em inglês como *Transfer Learning* (TL). Ela se baseia em utilizar modelos treinados em bancos de dados grandes e diversos, e, a partir do conhecimento prévio adquirido sobre aspectos físicos e visuais, refiná-los para uma aplicação específica. O processo de ajuste fino realiza o refinamento dos parâmetros do modelo para a nova aplicação, utilizando como entrada as imagens do sistema em estudo (17, 67).

A Figura 3.6 exemplifica esse processo, em que uma arquitetura CNN tem seu conhecimento transferido para classificação de lesões mamárias.

O TL pode ser aplicado de quatro formas distintas a modelos classificadores, variando conforme as partes do modelo original que são preservadas em termos de arquitetura e pesos gerados. Esse processo é denominado como congelamento total ou parcial da camada (17).

Em um extrator de características híbrido, as camadas de extração de característica são congeladas totalmente e as camadas de classificação são substituídas por novo modelo, como Máquinas de Vetores de Suporte, do inglês *Suport Vector Machine* (SVM), ou floresta aleatória (17).

No extrator de características, as camadas de extração continuam totalmente congeladas enquanto as camadas densas sofrem modificação parcial ou total da arquitetura e são retreinadas (17).

O ajuste fino, em inglês *fine-tuning*, é semelhante ao extrator de características, porém, parte das camadas de extração são retreinadas para ajuste de seus pesos. Por

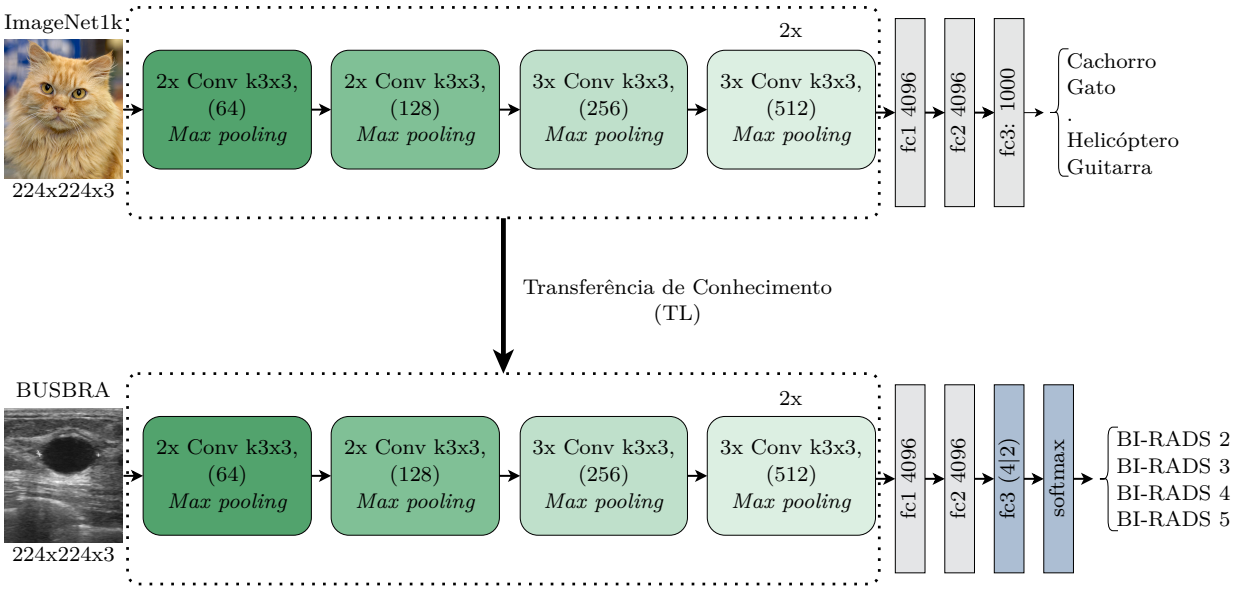


Figura 3.6. Exemplo de *transfer learning* (TL) entre domínio natural e a aplicação proposta.

último, existe o ajuste fino completo, onde todas as camadas do modelo são refinadas para a nova aplicação (17).

A escolha do tipo TL varia bastante de aplicação na área médica, não existindo atualmente um consenso quando aplicado á CNNs. Porém, o recomendado é começar pelo extrator de características, adicionando blocos de convolução mais profundos até chegar no caso de ajuste fino completo. No ajuste fino completo, o custo computacional é maior que outros casos devido a necessidade de treinar todas as camadas (17).

3.4 Aumento de dados

O aumento de dados, em inglês *data augmentation*, é uma outra técnica utilizada para melhorar o desempenho de DNNs quando a quantidade de dados de treinamento é insuficiente para alcançar bons resultados. A abordagem consiste no aumento artificial dos dados de treinamento com aplicação de diferentes transformações às imagens. Na aplicação em questão, o uso do aumento de dados para treinamento de modelos mais robustos é bastante usado, porém é necessária a correta escolha das transformações para melhora do desempenho (51).

Nesse sentido, a implementação aplicou mais transformações espaciais, como rotação e inversão, do que transformação de intensidade ou cor, como *color jitter*, com o intuito de evitar amostras artificiais fisicamente impossíveis de serem obtidas no mundo real. Para impedir a criação de viés e *overfitting* durante o treinamento, as transformações são aleatórias, evitando que duas amostras iguais sejam criadas.

No total é feito um aumento de dados de 29 vezes, realizando rotação com faixa variável de $[-30°, 30°]$ a $[-180°, 180°]$ utilizando transformação bilinear, inversão com probabilidade variável de 33 % ou 66 %, e ajuste de contraste com fator variável de 2 ou 3. Na Figura 3.7 são ilustrados 9 exemplos de dois pré-processamentos diferentes. Essa técnica é apenas aplicada para o *subset* de treinamento, em que seu número atualizado foi de 38.048 amostras.

Para implementação dos diferentes conjuntos de transformações é utilizada a função `Compose` do pacote *Torchvision*, em que todos são processados sequencialmente com a função `Sequential` do *PyTorch*.

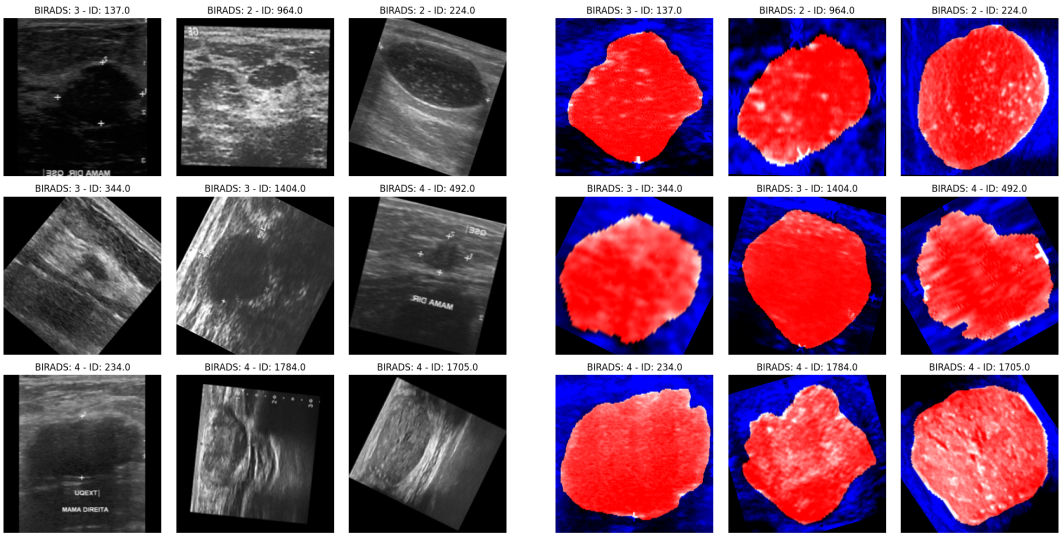


Figura 3.7. Exemplos do aumento de dados proposto. (a) São imagens do pré-processamento proposto #1 e (b) do pré-processamento #3.

3.5 Redes neurais convolucionais (CNNs)

Atualmente, as CNNs são muito utilizadas em soluções aplicadas a imagens e vídeos. Em um algoritmo de reconhecimento de padrões, 4 estágios básicos devem ser realizados: 1. aquisição; 2. pré-processamento; 3. extração de características; 4. classificação. A aquisição consiste no exame de ultrassonografia bilateral da mama, o qual gera as imagens brutas de entrada. O pré-processamento ajusta essas imagens para um padrão definido, com correção de contraste, redimensionamento para tamanho fixo e normalização. Na etapa de extração de características, são calculados atributos fundamentais para a diferenciação entre as classes. Por fim, a classificação é o processo que atribui à imagem a categoria com maior probabilidade (68).

A extração de características é o estágio mais difícil e que, geralmente, define a eficiência de um modelo de ML. Em uma arquitetura CNN aplicada a imagens, esse

estágio é feito pelas camadas de convolução 2-D, de ativação e de agrupamento 2-D (em inglês *2-D pooling*). Já a classificação é feita pelas camadas densas, em inglês *Fully Connected layers* (FC). Cada uma dessas camadas tem em sua composição parâmetros que são ajustados durante o treinamento supervisionado. No treinamento é utilizada a técnica de retropropagação (*backpropagation*) para ajustar retroativamente os parâmetros treináveis para todo os dados de treinamento (68).

A Figura 3.8 representa os componentes de um estágio de convolução, o qual é composto por três volumes: mapas de entrada, mapas de característica e mapas de agrupamento (*pooling*), em que o último não está presente em todos os estágios. Um volume é constituído de *arrays* 2-D de tamanho dependente das configurações para seu cálculo, porém todos têm o mesmo tamanho no mesmo volume. Em uma CNN, geralmente, múltiplos desses estágios — ou camadas que eles contêm — estão presentes e conectados em série (68).

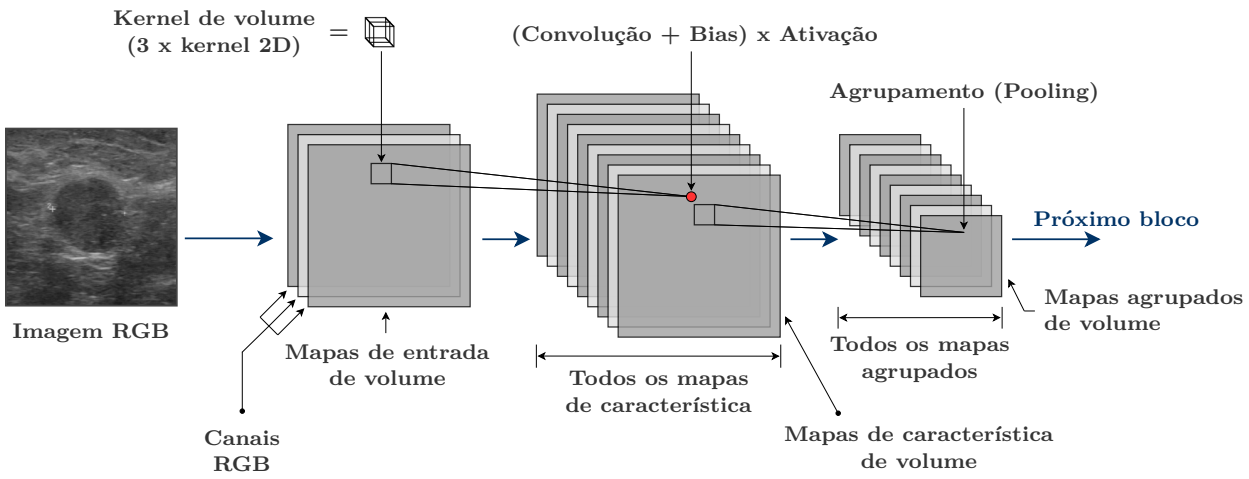


Figura 3.8. Componentes de uma camada de convolução em uma CNN. Adaptado de (69).

Uma imagem de ultrassom da mama, desconsiderando casos especiais como exames de Doopler de potência, é normalmente constituído de 1 canal (tons de cinza). Após passar pela etapa de pré-processamento, ela pode ser transformada para 3 canais (RGB), dependendo da aplicação. Caso a imagem de entrada tenha apenas 1 canal, o volume de entrada irá consistir em um mapa (tons de cinza). Caso haja 3 canais, igual ao mostrado na Figura 3.8, esse volume terá 3 mapas (vermelho, verde e azul) que são derivados dos canais RGB que constituem uma imagem RGB. No volume de entrada estão representados à altura, largura e profundidade, sendo a profundidade relacionada ao número de canais. A quantidade de mapas de volume de características depende da operação de convolução, porém o número de mapas de agrupamento sempre será igual ao de características da etapa anterior (68).

A operação de convolução, que dá nome a essas redes neurais, é aplicada no volume.

O kernel ou núcleo de volume, também denominado como resposta impulsional do filtro, se desloca apenas na altura e largura. Na Figura 3.8 temos uma representação do kernel de volume, que é composto de 3 kernels 2-D individuais que operam em cada um mapa de seus mapas de características (68). No caso de uma arquitetura para imagens de 1 canal, o kernel de volume seria um kernel 2-D.

Designa-se $\omega_{m,n,k}$ cada peso de um kernel 2-D, onde m e n são os índices que percorrem a altura e a largura do kernel, respectivamente, e k designa o índice do mapa de características ao qual o kernel está sendo aplicado. A operação de convolução entre o kernel e o k-ésimo mapa, em qualquer um dos pontos (x, y) nele contidos, é a soma dos produtos entre os pesos do kernel e os elementos do mapa em posições correspondentes a ele. Visto que a convolução volumétrica sobre uma entrada de 3 canais pode ser decomposta na aplicação de 3 kernels 2-D em paralelo, seu resultado em um ponto espacial (x, y) é a soma das convoluções de cada filtro 2-D com o respectivo canal do volume de entrada (68). Semelhante à forma que é calculado para um perceptron, também se pode adicionar ao somatório um viés (*bias*), b, para melhorar a convergência da rede. Em suma, pode-se definir a convolução volumétrica com *bias* como:

$$z_{x,y} = conv_{x,y} + b = \sum_i \omega_i v_i + b, \tag{3.1}$$

em que ω_i são os pesos do kernel, v_i são os valores respectivos dos elementos nos mapas de entradas que coincidem com o kernel, b representa o viés (*bias*), e $conv_{x,y}$ é a convolução volumétrica no mesmo ponto (x, y) para todos os mapas do volume de entrada (68).

O ponto vermelho, destacado na Figura 3.8, é o valor do escalar $z_{x,y}$ (Equação 3.1) obtido após aplicação da função de ativação ϕ. Uma função de ativação precisa ter três propriedades matemáticas para uso em redes neurais: não-linearidade, para que uma rede neural seja capaz de convergir para qualquer relação funcional entre variáveis, conforme estabelecido pelo Teorema da Aproximação Universal (70); intervalo finito ou infinito dependendo do método de treinamento; e ser uma função diferenciável. A última propriedade se baseia na melhoria de desempenho obtida pela otimização por gradiente descendente, porém não é necessária para todos os casos, considerando que existem funções que não atendem a esse quesito e apresentam resultado satisfatório, como a função *Rectified Linear Unit* (ReLU)(71). Dessa forma, o resultado obtido após adição da função de ativação ϕ a Equação 3.1 é dada por:

$$a_{x,y} = \phi(z_{x,y}) = \phi\left(\sum_i \omega_i v_i + b\right). \tag{3.2}$$

As funções de ativação mais usadas são a sigmoide $\phi(z) = 1/(1+\exp(-z))$ e a unidade retificada linear (ReLU) $\phi(z) = max(0, z)$, ambas amplamente utilizadas devido ao seu

bom desempenho em arquiteturas recentes. Após repetição do cálculo para todos os valores no mapa do volume de entrada, processo do qual se obtém o valor de $a_{x,y}$ (valor de ativação) para cada posição, é gerado um mapa de ativação ou mapa de características, conforme ilustrado na Figura 3.8. O tamanho do mapa de características vai depender do tamanho do kernel utilizado, do tamanho do preenchimento adicional feito na periferia do mapa de entrada (*padding*), e do passo (*stride*) do deslocamento do kernel entre cada valor do mapa de entrada (68).

Em uma imagem de ultrassom de tamanho 224×224, kernel 3×3 e *padding* de 1, ou seja, 1 fileira de zeros é adicionada em cada um dos lados. Caso o *stride* seja 1, uma convolução será calculada para cada posição, resultando em um mapa de características 224×224. Agora, utilizando um passo de 2 — efetivamente "pulando" um valor intermediário no mapa de entrada —, gera-se um mapa de tamanho 112×112. A quantidade de mapas de características em um volume é definida pela arquitetura, como exemplo ilustrado na Figura 3.8, onde são gerados 9 mapas de características (68).

O *pooling*, também chamado de subamostragem, realiza o processo de redução da resolução do mapa de características. Semelhante ao kernel 2-D, a subamostragem é definida por uma janela de determinado tamanho, na qual se realiza um cálculo sobre valores nela contidos, resultando em uma única variável. As operações matemáticas mais comuns são pela média, valor máximo (*max pooling*), e L2 *pooling*, este em que cada valor agrupado é a raiz quadrada da soma dos valores vizinhos ao quadrado (68).

A principais vantagens da camada de *pooling* são de reduzir o número de parâmetros treináveis na rede, aumentando a velocidade de treinamento em grandes bancos de dados, e diminuir a sensibilidade das características quanto à posição e rotação de objetos importantes, evitando *overfitting* do modelo final. A desvantagem principal é a diminuição do tamanho dos mapas de características, criando situações em que existe uma perda de informação excessiva, especialmente em arquiteturas profundas (68).

As arquiteturas CNN podem ser estruturadas de duas maneiras principais. A primeira utiliza apenas componentes de convolução em série, como estágio parcial ou completo da Figura 3.8, para segmentação da imagem de entrada por ROI. Como existe uma diminuição da dimensionalidade no decorrer das componentes, é necessário algum tipo de sobreamostragem que recupere o tamanho original da imagem (68). Um exemplo é a arquitetura U-net, geralmente aplicada na geração do contorno das lesões identificadas em uma imagem de ultrassom da mama, com base em ROIs manuais (72).

A segunda estruturação de uma CNN utiliza componentes de convolução para extrair mapas de características espaciais da entrada e adiciona, em série, uma camada densa para classificação. Essa camada, normalmente, é uma Perceptron Multicamadas, do inglês *Multilayer Perceptron* (MLP), na qual a entrada é um vetor de tamanho x e a saída é

a quantidade de categorias para classificação. O tamanho do vetor de entrada depende do vetor resultante da etapa anterior (convolução concatenada por linha para 1-D). Na Figura 3.9 é apresentado um exemplo de arquitetura CNN completa para classificação de quatro categorias BI-RADS.

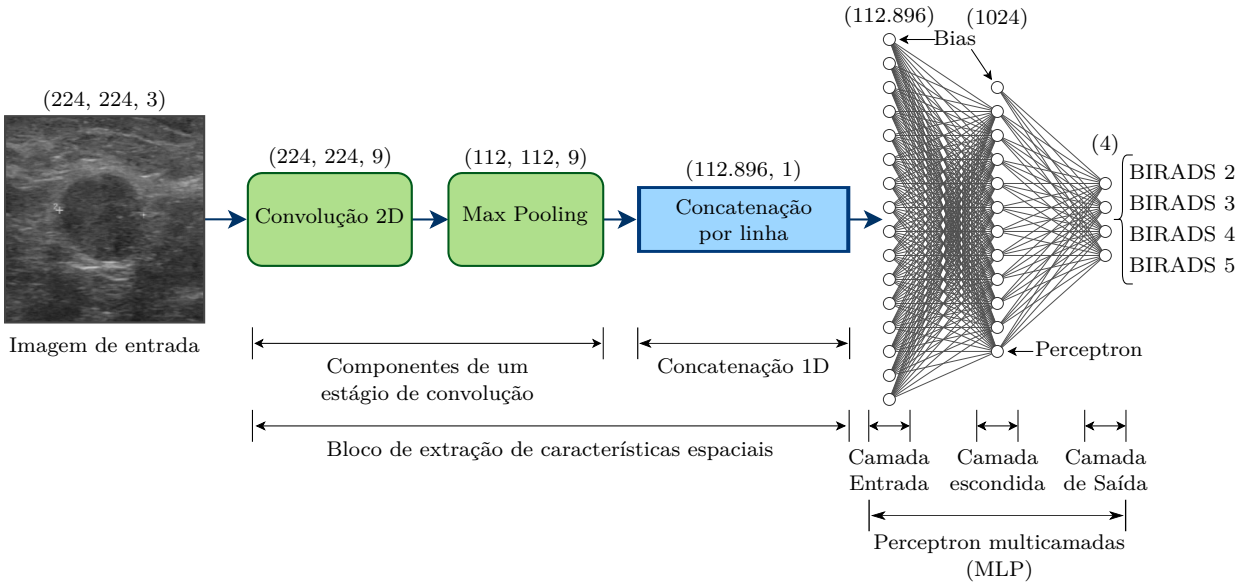


Figura 3.9. Exemplo de arquitetura CNN para classificação de quatro categorias BI-RADS. Adaptado de (69).

3.6 *Vision Transformers*

A arquitetura *transformer*, proposta por Vaswani *et al.* (47), utiliza o mecanismo de atenção como principal técnica para melhorar desempenho de modelos de linguagem natural, descartando o uso de blocos de convolução ou recorrência. O objetivo nessa arquitetura é paralelizar e diminuir o custo computacional para entender a relação entre entradas, principalmente quando estão longe uma das outras. Por exemplo, em outras arquiteturas aplicadas ao processamento de linguagem natural, quanto maior era a quantidade de entradas (palavras), maior era o custo computacional para calcular a relação entre a primeira e a última posição da entrada — constituída, neste caso, por unidades de dados que representam texto.

Na área de ML, um *token* é uma representação numérica de uma unidade dos dados de entrada. Para imagens, ele representaria um segmento da imagem de entrada ou de um mapa de características, dependendo da arquitetura, enquanto para texto seria uma palavra de uma frase (47, 48).

Um problema para implementação direta dos *transformers* para imagens é que nelas cada pixel seria um *token*. Consequentemente, imagens de dimensão relativamente pequena, como 224×224, necessitariam de uma arquitetura com entrada de 50.176 *tokens*

para entender a relação entre os pixels por atenção. Nessa dimensionalidade, o custo de treinamento e inferência da rede seria muito alto para o tamanho das imagens e já existem arquiteturas eficientes para processamento de imagens, como as CNNs.

Dentre soluções propostas para aplicação em imagens, temos os *vision transformers* de Dosovitskiy *et al.* (48). Eles propõem uma arquitetura que busca utilizar a maior parte da arquitetura *transformer*, porém aplicada de forma mais eficiente para imagens. Conforme ilustrado na Figura 3.10, uma arquitetura básica do *vision transformer* é formada de: um divisor de imagem que a separa em partes (*patches*) de tamanhos iguais; uma projeção linear das partes que foram achatadas; a incorporação de posições (1 a N) nos vetores lineares (*tokens*); a incorporação de posição (0) de um vetor (*token*) que representa a classe; o bloco do *encoder* do *transformer* com mesmas características definidas por (47); e o classificador, onde o vetor de saída do *encoder* na posição 0 (classe) é a entrada de uma MLP para definir a classe, semelhante a camada final de uma CNN (Figura 3.9).

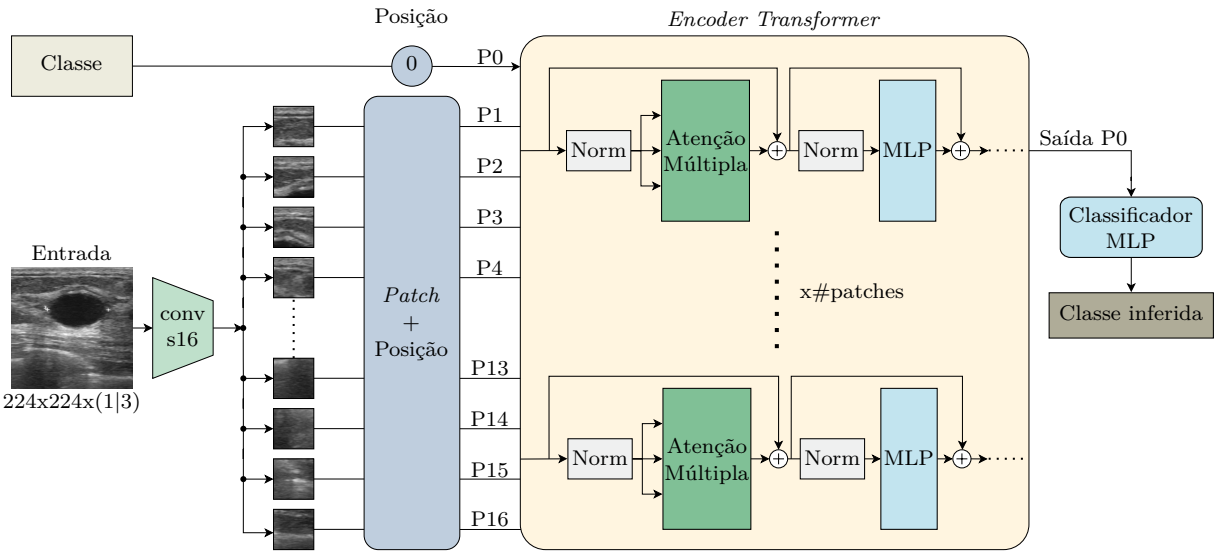


Figura 3.10. Arquitetura de um ViT aplicado para imagens. Imagem de ultrassom de (62). Adaptado de (48).

3.6.1 Imagem para partes (*patches*) e projeção linear

Uma imagem $x \in \mathbb{R}^{(L \times A \times C)}$ é dividida em partes menores $x_p^i \in \mathbb{R}^{(P \times P \times C)}$, em que (A, L e C) são a altura, largura e canal, respectivamente, P é o tamanho do *patch*, e i é o número do *patch* (48). Na Figura 3.10, uma imagem exemplo do ultrassom foi dividida em 16 *patches* ($i = 1, \ldots 16$).

Os *patches* são entradas sequenciais e numeradas de 1 a N da parte superior esquerda para a inferior direita de uma imagem. Em seguida, eles entram no bloco de projeção linear e passam pelas seguintes operações matemáticas: Primeiro o *patch* é redimensionado para um vetor (2-D para 1-D), nesse caso $x_p^i \in \mathbb{R}^{(P \times P \times C)} \rightarrow x_p^i \in \mathbb{R}^{1 \times (P^2 \cdot C)}$; em

seguida, o vetor é multiplicado por uma matriz de pesos (W) que gera o *patch* projetado linearmente, definido como z_1. Aplicando para a toda a imagem, é obtido:

$$z_1 = x_p \cdot W, \tag{3.3}$$

onde $z_1 \in \mathbb{R}^{N \times D}$, $W \in \mathbb{R}^{(P^2 \cdot C) \times D}$, $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, N é a quantidade de *patches* e D o tamanho do vetor de características (*token*). Essa operação ocorre em uma MLP sem *bias* (48), porém também pode ser implementada utilizando convolução 2-D.

A implementação por convolução 2-D evidencia-se mais eficiente que a MLP (73). Nesse caso, a convolução é aplicada na imagem de entrada com o kernel sendo os pesos (W), o *stride* igual o tamanho do *patch* (P) e a quantidade de canais de saída sendo o tamanho do vetor de características (D). Assim, a Equação 3.3 é reescrita como:

$$z_1^{i,D} = x_p^i * W_D, \tag{3.4}$$

em que i é a quantidade total de *patches*, $z_1^{1,D} \in \mathbb{R}$, $W_D \in \mathbb{R}^{P \times P \times C}$ e $x_p^i \in \mathbb{R}^{P \times P \times C}$.

3.6.2 Incorporação de posição (*position embeddings*) e classe

O *vision transformer* é invariante as transformações espaciais, como translação e escalonamento de objetos em uma imagem. Essa é uma característica presente na camada de *pooling* de CNNs; contudo, também existe uma noção de equivalência a transformações em suas camadas de convolução. Esta última é essencial quando o modelo é aplicado para segmentação e detecção de objetos, enquanto a invariância é essencial para classificação.

A incorporação de posição, do inglês *position embeddings*, realiza a identificação de cada *patch* ao relacionar um número que define sua posição. Esse processo adiciona ao *vision transformer* um viés indutivo em relação à posição unidimensional 1-D de cada *patch* $(1, \ldots, N)$, porque não houve melhoras ao se utilizar identificação 2-D. O restante das relações, como as espaciais intra e inter *patch*, são treináveis (48).

A implementação do *position embeddings* é feita aplicando uma codificação pela função senoide, conforme usado originalmente em (48), ou, mais comum nos modelos atuais, como um parâmetro treinável.

Por último, um vetor de zeros de tamanho D (*token*) é colocado na posição 0. Ele representa o vetor de características treináveis no *transformer*, em conjunto com os *tokens* de cada *patch* da imagem, que atribui o rótulo (classe). O vetor de saída do *encoder transformer* é a entrada do bloco classificador da arquitetura, como um MLP (48).

3.6.3 Encoder *Transformer*

O encoder *transformer* é formado de múltiplos blocos com dois componentes principais, atenção múltipla e MLP, e dois complementares, camada de normalização e de "pular conexão" (47).

A componente de atenção (A), em termos simplificados, apresenta função de aprender as relações e o contexto das suas entradas, neste caso as informações contidas nos *patches* e o contexto global da imagem. Ela é calculada conforme a expressão:

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3.5}$$

onde Q é o vetor de perguntas (*Query*), representando as características de interesse; K é o vetor de respostas (*Key*), representando as características que podem ser relevantes para as características de interesse; $\sqrt{d_k}$ é um fator de normalização, sendo que d_k é a dimensão dos vetores *keys* e *queries*; função `softmax` normaliza valores reais calculados em uma distribuição de probabilidades que soma 1; e V representa a matriz de características originais (valores) a ser escalonada pelas probabilidades resultantes da atenção (47).

No caso da atenção múltipla, os cálculos da Equação 3.5 são feitos múltiplas vezes em paralelo com distintos Q, K e V para cada *embedding*. Os resultados são mudanças distintas do *embedding* original, baseadas na atenção calculada. No final, a saída do bloco de atenção múltipla é a soma desses resultados mais o *embedding* original (mecanismo "pular conexão")(48).

Em seguida, um bloco que contém MLPs recebe a saída do bloco de atenção, onde cada entrada é uma MLP isolada do tipo *feedforward*. Essa MLP tem 2 camadas escondidas e utiliza a função de ativação *Gaussian-Error Linear Unit* (GELU), esta baseada em uma aproximação suave da ReLU (48).

O primeiro bloco complementar, a camada de normalização e presente antes do bloco de atenção múltipla e do MLP, conta com função de normalizar as entradas em cada bloco. A principal vantagem encontrada foi na diminuição do tempo de treinamento e sua estabilização, tornando a convergência do modelo mais suave (47). O segundo é o mecanismo de "pular conexão", que consiste em somar a representação original aos resultados calculados após cada bloco, seja ele de normalização e atenção múltipla, ou normalização e MLP. Na qual, seu uso apresentou maior desempenho no *transformer* quando as diferentes representações aprendidas são distribuídas para blocos mais longe do início (47).

4 Modelos propostos para classificação de lesões mamárias e metodologia de avaliação

A aplicação de redes neurais para classificação de imagens da mama, além da necessidade de um grande volume de dados como explicado anteriormente, requer arquiteturas e otimizações para atingir resultados de estado da arte. Os modelos propostos foram pré-treinados com diferentes bancos de dados do domínio natural, como ImageNet-1k, ImageNet-12K, ou LAION-2B (66, 74).

4.1 Redes Neurais Convolucionais (CNNs)

Conforme explicado no capítulo anterior, seção 3.5, as CNNs para classificação são compostas de um bloco de extração de características e um denso. Atualmente existem diversas variantes com diferentes abordagens quanto a estrutura do bloco de extração de características. Como as CNNs precisam de uma quantidade menor de dados em comparação aos *vision transformers*, elas são o foco principal deste estudo, no qual são propostas cinco arquiteturas distintas.

A escolha das cinco arquiteturas CNN foi, primeiramente, definida pelo *hardware* de treinamento disponível, GPUs com 8GB de memória VRAM. Essa quantidade de memória para imagens de entrada com até 300x300 pixels de tamanho em um *batch* de 32 amostras limita a complexidade das arquiteturas escolhidas. Ademais, foram utilizados os seguintes critérios para cada arquitetura CNN:

- **VGG16**: Apresentou bom desempenho na classificação de imagens de mamografia mamária em (75) e foi utilizado como modelo referência para comparação em (76) e (54).
- **ResNet50D**: O modelo ResNet50 é utilizado como referência em diversos estudos (52, 53, 54, 62, 76, 77, 78, 79) e pode representar como base de desempenho na arquitetura CNN. Em específico, a variante ResNet50D apresentou maior desempenho que a ResNet50 na classificação do *benchmark* ImageNet1K (65), mas não foi avaliado para imagens de ultrassom da mama.

- **EfficientNet-B3**: Nos resultados de validação do ImageNet-1k (80), esse modelo apresentou a maior acurácia (82,1 %) para imagens de entrada com tamanho 300×300 pixels ou menor.
- **RegNetY-3.2GF**: Nos resultados de validação do ImageNet-1k (80), o "regnet_32" foi o segundo com maior acurácia (82,0 %). Considerando os parâmetros treináveis reportados, 19,44 milhões, foi considerado que se trata da variação Y da família RegNet, ou seja, o modelo RegNetY-3.2GF.
- **MixNet-XL**: Nos resultados de validação do ImageNet-1k (80), o MixNet-XL foi o sétimo com maior acurácia (82,1 %). Desconsiderando modelos com maior acurácia (EfficientNet-B3, RegNetY-3.2GF e ResNet50D), os não encontrados na biblioteca devido à mudança de versão ("skresnext50d_32x4d" e "seresnext50d_32x4d") e a necessidade de ser uma arquitetura base diferente ("efficientnet_b2a"), esse modelo é melhor comparado aos demais.

4.1.1 VGG16

O *Visual Geometry Group* (VGG) é uma das primeiras famílias de CNNs que utilizaram várias camadas e alcançaram resultados de estado da arte para ImageNet-1k. Desenvolvido por pesquisadores da universidade de Oxford, cada VGG se baseia no uso de 16 a 19 camadas, com a maioria sendo de convolução (81).

Em específico, a VGG16 contém 13 camadas convolucionais com kernel 3x3 e 5 de *max pooling* 2×2, divididas em 5 componentes. O bloco denso é composto de 2 camadas escondidas com 4096 perceptrons cada e uma de saída para as 1000 classes da ImageNet-1k, esta que é normalizada pela função `softmax` (81). Para aplicação médica, a VGG16 apresentou resultados promissores para classificação de lesões em mamografias (75). Portanto, sua escolha foi para ser a base de desempenho para outros modelos CNN propostos.

Para aplicação proposta, é necessário alterar o bloco denso para duas variações: classificação de quatro classes BI-RADS e binária, esta usada para diferentes combinações do BI-RADS ou quanto à malignidade. Assim, é alterada a camada de saída de 1000 para 4 ou 2, dependendo do tipo de classificação proposta e é aplicada para todos os modelos propostos restantes.

O modelo VGG16 teve seus pesos pré-treinados para ImageNet-1k (64, 81). A Figura 4.1 apresenta a arquitetura da VGG16 alterada para classificação de 4 classes.

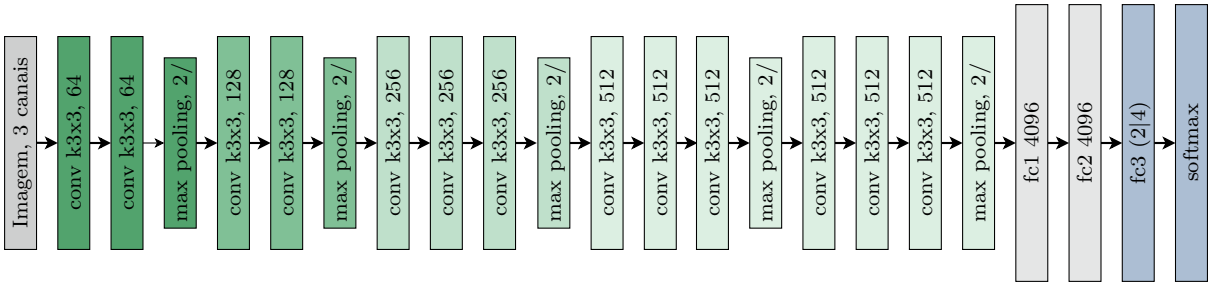


Figura 4.1. Arquitetura VGG16 proposta para classificação de duas ou quatro classes, dependendo da aplicação. Adaptado de: (81).

4.1.2 ResNet50D

O sucesso da VGG abriu oportunidades de pesquisa para treinamento de CNNs mais profundas, pois ela continha problemas de degradação com aumento da profundidade (82). Conforme proposto por (82), uma solução encontrada foi utilizar o princípio de aprendizado residual e aplicá-lo a uma nova arquitetura, denominada de ResNet.

O aprendizado residual é baseado na hipótese que múltiplas camadas não lineares, como convolucionais, podem se aproximar a funções residuais assintoticamente quando a entrada e a saída têm o mesmo tamanho. Por exemplo, digamos que o resultado de múltiplas camadas de convolução $a_{x,y}(entrada)$ (Equação 3.2) se aproxima a $H(entrada)$. Portanto seria possível afirmar que esse resultado também se aproximaria a uma função residual, como $H(entrada) + entrada$ (82).

A arquitetura ResNet aplica o aprendizado residual adicionando conexões de atalho entre duas ou três camadas de convolução, conforme Figura 4.2. Outra mudança foi que, quando o tamanho do mapa de características diminui pela metade, o número de filtros dobra, assim preservando a complexidade temporal. Assim como na VGG, em uma ResNet, também são removidas as camadas de *pooling*, nas quais a subamostragem é feita por convoluções com *stride* 2 (82).

O modelo avaliado foi o ResNet50D, uma implementação com pequenas modificações do original de (82) que resultam em melhor desempenho de acurácia para ImageNet (83). A variante D modifica as camadas de subamostragem que realizam convolução com *stride* de 2 e 2048 kernels 1×1, para duas camadas: primeiro uma de *pooling* de média 2×2 e depois a mesma de convolução com *stride* de 1.

O modelo utilizado teve seus pesos pré-treinados para ImageNet-1k com os aprimoramentos de treino de (65) e implementado para (64). A arquitetura foi modificada para classificação de 4 classes é ilustrado na Figura 4.2.

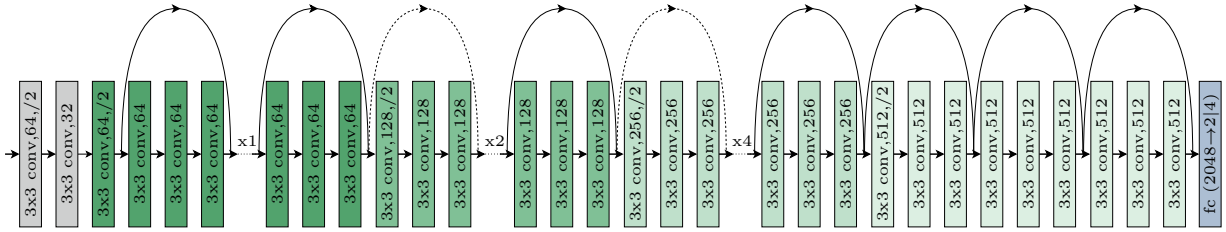


Figura 4.2. Arquitetura ResNet50D proposta para classificação de duas ou quatro classes, dependendo da aplicação. Adaptado de: (82).

4.1.3 EfficientNet-B3

As CNNs mais profundas melhoraram o desempenho para classificação aplicada ao ImageNet, porém ocasionou necessidades computacionais cada vez maiores (84). Devido a isso, sua aplicação em sistemas CADx, principalmente para regiões vulneráveis onde seja necessário processamento local, é inviabilizada por necessitar de soluções em nuvem e internet.

Com objetivo de melhorar desempenho sem necessariamente aumentar o custo computacional, Tan *et al.* (84) propuseram uma família de arquiteturas mais eficientes, denominadas de EfficientNet. Baseada na arquitetura MnasNet, proposta por (85), esses modelos buscam otimizar tanto a acurácia como custo computacional, *Floating-Points Operations Per Second* (FLOPS), utilizando a técnica de busca espacial de (85) nesse novo objetivo. Essa técnica, denominada como Busca Automática de Redes Neurais Artificiais, do inglês *Neural Architecture Search* (NAS), busca a melhor arquitetura para um tipo de aplicação e objetivo de forma automatizada.

No caso do EfficientNet, a busca proposta foi variação uniforme das dimensões de largura, profundidade e resolução presentes nos componentes convolucionais de CNNs, denominada como método de escala composta. O resultado gera uma arquitetura base (EfficientNet-B0) que contém 5,3M parâmetros treináveis (0.39B FLOPS) e alcança resultados melhores que ResNet50, que precisa de 4,9x mais parâmetros (11x mais FLOPS) (84).

A partir da arquitetura base, são criadas arquiteturas maiores até a B7, essa com 66M de parâmetros. A arquitetura implementada nesse trabalho foi a EfficientNet-B3, que contém 12M de parâmetros treináveis e custo de 1,8B FLOPS (84). O modelo final com adição da classificação de quatro classes é apresentado na Tabela 4.1.

A arquitetura original teve seus pesos pré-treinados para ImageNet-1k com os aprimoramentos do treinamento de (65) e implementado para (64).

Tabela 4.1. Arquitetura EfficientNet-B3 proposta. Adaptado de: (84).

#	Operador	Resolução	# Canais	Camadas
1	Conv2D (3x3)	288×288	40	1
2	MBConv1, k3x3	144×144	24	1
3	MBConv6, k3x3	144×144	32	3
4	MBConv6, k3x3	72×72	48	3
5	MBConv6, k3x3	36×36	96	5
6	MBConv6, k5x5	18×18	136	5
7	MBConv6, k5x5	18×18	232	6
8	MBConv6, k3x3	9×9	384	2
9	Conv1x1&*Pooling*&FC	9×9	1536	4

4.1.4 RegNetY-3.2GF

A arquitetura RegNet foi desenvolvida com uma técnica semelhante ao do EfficientNet para encontrar modelos mais eficientes. Nesse caso, Radosavovic *et al.* (86) propuseram o processo de design de espaço de design, este que busca simplificar uma população de modelos em um espaço de design sem precisar avaliar cada um individualmente.

O espaço de design do RegNet é uma versão restrita da primeira proposta dos autores, o AnyNet, em que é aplicada uma parametrização linear para as larguras do bloco original como restrição, explicado em mais detalhes por (86).

Como resultado, (86) propuseram a arquitetura RegNetX e, quando avaliadas utilizando a blocos de *squeze-and-excitation* (87), a RegNetY. O bloco de *squeze-and-excitation* calcula a interdependência entre os canais do mapa de característica (comprimir) e destaca as características mais importantes (excitação) (87).

Quando comparado ao EfficientNet, o RegNetY é melhor nas variantes com média necessidade computacional, como a RegNetY-1.6GF e EfficientNet-B3 com 1,6B e 1,8B FLOPS, respectivamente. No caso foram testados a RegNet-3.2GF, o que equivale em desempenho entre EfficientNet-B3 e B4 e contém 19M de parâmetros treináveis.

O modelo utilizado para TL foi pré-treinado com ImageNet1K-V2 com treinamento otimizado e pesos no *PyTorch*, especificamente no *Torchvision* (64).

4.1.5 MixNet-XL

As camadas convolucionais das CNNs geralmente utilizam um tamanho de kernel na mesma camada. Com proposta por (88) para verificar desempenho de múltiplos kernels na mesma camada, a convolução profunda mista (MixConv) realiza operações de convolução paralelas com diferentes grupos de kernels em diferentes profundidades da rede.

A definição das três arquiteturas MixNet propostas originalmente por (88) utilizou a busca automatizada por NAS, semelhante às arquiteturas anteriores implementadas.

Como exemplo, a arquitetura MixNet-L contém diferentes grupamentos de tamanhos de kernel que variam de 3×3 até {3×3, 5×5, 7×7, 9×9}. Foram alcançados resultados de estado da arte para ImageNet1K com custos computacionais abaixo da aplicação a dispositivos móveis (<600M FLOPS) (88).

A arquitetura implementada foi a MixNet-XL, baseada nas originais de (88), que contém 11,9M parâmetros e custo computacional de ≈ 900M FLOPS. A Figura 4.3 ilustra a arquitetura proposta em termos das combinações de kernels.

O modelo utilizado foi pré-treinado com ImageNet-1k com otimizações de treinamento de (83) (64).

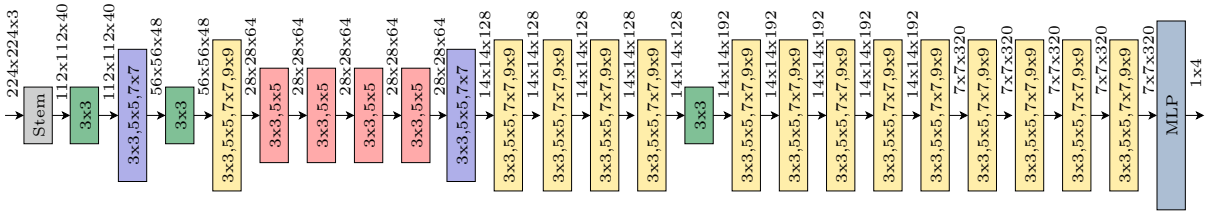


Figura 4.3. Arquitetura MixNet-XL proposta em termos das combinações de kernels de convolução. Adaptado de: (88).

4.2 *Vision Transformer* (ViT)

Conforme a Seção 3.6, os *vision transformers* (ViTs) dizem respeito a um tipo de arquitetura que utiliza um *encoder transformer* em *patches* da imagem de entrada para classificação (48).

O desafio para sua implementação vem da necessidade de um grande volume de dados e recurso computacional para alcançar resultados de estado da arte (48). Nesse sentido, o modelo avaliado nessa pesquisa, versão ViT base com 16 *patches* de (48), foi inicialmente pré-treinado utilizando a técnica CLIP (*Contrastive Language-Image Pre-training*) de (89), disponível em (90).

A técnica CLIP, do inglês *Contrastive Language-Image Pre-training* aplica o pré-treinamento contrastante para melhorar o desempenho de modelos para classificação de imagens utilizando dados de texto (89). Essa técnica evidenciou melhora expressiva ao estado da arte para aplicação médica de identificação de anomalias em raio-X do tórax (91).

Para pré-treinamento do ViT foram utilizados o *encoder* de imagem por *patches* em conjunto com um de texto, este que representa uma frase descrevendo a imagem e é

um *transformer*. Ambos são entrada em um *encoder transformer* que calcula a atenção entre os diferentes *patches* da imagem com as características calculadas pelo *encoder* de texto (89). Os dados de pré-treinamento é a base de dados LAION-2B, que é pública e contém 2,2 bilhões de amostras imagem-texto (74).

Em seguida, o modelo pré-treinado é refinado por TL para classificação *one-shot* de imagens, esta que é só uma denominação diferente do que é realizado nessa pesquisa (89). No modelo ViT-Base com 16 *patches*, implementado nesta pesquisa, seguiu um processo de treinamento que incluiu: pré-treinamento com LAION-2B, um primeiro ajuste fino no ImageNet-12k e, posteriormente, um segundo ajuste fino no ImageNet-1k (64, 92).

A arquitetura proposta altera o classificador MLP de 1000 para 4 (ou 2) perceptrons de saída e adiciona a função `softmax`, conforme Figura 3.10.

4.3 Definição dos Hiperparâmetros

Os hiperparâmetros, em inglês *hyperparameters*, representam as configurações dos algoritmos utilizados para o treinamento de uma rede neural. A escolha tanto do algoritmo como dos hiperparâmetros define a capacidade que uma rede neural terá de se ajustar aos dados de treinamento e, consequentemente, seu desempenho de inferência.

O otimizador utilizado nessa pesquisa foi escolhido com base em resultados encontrados em modelos protótipo realizados durante o desenvolvimento. Foram testadas variações de uma arquitetura CNN customizada e no ResNet50D algoritmos de otimização comuns da literatura, como SGD, ADAM, ADAMW e RMSProp (93, 94, 95, 96). O ADAMW apresentou melhores resultados tanto na classificação multiclasse como binária, esta que utiliza uma implementação de quantização `8 bits` de (97). Os hiperparâmetros utilizados foram: taxa de aprendizado $lr = LR_0$; betas [0,9, 0,999]; $eps = 1e^{-5}$; coeficiente de decaimento dos pesos $weight\_decay = 1e^{-2}$.

O algoritmo de agendamento da taxa de aprendizado e que tem função de variá-la durante o decorrer do treinamento, foi o *StepLR*, em que a variação do aprendizado é dada por:

$$LR(x) = LR_0 \cdot \gamma^x, \tag{4.1}$$

sendo x a época do treinamento, LR_0 a taxa de aprendizado inicial, e γ o fator de decaimento. Nos resultados apresentados todos os modelos com identificação de amostrador 1 e 2 foram treinados com $LR_0 = 1e^{-6}$ e $\gamma = 0,99$. Enquanto, os com amostrador 3 e 4, utilizaram $LR_0 = 1e^{-4}$ para maioria com exceção do VGG16, este utilizando $1e^{-3}$ e γ variável de 0,5 para VGG16, 0,35 para ResNet50D e EfficientNet-B3, 0,25 para MixNet-XL, e 0,2 para o RegNetY-3.2GF.

No algoritmo *StepLR* foi utilizado um tempo de aquecimento de 3 (*warm_t*) e tempo de decaimento de 1 (*decay_t*) (64). Para os experimentos com amostrador 1 e 2 com modelo ViT o total de épocas treinadas foi de 125 e o restante foi 50. No caso dos amostradores identificados como 3 e 4, os modelos foram treinados em 10 épocas com tempo de aquecimento 1 e decaimento 1.

Na camada reguladora de *dropout*, antes da MLP de classificação, foram feitos testes pontuais para verificar melhora de desempenho e diminuição de *overfitting*. Porém, optou-se por não utilizá-lo nessa pesquisa ($P_{dropout} = 0$).

Para classificação multiclasse foi utilizada a função de custo de entropia cruzada. Enquanto para a binária foram encontrados os melhores resultados ao treinar utilizando a entropia cruzada binária com função logit, a qual é expressa por:

$$\text{logit}(p) = \ln \frac{p}{1-p}, \tag{4.2}$$

em que p é a probabilidade resultante dos modelos implementados devido uso da função `softmax` (98).

No banco de dados BUS-BRA, conforme descrito na Seção 3.1.3, existe um desbalanceamento entre as classes que pode adicionar um viés para os modelos com ele treinados. Para verificar esse impacto, foi aplicado um amostrador aleatório ponderado no algoritmo de carregamento dos dados de treinamento. Nesse caso, a ponderação é feita pela quantidade total de cada classe. Os modelos identificados com amostrador 1 e 3 utilizaram o ponderado, enquanto 2 e 4 utilizaram o amostrador aleatório normal.

Os códigos para implementação dos modelos propostos utilizaram PyTorch 2.5.1 no Python 3.12.3. Eles foram executados, em grande parte, em máquina local com Ubuntu 22.04.4, NVIDIA GTX 1080 8GB como placa gráfica, 32 GB de memória RAM e processador Intel i7 13700k. Foram também utilizadas, nas versões finais, três máquinas locais: duas com NVIDIA RTX 3080 10GB e uma com NVIDIA RTX 3070 8GB.

Devido à limitação entre 8 e 10GB de memória dedicada na placa gráfica, todos os modelos foram treinados com precisão mista. As imagens de entrada são `float 16` e, para as classificações binárias, foi usada uma quantização de `8 bits`.

4.4 Experimentos

Os experimentos realizados são divididos em duas abordagens. A primeira aplicada aos modelos baseados em CNNs, a segunda para o *vision transformer*. Os dois experimentos englobam todos os treinamentos realizados para os diferentes tipos de pré-processamento e as classificações.

4.4.1 Experimentos em CNNs

Para as cinco CNN implementadas (VGG16, ResNet50D, EfficientNet-B3, RegNetY-3.2GF, MixNet-XL) foi realizado o TL de ajuste fino completo. Em que são treinados modelos para as seis classificações diferentes conforme sumário na Figura 4.4.

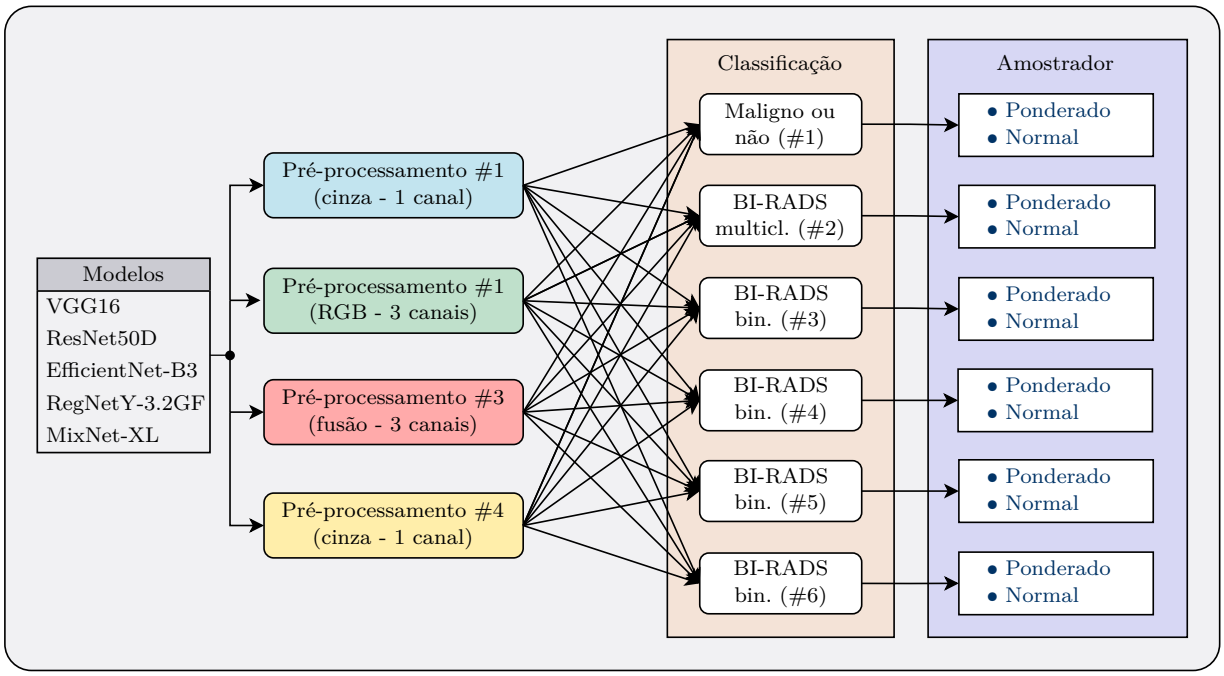


Figura 4.4. Sumário dos experimentos aplicados aos cinco modelos CNNs propostos para as classificações na Tabela 3.1.

4.4.2 Experimentos em *vision transformers*

Para o modelo de *vision transformer* implementado, ViT-base de 16 *patches*, foram testados quatro diferentes tipos de TL: um de ajuste fino completo e três de ajuste fino em que são testados diferentes níveis de congelamento do *encoder* da imagem com o *encoder transformer* congelado. As variações de congelamento do *encoder* são:

1. *Encoder* de imagem e *transformer* congelado, ajuste fino do classificador.
2. Parte do *encoder* de imagem e *transformer* congelado, ajuste fino da camada de separação das partes e projeção linear (conv2D com *stride* e kernel 16×16) e do classificador.
3. *Encoder transformer* congelado, ajuste fino do *encoder* de imagem e do classificador.

Todas as variações foram aplicadas no ViT para mesma quantidade de experimentos vista anteriormente na Figura 4.4. Com exceção dos pré-processamentos que geram

imagem de 1 canal (1 e 4) não ser realizada a variação 1. De fato, é necessário alterar a camada de entrada para suportar 1 canal e, consequentemente, seu ajuste fino.

4.4.3 Comparação entre arquiteturas

Após treinamento dos modelos e variações propostas, cada classificação foi avaliada no subconjunto de teste do BUS-BRA e todas as amostras compatíveis do BrEaST. O OASBUD foi testado para maioria das classificações, exceção da #6 por não existir rótulo BI-RADS 2 para ser isolado. Para o BUSI, apenas os modelos treinados para classificação quanto à malignidade (#1) foram avaliados.

Para definição do melhor modelo de cada arquitetura CNN foi aplicada a seguinte regra: na classificação BI-RADS multiclasse, as métricas de acurácia (macro), acurácia (micro), F1-*Score* (macro), e AUC são utilizadas com igual peso para ranqueamento; enquanto, nas classificações binárias, são utilizadas a acurácia (binária), sensibilidade (binária), F1-*Score* (binário), e AUC.

5 Resultados e Discussão para modelos CNN

Nesse capítulo são apresentados os resultados obtidos para cinco modelos de arquitetura CNN: VGG16, ResNet50D, EfficientNet-B3, RegNetY-3.2GF, e MixNet-XL. Esses modelos foram treinados e testados para seis classificações diferentes com intuito de verificar a eficácia na aplicação de imagens de ultrassom da mama.

5.1 Classificação quanto à Malignidade (#1)

Na classificação quanto à malignidade, os modelos foram treinados para identificar se a lesão em uma imagem é maligna ou não. Os modelos treinados utilizaram apenas amostras do banco de dados BUS-BRA, este define esse grupo de classes utilizando o padrão ouro (biópsia). Ou seja, os modelos implementados não são influenciados por subjetividade e experiência dos profissionais que definiram o tipo de lesão apenas pela imagem.

A Tabela 5.1 apresenta as métricas dos melhores resultados de cada modelo. Em que, a acurácia balanceada (macro), acurácia binária, precisão, sensibilidade, especificidade, F1-*Score* foram as métricas avaliadas.

Tabela 5.1. Melhores resultados da classificação #1 entre os modelos propostos.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#3	1	Custo	0,862	0,894	**0,873**	0,775	**0,948**	0,821	**0,942**
ResNet50D	#3	2	Custo	0,860	0,879	0,809	0,809	0,912	0,809	0,924
EfficientNet-B3	#3	4	Último	**0,894**	0,894	0,817	**0,854**	0,912	0,835	0,925
RegNetY-3.2GF	#3	2	*Acurác.	0,882	**0,904**	0,869	0,820	0,943	**0,844**	0,927
MixNet-XL	#3	2	Último	0,872	0,883	0,798	0,843	0,902	0,820	0,915

*: Acurácia e Custo com mesmo desempenho.

Os melhores modelos nessa classificação utilizaram o pré-processamento #3, originalmente proposto por (62). Quando comparamos a Tabela 5.1 com os de melhor desempenho com outros pré-processamentos, Tabela 5.2, o desempenho diminui para todos os casos. Isso sugere que, para alcançar o melhor desempenho nos modelos testados, é necessário realizar um pré-processamento que depende de uma ROI. Consequentemente, para uma aplicação clínica que busca o melhor desempenho, é essencial uma etapa prévia robusta para a extração da ROI da imagem de entrada. Também é levantada a hipótese

que o uso de alongamento de contraste diminui a diferença entre imagens de equipamentos distintos e torna a classificação mais precisa. No entanto, foi testado apenas em conjunto com a segmentação do tumor por ROI; portanto, é necessário avaliar sua contribuição de forma isolada.

Entre os modelos de arquitetura CNN, a RegNetY-3.2GF foi a melhor no que diz respeito à acurácia, sensibilidade, F1-*Score*, e AUC. Destaca-se que, quando considerada a aplicação para detecção de malignidade com menor quantidade de falso negativo junto com alta precisão, esse é o melhor modelo obtido. Também pode-se destacar o desempenho encontrado para VGG16, que é uma arquitetura mais antiga comparada as outras, e a EfficientNet-B3. Esta apresenta melhor resultado para identificação de caso positivo (sensibilidade) e acurácia média por classe.

Avaliando os modelos para uma aplicação clínica de triagem, a VGG16 se destaca com a melhor especificidade, $94,8\,\%$, seguido da RegNetY-3.2GF, com $94,3\,\%$. Valores mais altos dessa métrica indicam reduzida probabilidade de que pacientes saudáveis sejam erroneamente classificados como suspeitos de câncer. Em triagem, também se deseja sensibilidade alta.

A análise do amostrador revela que não houve uma dominância entre o amostrador ponderado (1 e 3) e o normal (2 e 4) nos modelos testados. Dessa forma, quando aplicado durante o treinamento, é necessário verificar ambos os casos para resultar no melhor modelo.

A Figura 5.1 apresenta as curvas de treinamento para o RegNetY-3.2GF de melhor desempenho, no qual se verifica o distanciamento entre o treinamento e validação nas épocas iniciais. Esse fenômeno pode indicar uma necessidade de maior quantidade de dados de treinamento distintos entre si.

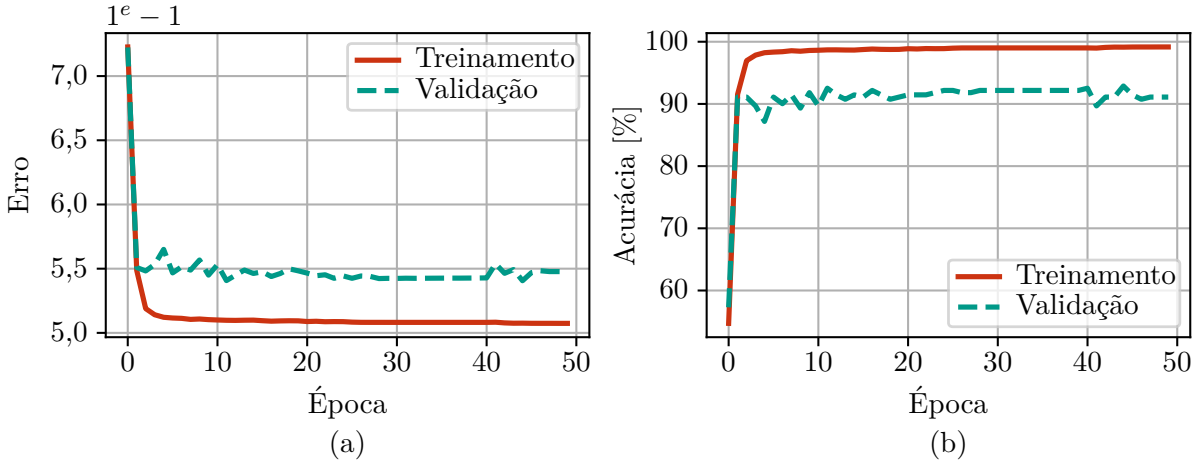


Figura 5.1. Gráficos de função de erro (a) e acurácia (b) por épocas de treinamento para o RegNetY-3.2GF na classificação #1.

Conforme apresentado na Tabela 5.2, os melhores modelos, quando desconsiderado o

pré-processamento #3, utilizaram o pré-processamento proposto (#2). Nesse contexto, o uso de imagens de 3 canais (#2) mostrou-se mais eficaz para todos os casos dessa aplicação, quando comparado ao de 1 canal (#1) e à imagem original (#4).

Tabela 5.2. Melhores resultados da classificação #1 entre os modelos propostos após exclusão do pré-processamento #3.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#2	2	Último	0,778	0,787	0,638	0,753	0,803	0,691	0,865
ResNet50D	#2	4	*Acurác.	0,862	0,862	0,798	0,753	0,912	0,775	0,879
EfficientNet-B3	#2	4	*Acurác.	0,883	0,883	0,841	**0,775**	0,933	**0,807**	0,909
RegNetY-3.2GF	#2	1	Acurác.	0,844	0,887	**0,890**	0,730	**0,959**	0,802	**0,941**
MixNet-XL	#2	3	*Acurác.	**0,887**	0,887	0,880	0,742	0,953	0,805	0,906

*: Acurácia, Custo e Último com mesmo desempenho.

A limitação dos pré-processamentos anteriores é a dependência de uma solução robusta para identificação da ROI para sua aplicação. Nesse sentido, a Tabela 5.3 apresenta os melhores modelos considerando exclusivamente o pré-processamento #4, que consiste no uso da imagem original redimensionada como entrada.

Tabela 5.3. Melhores resultados da classificação #1 entre os modelos propostos para pré-processamento #4 (imagem original).

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#4	1	*Acurác.	0,777	0,840	0,844	0,607	**0,948**	0,706	0,875
ResNet50D	#4	1	Acurác.	0,778	0,809	0,697	0,697	0,860	0,697	0,839
EfficientNet-B3	#4	3	*Acurác.	**0,851**	0,851	0,790	0,719	0,912	0,753	0,892
RegNetY-3.2GF	#4	1	Último	0,838	**0,869**	**0,817**	**0,753**	0,922	**0,784**	**0,913**
MixNet-XL	#4	3	Último	0,844	0,844	0,759	0,742	0,891	0,750	0,890

*: Acurácia e Custo com mesmo desempenho.

Quando comparamos os resultados da Tabela 5.3 com os melhores obtidos (Tabela 5.1), verifica-se uma redução de $\approx 10\,\%$ em várias métricas para parte dos modelos. Contudo, a RegNetY-3.2GF, a EfficientNet-B3, e a MixNet-XL se mantém com desempenho semelhante ao pré-processamento #3. Destaca-se uma perda de $6,0\,\%$ no F1-*Score* e $6,7\,\%$ na sensibilidade Da RegNetY-3.2GF, que se mantém como a melhor CNN, semelhante aos casos anteriores.

Para validação dos resultados obtidos, todos os modelos foram testados para os bancos de dados BrEaST, OASBUD, e BUSI para classificação #1. Todas as amostras compatíveis são testadas sem aplicação de ajuste fino nos modelos com os novos dados. A Tabela 5.4 apresenta os melhores modelos utilizando o mesmo ranqueamento feito nas tabelas anteriores.

Analisando a Tabela 5.4 para o banco de dados BrEaST, observa-se uma perda de desempenho geral de $\approx 10\,\%$ para todos os modelos. O modelo VGG16 apresentou o melhor desempenho, com as métricas mais elevadas, seguido pelo modelo MixNet-XL, que se destacou pela maior sensibilidade. Também ocorre uma mudança no melhor grupo de

Tabela 5.4. Melhores resultados da classificação #1 entre os modelos propostos para os bancos de dados BrEaST, OASBUD e BUSI.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
BrEaST										
VGG16	#3	1	Acurác.	**0,790**	**0,790**	0,683	0,857	0,747	**0,760**	**0,868**
ResNet50D	#3	2	Acurác.	0,786	0,786	**0,686**	0,827	0,760	0,750	0,854
EfficientNet-B3	#3	4	Último	0,746	0,746	0,631	0,837	0,688	0,719	0,848
RegNetY-3.2GF	#3	4	*Acurác.	0,766	0,766	0,661	0,816	0,734	0,731	0,844
MixNet-XL	#3	4	Último	0,762	0,762	0,646	0,857	0,701	0,737	0,856
OASBUD										
VGG16	#3	1	Custo	0,700	0,700	0,855	0,510	0,906	0,639	0,788
ResNet50D	#3	3	**Último	0,665	0,665	0,718	0,587	0,750	0,646	0,743
EfficientNet-B3	#4	3	*Acurác.	0,635	0,635	0,746	0,452	0,833	0,563	0,745
RegNetY-3.2GF	#2	1	Custo	**0,775**	**0,775**	**0,915**	**0,625**	**0,938**	**0,743**	**0,856**
MixNet-XL	#3	2	Último	0,645	0,645	0,714	0,529	0,771	0,608	0,720
BUSI										
VGG16	#3	1	Último	0,927	0,927	0,971	0,800	0,989	0,877	0,970
ResNet50D	#3	4	*Acurác.	0,932	0,932	0,994	0,795	**0,998**	0,884	0,981
EfficientNet-B3	#3	4	*Acurác.	0,920	0,920	0,887	0,862	0,947	0,874	0,965
RegNetY-3.2GF	#3	1	*Acurác.	0,934	0,934	0,915	**0,876**	0,961	0,895	0,974
MixNet-XL	#3	4	Último	**0,941**	**0,941**	**0,943**	0,871	0,975	**0,906**	**0,982**

*: Acurácia e Custo com mesmo desempenho. **: Acurácia, Custo e Último com mesmo desempenho.

hiperparâmetros utilizados para o RegNetY e o MixNet-XL. Anteriomente, utilizava-se uma taxa de amostragem fixa de $1e^{-6}$ com mais épocas de treinamento (Amost. 2), mas passaram a ser adotados os hiperparâmetros com γ variável e taxa de $1e^{-4}$ em menos épocas de treinamento (Amost. 4).

Para o banco de dados OASBUD ocorreu uma maior perda de desempenho para todos os modelos. No entanto, esse resultado não é atípico quando não se realiza um ajuste fino com uma pequena porcentagem dos novos dados (77). Diferente do banco de dados anterior, o RegNetY-3.2GF se mantém como o melhor modelo, mas utiliza o pré-processamento #2 ao invés do #3. Destaca-se a especificidade de $93,8\,\%$.

Comparado à (77), existe uma melhora em todas as métricas, como F1-*Score* $26\,\%$ maior para o RegNetY-3.2GF, considerando que nenhuma amostra do OASBUD é utilizada para ajuste fino do modelo. Contudo, o desempenho foi menor comparado ao estudo de (18), que obteve sensibilidade de $80{,}7\,\% \pm 3,9$ utilizando VGG19 com *transfer learning*.

Mesmo apresentando problemas, conforme estudo de (63), o BUSI é atualmente o banco de dado público mais citado na literatura e, consequentemente, muito utilizado para comparação. Conforme resultado da Tabela 5.4, os modelos propostos apresentam desempenho superior ao compará-los com o subconjunto de teste do BUS-BRA. Este é um resultado esperado devido à existência de múltiplas amostras do mesmo caso (63). O modelo com arquitetura MixNet-XL obteve o melhor desempenho, seguido do RegNetY-3.2GF, cujos resultados foram próximos, com uma diferença de $\approx 1\,\%$. Quando comparado à (77), ocorreu uma melhora de $3\,\%$ na acurácia (macro), acurácia binária, e precisão,

mantendo desempenho semelhante do F1-*Score* ($\approx 91\,\%$).

No estudo de (76), em que foram aplicados diferentes pré-processamentos, se observa desempenho igual na maioria das métricas entre seu melhor resultado e o MixNet-XL. A diferença consiste em uma menor sensibilidade, mas maior precisão, quando se utiliza um pré-processamento semelhante ao #3 deste estudo.

5.2 Classificação BI-RADS Multiclasse (#2)

A classificação #2 engloba as quatro categorias BI-RADS disponíveis no banco de dados de treinamento BUS-BRA: categorias 2, 3, 4, 5. É avaliado o desempenho de arquiteturas CNN para definir em qual categoria uma lesão se encontra, e, consequentemente, possíveis sugestões de controle. Os modelos com melhor desempenho para cada arquitetura estão listados na Tabela 5.5.

Tabela 5.5. Melhores resultados da classificação #2 entre os modelos propostos.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Prec. (macro)	Sensib. (macro)	Especif. (macro)	F1 (macro)	Acurác. (micro)	AUC
VGG16	#3	1	Custo	0,620	0,661	0,620	0,892	0,634	0,699	0,887
ResNet50D	#3	3	Custo	0,846	0,660	0,646	0,891	0,652	0,691	0,887
EfficientNet-B3	#3	4	*Acurác.	0,849	0,726	0,601	0,889	0,626	0,699	0,871
RegNetY-3.2GF	#2	3	**Último	0,849	0,690	**0,670**	0,892	**0,678**	0,699	0,880
MixNet-XL	#3	3	*Acurác.	**0,856**	**0,744**	0,645	**0,894**	0,675	**0,713**	0,887

*: Acurácia e Custo com mesmo desempenho. **: Acurácia, Custo e Último com mesmo desempenho.

Os resultados para as arquiteturas CNN apresentam desempenho mais baixo que a classificação quanto à malignidade, principalmente na sensibilidade, precisão e, consequentemente, F1-*Score* médio por classe. Todavia, observa-se uma predominância do pré-processamento #3 no melhor desempenho, em que o #2 foi melhor apenas no modelo RegNetY-3.2GF.

A especificidade e a acurácia de média por classe (macro) foram as únicas métricas mais elevadas que as demais. Esse comportamento é esperado devido à forma de cálculo. Nela, os verdadeiros negativos (VNs) de cada classe englobam os valores da diagonal principal que não estão relacionados ao valor real de cada caso. Consequentemente, o aparente bom desempenho para aplicação em triagem de paciente é inviesado por uma quantidade bem maior de VNs que VPs. Levando em consideração a acurácia balanceada (macro), a F1-*Score* (macro), a acurácia (micro), e a AUC proporcionalmente, o melhor modelo implementado é o MixNet-XL.

Analisando mais detalhadamente o MixNet-XL, pode-se destacar sua matriz de confusão, conforme Figura 5.2. O primeiro resultado observado é uma maior dificuldade no modelo na distinção entre as categorias 5 e 4 e entre 2 e 3.

A categoria 4 representa suspeita de câncer entre uma probabilidade de 5 % a 95 %.

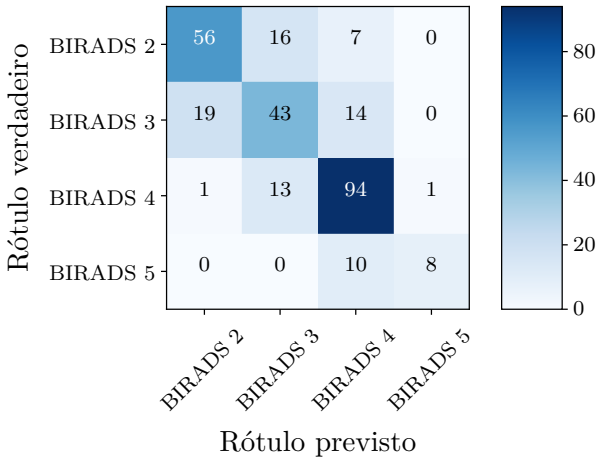


Figura 5.2. Matriz de confusão para modelo MixNet-XL na classificação BI-RADS multiclasse com total de 282 amostras.

A expectativa é que essa categoria contenha tanto casos malignos quanto benignos, pendentes de confirmação por biópsia. Nesse sentido, verifica-se que, para as amostras que a classe verdadeira é 4 mas o modelo identificou como 2 e 3, 71 % eram benignas (10 amostras) e 29 % malignas (4 amostras). O pior tipo de erro, a identificação de uma lesão maligna como benigna pelo modelo, não ocorreu neste conjunto de dados. Porém, o modelo classificou quatro amostras malignas com alta probabilidade de serem benignas. Dessas, três tinham histologia de carcinoma ductal invasivo e uma de linfoma. Esse é um tumor raro, apresentando baixa frequência no banco de dados BUS-BRA (apenas $0,3$ % do total de amostras).

A curva ROC é uma métrica muito utilizada para verificar o desempenho de um modelo na previsão de classes específicas. A Figura 5.3 apresenta as curvas ROC do modelo MixNet-XL para as quatro categorias BI-RADS. Observa-se que o modelo implementado tem melhor desempenho para identificar BI-RADS 4 e 5, ≥ 91 % para ambos os casos.

Os modelos na Tabela 5.6 são os com maior desempenho quando se usa apenas a imagem original como entrada para classificação multiclasse, ou seja, o Pré-proc. #4. Isso indica um desempenho próximo entre as arquiteturas na maioria dos casos. O primeiro modelo, VGG16, se destaca por ser o com melhor desempenho quando descartado aqueles treinados com pré-processamento #3, superando o desempenho da segmentação por ROI proposta nesta pesquisa (#1 e #2).

Para comparar com os resultados do subconjunto de testes do BUS-BRA, também foram realizados testes com os bancos de dados BrEaST e OASBUD, envolvendo a classificação das categorias BI-RADS 2 a 5 e 3 a 5, respectivamente. As métricas multiclasse dos melhores modelos são apresentados na Tabela 5.7.

Comparando os resultados da Tabela 5.7 com os do subconjunto de teste (BUS-BRA), observa-se uma perda expressiva na acurácia (micro), precisão, sensibilidade e,

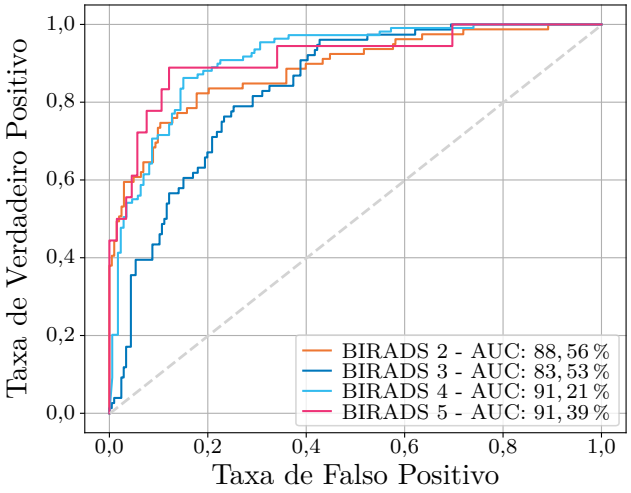


Figura 5.3. Curvas ROC por classe para modelo MixNet-XL na classificação BI-RADS multiclasse com total de 282 amostras.

Tabela 5.6. Melhores resultados da classificação #2 para pré-processamento #4 (imagem original).

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Prec. (macro)	Sensib. (macro)	Especif. (macro)	F1 (macro)	Acurác. (micro)	AUC
VGG16	#4	1	Custo	0,634	**0,657**	**0,634**	**0,879**	**0,642**	**0,663**	0,842
ResNet50D	#4	3	Último	0,793	0,598	0,593	0,851	0,594	0,585	0,797
EfficientNet-B3	#4	3	Último	0,817	0,602	0,573	0,868	0,583	0,635	**0,845**
RegNetY-3.2GF	#4	3	Último	0,807	0,600	0,591	0,861	0,594	0,613	0,807
MixNet-XL	#4	3	*Acurác.	**0,819**	0,620	0,610	0,869	0,614	0,638	0,805

*: Acurácia e Custo com mesmo desempenho.

consequentemente, no F1-*Score* para todos os casos. No BrEaST, a acurácia (macro) e AUC sofreram menos perdas, porém o melhor modelo se altera do MixNet-XL para o EfficientNet-B3 com ResNet50D em sequência. Para as amostras do OASBUD, o RegNetY-3.2GF apresenta desempenho superior na sensibilidade e acurácia (micro), o que o torna com melhor ranking entre os modelos. Apesar de uma diminuição substancial na acurácia (micro), o modelo MixNet-XL se destaca pelo melhor desempenho em outras métricas, como a especificidade, que foi a mais alta tanto para o OASBUD quanto para o BrEaST. Considerando os resultados de especificidade nos três testes, o MixNet-XL é o modelo mais indicado para aplicação de triagem clínica entre aqueles com arquitetura CNN na classificação multiclasse BI-RADS.

Por se tratar de um banco de dados recente, não foram encontrados estudos para classificação BI-RADS que utilizaram a base BrEaST para comparação. No caso do OASBUD, os resultados obtidos pelo RegNetY-3.2GF são semelhantes ao estudo de (99), principalmente ao considerar apenas um pré-processamento. Contudo, quando comparado ao uso de um conjunto de diferentes modelos, proposto por (99), a precisão, sensibilidade, F1-*Score* e AUC são inferiores. Uma hipótese levantada é que a diferença de desempenho ser devido as vantagens do uso de múltiplos modelos com votação suave e

Tabela 5.7. Melhores resultados da classificação #2 entre os modelos propostos para os bancos de dados BrEaST (classes 2–5) e OASBUD (classes 3–5).

BrEaST										
Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Prec. (macro)	Sensib. (macro)	Especif. (macro)	F1 (macro)	Acurác. (micro)	AUC
VGG16	#3	1	Acurác.	0,728	0,427	**0,445**	0,794	0,418	0,456	**0,737**
ResNet50D	#3	2	Acurác.	0,756	**0,462**	0,419	0,798	0,411	0,512	0,714
EfficientNet-B3	#3	4	Último	0,756	0,438	0,436	0,797	**0,435**	0,512	0,699
RegNetY-3.2GF	#3	4	Custo	0,756	0,432	0,432	0,801	0,426	0,512	0,688
MixNet-XL	#2	3	Último	**0,774**	0,455	0,414	**0,802**	0,424	0,548	0,663
OASBUD										
VGG16	#1	1	*Acurác.	0,668	0,462	0,378	0,771	0,393	0,425	0,631
ResNet50D	#3	3	Acurác.	0,685	**0,727**	0,319	0,813	0,355	0,375	0,667
EfficientNet-B3	#4	3	*Acurác.	0,650	0,471	0,383	0,747	0,410	0,415	0,630
RegNetY-3.2GF	#2	3	Último	0,667	0,510	**0,424**	0,721	**0,414**	**0,490**	0,603
MixNet-XL	#3	4	**Custo	**0,688**	0,567	0,365	**0,870**	0,412	0,335	**0,681**

*: Acurácia e Custo com mesmo desempenho. **: Custo e Último com mesmo desempenho.

o número superior de amostras de treinamento de um banco de dados privado (UCC), 6774 imagens (3867 malignas e 2907 benignas).

5.3 Classificação BI-RADS (#3)

A classificação BI-RADS #3 representa a primeira variação implementada do conjunto original de classes, em que os modelos são treinados e testados apenas com amostras do BI-RADS 2 e 5. A princípio, espera-se valores mais altos para as métricas avaliadas, pois se trata de um grupo bem distinto de classes, onde BI-RADS 2 são lesões com probabilidade próxima a 100 % de serem benignas e BI-RADS 5 com $\geq 95\,\%$ de serem malignas.

Devido à remoção de amostras das categorias BI-RADS 3 e 4, os modelos foram testados em um subconjunto de 109 imagens de teste (88 BI-RADS 2 e 21 BI-RADS 5). Posto isso, a Tabela 5.8 apresenta os melhores resultados para as arquiteturas CNNs implementadas.

Tabela 5.8. Melhores resultados da classificação #3 entre os modelos propostos.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#3	*1	**Custo	1,000	1,000	1,000	1,000	1,000	1,000	1,000
ResNet50D	#3	4	Último	1,000	1,000	1,000	1,000	1,000	1,000	1,000
EfficientNet-B3	#3	*1	**Acurác.	1,000	1,000	1,000	1,000	1,000	1,000	1,000
RegNetY-3.2GF	#3	1	***Último	1,000	1,000	1,000	1,000	1,000	1,000	1,000
MixNet-XL	#3	1	Último	0,994	0,991	0,955	1,000	0,989	0,977	0,999

*: Amostrador 1 e 2 com mesmo desempenho. **: Custo, Acurácia e Último com mesmo desempenho. ***: Custo e Acurácia com mesmo desempenho. ****: Último e Custo com mesmo desempenho.

Semelhante as classificações #1 e #2, o pré-processamento #3 obteve as métricas mais elevadas na classificação #3. Os modelos VGG16, ResNet50D, RegNetY-3.2GF e EfficientNet-B3 classificaram corretamente todas as amostras analisadas. Na maioria das

arquiteturas, o amostrador aleatório ponderado (Amost.-1) gerou o modelo com melhor desempenho.

Para o MixNet-XL alcançar os mesmos resultados que as outras arquiteturas, é levantada a hipótese que seja necessário maior refinamento ou uso de outras técnicas de treinamento, como ajuste da taxa de aprendizagem utilizando a função cosseno (83). Entretanto, as curvas de treinamento e validação para o MixNet-XL na Figura 5.4 indicam que as otimizações feitas estão corretas, uma vez que ambas permanecem próximas durante as épocas e tendem ao resultado "ideal".

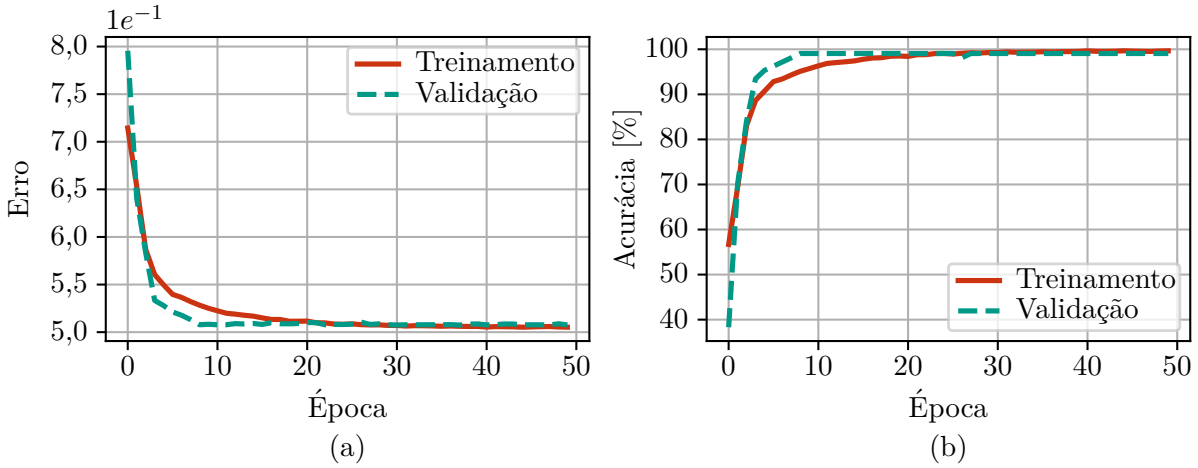


Figura 5.4. Gráficos de função de erro (a) e acurácia (b) por épocas de treinamento para o MixNet-XL na classificação #3.

Avaliando os melhores resultados para os pré-processamentos propostos #1 e #2, conforme primeira parte da Tabela 5.9, é observada uma perda de desempenho para todas as métricas. O MixNet-XL tem uma diminuição mais expressiva que os outros modelos, reforçando a necessidade de refinamento dos hiperparâmetros para uma possível melhora dessa arquitetura.

Tabela 5.9. Melhores resultados da classificação #3 entre os modelos propostos para os pré-processamentos #1 e #2 na primeira parte e #4 na segunda parte.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#1	2	Último	0,930	0,945	0,826	0,905	0,955	0,864	0,982
ResNet50D	#1	1	Último	0,959	0,963	0,870	0,952	0,966	0,909	**0,995**
EfficientNet-B3	#2	3	Custo	0,963	0,963	0,905	0,905	0,977	0,905	0,990
RegNetY-3.2GF	#1	3	*Último	**0,982**	**0,982**	**0,952**	0,952	**0,989**	**0,952**	0,992
MixNet-XL	#2	4	Último	0,954	0,954	0,900	0,857	0,977	0,878	0,972
VGG16	#4	1	Custo	0,894	0,917	0,750	0,857	0,932	0,800	0,946
ResNet50D	#4	4	Custo	0,927	0,927	0,760	0,905	0,932	0,826	0,970
EfficientNet-B3	#4	3	Custo	**0,963**	**0,963**	**0,905**	**0,905**	**0,977**	**0,905**	**0,981**
RegNetY-3.2GF	#4	4	Acurác.	0,945	0,945	0,800	0,952	0,943	0,870	0,977
MixNet-XL	#4	3	Último	0,927	0,927	0,783	0,857	0,943	0,818	0,946

*: Último e Custo com mesmo desempenho.

Quando são avaliadas todas as métricas para os pré-processamentos #1 e #2, Tabela 5.9, pode-se concluir que o RegNetY-3.2GF foi o melhor modelo para a classificação

#3. Este é o mesmo modelo que apresentou o melhor desempenho para classificação da lesão como maligna ou não (#1), estando estas diretamente relacionadas com BI-RADS 5 e 2, respectivamente. Quando levado em consideração apenas o pré-processamento #4, o melhor modelo foi o EfficientNet-B3. Neste modelo, o uso da imagem original resulta em alta sensibilidade para predição de malignidade (90 %) e alta precisão (90 %) comparado aos outros modelos.

Ao realizar testes nos bancos de dados BrEaST e OASBUD para todos os modelos foram obtidos os resultados da Tabela 5.10. De forma similar aos resultados das amostras de teste, o pré-processamento #3 gerou os modelos com maiores desempenhos em ambos os dados. Também se observam valores próximos na maioria das métricas para todas as arquiteturas testadas no BrEaST. Os modelos RegNetY-3.2GF e ResNet50D se destacam com F1-*Score* de 91, 7 % e 91, 5 %, respectivamente.

Tabela 5.10. Melhores resultados da classificação #3 entre os modelos propostos para os bancos de dados BrEaST e OASBUD.

BrEaST										
Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#3	1	Acurác.	0,868	0,868	0,860	0,935	0,767	0,896	0,927
ResNet50D	#3	2	*Acurác.	0,895	0,895	**0,896**	0,935	**0,833**	0,915	**0,964**
EfficientNet-B3	#3	4	**Custo	0,868	0,868	0,875	0,913	0,800	0,894	0,936
RegNetY-3.2GF	#3	4	Acurác.	0,895	0,895	0,880	**0,957**	0,800	**0,917**	0,946
MixNet-XL	#3	3	Último	0,868	0,868	0,875	0,913	0,800	0,894	0,909
OASBUD										
VGG16	#3	1	Acurác.	0,848	0,848	0,923	0,720	0,952	0,809	0,881
ResNet50D	#3	3	Custo	0,821	0,821	0,826	**0,760**	0,871	0,792	0,859
EfficientNet-B3	#3	4	**Custo	0,813	0,813	**1,000**	0,580	**1,000**	0,734	0,880
RegNetY-3.2GF	#3	1	Custo	**0,857**	**0,857**	0,972	0,700	0,984	**0,814**	**0,895**
MixNet-XL	#3	2	Custo	0,830	0,830	0,897	0,700	0,935	0,787	0,850

*: Acurác., Custo e Último com mesmo desempenho. **: Custo e Último com mesmo desempenho.

Ao contrário do observado no BrEaST, os modelos testados no OASBUD obtiveram um desempenho inferior, principalmente em termos da sensibilidade. O RegNetY-3.2GF se mantém como o melhor modelo CNN com AUC, F1, e acurácia binária de 89, 5 %, 81, 4 %, e 85, 7 %, respectivamente.

5.4 Classificação BI-RADS (#4)

A classificação BI-RADS #4 é a segunda variante binária proposta, onde são agrupadas as categorias BI-RADS 2 e 3 na classe A e BI-RADS 4 e 5 na classe B. O objetivo é verificar o desempenho para classificação em duas categorias imaginárias mais simples, um com viés mais benigno (A) e outro mais maligno (B).

Na aplicação clínica de triagem, um paciente classificado na classe A indicaria ausência de câncer, enquanto, na classe B, o paciente seria encaminhado para uma avaliação mais

aprofundada, como o exame de confirmação por biópsia. A Tabela 5.11 apresenta o melhor desempenho para as arquiteturas CNN.

Tabela 5.11. Melhores resultados da classificação #4 entre os modelos propostos.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#3	2	Acurác.	0,894	0,897	**0,902**	0,866	**0,923**	0,884	0,939
ResNet50D	#3	3	Custo	0,887	0,887	0,874	0,874	0,897	0,874	0,946
EfficientNet-B3	#3	1	Acurác.	0,867	0,865	0,830	0,882	0,852	0,855	0,926
RegNetY-3.2GF	#3	3	Último	**0,908**	**0,908**	0,886	**0,913**	0,903	**0,899**	**0,949**
MixNet-XL	#3	4	Último	0,897	0,897	0,866	0,913	0,884	0,889	0,948

Quando comparamos os melhores resultados da classificação #4 (Tabela 5.11) com a multiclasse BI-RADS (Tabela 5.5), observamos um aumento superior a 25 % na acurácia balanceada e de 10 % para o AUC. Quando avaliadas as métricas binárias, verifica-se que o RegNetY-3.2GF foi o modelo com melhor desempenho, em que sua capacidade de identificar BI-RADS 4 e 5 (B) com precisão alcança $\approx 90\,\%$ (F1-*Score*). Contudo, é necessária uma implementação robusta para definição do contorno (ROI) na aplicação clínica, pois esse desempenho é apenas possível com o pré-processamento #3. Para aplicação em triagem, o VGG16 apresenta o maior desempenho quanto à especificidade $(92,3\,\%)$, seguido do RegNetY-3.2GF $(90,3\,\%)$.

Considerando uma implementação que não é composta de uma boa solução para definição da ROI, os modelos CNN para classificação #4 alcançam as métricas apresentadas na Tabela 5.12. O RegNetY-3.2GF foi o único modelo que não sofreu uma perda mais clara de desempenho, sendo a melhor escolha para esse caso. Também foi possível observar que em todas as arquiteturas o melhor modelo final foi aquele com maior quantidade de épocas de treinamento (último). Isso sugere uma possível necessidade de mais épocas de treinamento para o pré-processamento #4.

Tabela 5.12. Melhores resultados da classificação #4 entre os modelos propostos para pré-processamento #4 (imagem original).

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#4	1	Último	0,791	0,798	0,813	0,717	0,865	0,762	0,827
ResNet50D	#4	1	Último	0,769	0,773	0,756	0,732	0,806	0,744	0,840
EfficientNet-B3	#4	3	Último	0,784	0,784	0,766	0,748	0,813	0,757	0,851
RegNetY-3.2GF	#4	2	Custo	**0,829**	**0,830**	**0,806**	**0,819**	**0,839**	**0,813**	**0,881**
MixNet-XL	#4	4	*Último	0,787	0,787	0,741	0,811	0,768	0,774	0,832

*: Último, Acurácia e Custo com mesmo desempenho.

Ao avaliar os modelos para classificação #4 com um banco de dados não relacionado com o de treinamento, como o BrEaST e OASBUD, foram obtidos os resultados da Tabela 5.13. Quando comparado aos resultados de teste, a principal diferença observada é uma diminuição de especificidade para todos os modelos. Isso sugere boa generalização para identificação de amostras nas categorias BI-RADS 4 e 5, mas com perdas na identificação de categorias benignas (2 e 3). Ou seja, na aplicação clínica, ocorreria uma

situação de sobrediagnóstico onde pacientes seriam submetidos a biópsia de forma desnecessária devido à lesão ter sido identificada como potencialmente maligna quando era benigna.

Tabela 5.13. Melhores resultados da classificação #4 entre os modelos propostos para os bancos de dados BrEaST e OASBUD.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
BrEaST										
VGG16	#3	2	Último	0,734	0,734	0,839	0,789	0,582	0,813	0,769
ResNet50D	#3	2	Custo	0,738	0,738	**0,870**	0,757	**0,687**	0,809	0,790
EfficientNet-B3	#3	2	Acurác.	0,746	0,746	0,807	**0,859**	0,433	0,832	0,798
RegNetY-3.2GF	#3	4	*Acurác.	0,750	0,750	0,851	0,800	0,612	0,825	**0,820**
MixNet-XL	#3	1	Acurác.	**0,766**	**0,766**	0,842	0,838	0,567	**0,840**	0,784
OASBUD										
VGG16	#1	1	Acurác.	0,750	0,750	0,782	**0,884**	0,452	0,830	0,712
ResNet50D	#2	1	Último	0,725	0,725	0,817	0,775	0,613	0,796	0,777
EfficientNet-B3	#3	4	Último	0,760	0,760	**0,888**	0,746	**0,790**	0,811	**0,829**
RegNetY-3.2GF	#3	4	Acurác.	**0,790**	**0,790**	0,853	0,841	0,677	**0,847**	0,823
MixNet-XL	#2	3	Custo	0,720	0,720	0,770	0,848	0,435	0,807	0,753

*: Acurác. e Último com mesmo desempenho.

Para o teste aplicado no BrEaST, Tabela 5.13, o melhor modelo é o EfficientNet-B3, considerando a acurácia binária, sensibilidade, F1-*Score* e AUC para ranqueamento. Porém, quando incluída a especificidade no ranqueamento, os modelos RegNetY-3.2GF e MixNet-XL empatam na primeira posição, com F1-*Score* de $82,5\,\%$ e $84,0\,\%$ e AUC de $82,0\,\%$ e $78,4\,\%$, respectivamente. Considerando os resultados obtidos no OASBUD, o modelo com melhor desempenho geral entre as CNNs é o RegNetY-3.2GF, alcançando F1-*Score* e AUC de $84,7\,\%$ e $82,3\,\%$, respectivamente.

5.5 Classificação BI-RADS (#5)

Na classificação BI-RADS #5, amostras classificadas como BI-RADS 5 (categorias que indica alta probabilidade de malignidade da lesão) são isoladas na classe B, enquanto o restante é definido como A. O objetivo dessa classificação é verificar a capacidade de distinção que os modelos tem de isolar a classe 5 das demais. Visto que foi possível observar na classificação BI-RADS #2 (Figura 5.2) uma dificuldade em separar o BI-RADS 5 do 4. Entretanto, os modelos treinados para essa classificação não podem ser aplicados para triagem de pacientes, considerando que a classe negativa (A) inclui a categoria BI-RADS 4 (suspeita de câncer).

Na Tabela 5.14, estão apresentados os melhores modelos para as métricas analisadas. O principal resultado encontrado foi uma variação do pré-processamento entre os modelos, pois, em todas as classificações anteriores, o #3 se sobressaiu comparado aos demais. Um destaque foi o melhor modelo MixNet-XL utilizar o pré-processamento #4, na qual é usado apenas as imagens originais como entrada.

Tabela 5.14. Melhores resultados da classificação #5 entre os modelos propostos.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#1	1	Último	0,742	0,954	0,692	0,500	0,985	0,581	0,872
ResNet50D	#3	3	Custo	0,947	0,947	0,615	0,444	0,981	0,516	0,947
EfficientNet-B3	#1	3	Acurác.	**0,957**	0,957	0,714	0,556	0,985	0,625	0,943
RegNetY-3.2GF	#2	1	Último	0,772	**0,961**	**0,769**	0,556	**0,989**	**0,645**	**0,953**
MixNet-XL	#4	3	Custo	0,947	0,947	0,600	0,500	0,977	0,545	0,884

Analisando a Tabela 5.14 para a identificação do BI-RADS 5 (B), constata-se uma dificuldade generalizada para predição com baixa quantidade de falso negativo e alta precisão (F1). Alguns modelos resultam em alta sensibilidade mas baixa precisão, enquanto outros tem alta precisão mas baixa sensibilidade. Como o F1-*Score* leva em consideração ambos, o modelo com melhor desempenho para classificação do BI-RADS 5 é o RegNetY-3.2GF.

Quando se considera apenas o pré-processamento #4, o melhor modelo se mantém, conforme resultado da Tabela 5.15. Nesse caso, a RegNetY-3.2GF não é o mesmo treinado da avaliação, uma vez que foi necessário o refinamento dos hiperparâmetros de treinamento para alcançar esses resultados. A sensibilidade foi igual ao pré-processamento #2, mas ocorreu a redução de 18 % da precisão, refletindo na diminuição de ≈ 7 % do F1-*Score*.

Tabela 5.15. Melhores resultados da classificação #5 entre os modelos propostos para pré-processamento #4 (imagem original).

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#4	1	*Custo	0,715	0,950	0,667	0,444	0,985	0,533	**0,876**
ResNet50D	#4	3	**Acurác.	0,926	0,926	0,440	**0,611**	0,947	0,512	0,855
EfficientNet-B3	#4	3	**Acurác.	0,936	0,936	0,500	0,389	0,973	0,438	0,864
RegNetY-3.2GF	#4	3	**Acurác.	0,947	0,947	0,588	0,556	0,973	**0,571**	0,871
MixNet-XL	#4	3	Custo	0,947	0,947	**0,600**	0,500	0,977	0,545	0,884

*: Último e Custo com mesmo desempenho. **: Acurácia e Custo com mesmo desempenho.

Para verificar o desempenho para amostras de outros bancos de dados, todos os modelos foram testados no BrEaST e OASBUD, com a classe A contendo, respectivamente, as categorias 2, 3, 4 e 3, 4. Os melhores modelos, segundo ranqueamento aplicado anteriormente e excluindo aqueles com AUC abaixo de 70 %, caso aplicável, estão listados na Tabela 5.16. Em ambos os casos, observa-se o mesmo fenômeno de variabilidade da precisão e sensibilidade vista anteriormente, em que alguns modelos têm com precisão maior e sensibilidade menor enquanto outros o contrário. Consequentemente, o F1-*Score* é mais baixo, principalmente para algumas arquiteturas aplicadas no OASBUD.

Ao comparar os resultados de ambos os conjuntos de dados (Tabela 5.16), pode-se destacar o ResNet50D com melhor desempenho. Na qual, o pré-processamento com a imagem original (#4) obteve a maior performance no BrEaST e a imagem segmentada

Tabela 5.16. Melhores resultados da classificação #5 entre os modelos propostos para os bancos de dados BrEaST e OASBUD.

BrEaST										
Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#1	1	*Custo	0,688	0,821	0,512	0,478	0,898	0,494	0,764
ResNet50D	#4	1	Acurác.	0,727	**0,829**	**0,531**	0,565	0,888	**0,547**	**0,807**
EfficientNet-B3	#3	4	Último	0,806	0,806	0,463	0,413	0,893	0,437	0,790
RegNetY-3.2GF	#1	3	Último	0,770	0,770	0,417	**0,652**	0,796	0,508	0,776
MixNet-XL	#1	4	Último	**0,821**	0,821	0,514	0,413	**0,913**	0,458	0,749
OASBUD										
VGG16	#2	2	**Custo	0,695	0,695	0,430	**0,680**	0,700	0,527	0,735
ResNet50D	#3	2	Acurác.	0,775	0,775	0,593	0,320	0,927	0,416	**0,739**
EfficientNet-B3	#3	4	Último	**0,785**	**0,785**	**0,733**	0,220	0,973	0,338	0,714
RegNetY-3.2GF	#3	1	Último	0,760	0,760	0,600	0,120	0,973	0,200	0,699
MixNet-XL	#3	4	Último	0,750	0,750	0,500	0,360	0,880	0,419	0,740

*: Acurác. e Custo com mesmo desempenho. **: Custo e Último com mesmo desempenho.

por fusão (#3) para o OASBUD. O uso do pré-processamento #3 no BrEaST resultou em especificidade excepcional de $89,1\,\%$ e AUC de $82,9\,\%$, porém o F1-*Score* foi inferior ($50\,\%$).

5.6 Classificação BI-RADS (#6)

A classificação BI-RADS #6 foi o último experimento realizado, derivado da mesma proposição do #5, isolando a categoria BI-RADS 2 (A) das restantes (B). Devido à grande semelhança da categoria 2 e 3 — sendo a primeira com $\approx 100\,\%$ de probabilidade e a segunda com $\geq 95\,\%$ de uma lesão ser benigna — há uma dificuldade para os modelos implementados em distinguí-las. Esse problema foi exemplificado na Figura 5.2, pois um número expressivo de amostras BI-RADS 3 ou 4 foi identificado como 2 e o mesmo ocorreu para amostras classificadas como 2 quando eram, na verdade, 3 ou 4, com taxas de $\approx 28\,\%$ e $\approx 30\,\%$ respectivamente.

Semelhante ao observado na #5, esta classificação também não resulta em modelos clinicamente aplicáveis para a triagem de pacientes. Isso ocorre porque a categoria BI-RADS 3 (lesões provavelmente benignas) é agrupada com as categorias BI-RADS 4 (lesões suspeitas) e 5 (lesões com alta probabilidade de malignidade).

Na Tabela 5.17 são apresentados os melhores modelos treinados para cada arquitetura, conforme avaliação no subconjunto de testes. Observa-se um fenômeno semelhante a da classificação #5 quanto ao tipo de pré-processamento usado A diferença reside no fato de que parte dos modelos utilizou o pré-processamento #3, e todos empregaram imagens de 3 canais como entrada.

Também se observa uma dificuldade em distinção da classe A, que contém BI-RADS 2, com a classe B, uma vez que a especificidade tem valores baixos comparados a sensi-

Tabela 5.17. Melhores resultados da classificação #6 entre os modelos propostos.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#3	2	Acurác.	0,730	0,823	0,834	0,941	0,519	0,884	0,893
ResNet50D	#3	4	Acurác.	0,844	0,844	0,860	0,936	0,608	0,896	0,871
EfficientNet-B3	#2	4	Acurác.	0,848	0,848	0,864	0,936	0,620	0,898	0,846
RegNetY-3.2GF	#3	1	Acurác.	0,812	**0,869**	**0,884**	0,941	**0,684**	**0,912**	**0,901**
MixNet-XL	#2	4	Acurác.	**0,851**	0,851	0,864	0,941	0,620	0,901	0,859

bilidade. Esse foi o mesmo fenômeno presente na classificação #5, porém para a métrica oposta. A precisão e sensibilidade levam em consideração acertos da classe positiva (B) e, consequentemente, têm valor elevado na classificação #6. Por essa razão, elas não apresentam os mesmos baixos valores quando a classe B era definida unicamente pela categoria 5. Essa diferença é notável devido o aumento global de 40 % ou mais para várias métricas nesta classificação, na qual a classe B engloba as categorias BI-RADS 3, 4 e 5.

Considerando a mesma abordagem de ranqueamento usada para as classificações anteriores, onde a identificação de câncer é mais importante, o melhor modelo foi o RegNetY-3.2GF, com MixNet-XL em segundo e o ResNet50D em terceiro para classificação #6.

Contudo, uma avaliação mais apropriada para a classificação #6 requer uma abordagem distinta das classificações anteriores. Considerando a aplicação clínica, a classe positiva geralmente se associa à presença de câncer, de modo que a classe B é considerada a positiva nesse caso específico. Portanto, a especificidade é a métrica mais adequada para ranqueamento dos modelos da classificação #6, em conjunto com acurácia balanceada, F1-*Score* e AUC. Aplicando esse novo tipo de avaliação são obtidos os modelos apresentados na Tabela 5.18.

Tabela 5.18. Melhores resultados da classificação #6 entre os modelos propostos.

Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	#1	1	Último	0,778	0,837	0,869	**0,911**	0,646	0,889	0,848
ResNet50D	#2	1	Acurác.	0,789	0,840	0,876	0,906	0,671	0,891	0,858
EfficientNet-B3	#2	3	Acurác.	0,826	0,826	0,877	0,882	0,684	0,880	0,878
RegNetY-3.2GF	#2	1	Último	**0,846**	**0,872**	**0,915**	0,906	**0,785**	**0,911**	0,877
MixNet-XL	#3	1	Acurác.	0,789	0,830	0,882	0,882	0,696	0,882	**0,882**

Semelhante ao que ocorreu na Tabela 5.17, o ranqueamento com especificidade sugere que o pré-processamento #3 não é o mais eficiente quando aplicado na identificação do BI-RADS 2 em relação ao restante. Mesmo com um novo ranqueamento, o melhor modelo se manteve o mesmo, o RegNetY-3.2GF. Ele alcançou especificidade superior comparado aos outros, com média de $\approx 11,1\,\%$.

Ao realizar os mesmos testes, porém aplicados aos dados do BrEaST, foram obtidos os modelos apresentados na Tabela 5.19. No geral, a dificuldade em discernir a categoria BI-RADS 2 das demais agrava-se quando os modelos são aplicados a um novo conjunto

de dados. Nesse cenário, observa-se uma melhora na acurácia, precisão e sensibilidade. Contudo, o desempenho da especificidade é consideravelmente inferior ao registrado anteriormente, mesmo ao se utilizar essa métrica no ranqueamento. Uma suposição é que as categorias BI-RADS 2 e 3 têm características muito semelhantes entre si, ocasionando em uma maior dificuldade em distinção entre elas. Nesse caso, outras métricas, tal como a sensibilidade, apresentaram valores mais elevados, visto que a classe B corresponde a $88,1\,\%$ do total de amostras testadas. Enquanto o subconjunto de teste do BUS-BRA são mais balanceadas, com a classe B contendo $45\,\%$ do total de amostras.

Tabela 5.19. Melhores resultados da classificação #6 entre os modelos propostos para o banco de dados BrEaST.

BrEaST											
Aval.	Modelo	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
Sensib.	VGG16	#3	2	Acurác.	0,881	0,881	0,893	0,982	0,133	0,936	**0,795**
	ResNet50D	#3	2	Acurác.	0,869	0,869	**0,906**	0,950	**0,267**	0,927	0,774
	EfficientNet-B3	#3	4	*Acurác.	0,853	0,853	0,897	0,941	0,200	0,919	0,735
	RegNetY-3.2GF	#2	3	Último	0,881	0,881	0,903	0,968	0,233	0,935	0,711
	MixNet-XL	#1	3	Último	**0,893**	**0,893**	0,895	**0,996**	0,133	**0,942**	0,603
Especif.	VGG16	#3	2	Acurác.	0,881	0,881	0,893	0,982	0,133	0,936	**0,795**
	ResNet50D	#3	2	Acurác.	0,869	0,869	0,906	0,950	0,267	0,927	0,774
	EfficientNet-B3	#3	4	*Acurác.	0,853	0,853	0,897	0,941	0,200	0,919	0,735
	RegNetY-3.2GF	#3	2	*Acurác.	0,857	0,857	**0,919**	0,919	**0,400**	0,919	0,786
	MixNet-XL	#1	3	*Acurác.	**0,889**	**0,889**	0,894	**0,991**	0,133	**0,940**	0,611

*: Acurác. e Custo com mesmo desempenho.

Considerando ambas as avaliações na Tabela 5.19, o melhor modelo CNN para a classificação #6 foi o VGG16. Destaca-se o desempenho do RegNetY-3.2GF quando considerada a especificidade no ranqueamento, obtendo F1-*Score* de $91,9\,\%$ e a maior especificidade $(40,0\,\%)$.

6 Resultados e Discussões para o modelo ViT

Nesse capítulo são apresentados e discutidos os resultados encontrados para variações de *transfer learning* (TL) aplicadas no treinamento da arquitetura ViT-Base com 16 *patches*. O objetivo é avaliar o desempenho encontrado quando se realiza o ajuste fino completo (1), ajuste fino apenas do classificador (2), ajuste fino do classificador e parte do *encoder* de imagem (3), e ajuste fino completo do classificador e do *encoder* de imagem (4). Todos os casos onde é realizado um ajuste fino parcial (2 a 4), o *encoder transformer* é preservado para não alterar o conhecimento aprendido da aplicação original, neste caso ImageNet1K com pré-treino no ImageNet12K e no LAION-2B.

Para cada classificação proposta, também se compara o melhor modelo ViT-Base com os cinco de arquitetura CNN, apresentados e discutidos no Capítulo 5.

6.1 Classificação quanto à Malignidade (#1)

A classificação quanto à malignidade tem como objetivo que os modelos identifiquem se a lesão na imagem de ultrassom é benigna ou maligna, considerando que ela contém, pelo menos, uma lesão. Como o modelo ViT treinado, assim como as CNNs, utilizou uma base de dados cuja classe foi validada por biópsia, considera-se que as métricas resultantes não contêm viés de erro intrínseco à experiência dos profissionais que realizam a avaliação do exame.

Primeiramente, são apresentados os modelos ViT com melhor desempenho para cada variação de TL na Tabela 6.1. Nesta tabela, os modelos são apresentados em ordem decrescente de performance, conforme o ranqueamento proporcional das métricas de acurácia (binária), sensibilidade, F1, e AUC.

Tabela 6.1. Melhores resultados da classificação #1 entre as variações de TL no ViT-Base.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
ViT-Base	1	#3	2	Acurác.	**0,897**	**0,897**	**0,849**	0,820	**0,933**	**0,834**	**0,930**
ViT-Base	4	#3	2	Acurác.	0,887	0,887	0,813	**0,831**	0,912	0,822	0,924
ViT-Base	3	#3	2	Acurác.	0,865	0,865	0,807	0,753	0,917	0,779	0,919
ViT-Base	2	#3	1	Acurác.	0,851	0,851	0,747	0,798	0,876	0,772	0,903

Analisando a Tabela 6.1, é identificado que o pré-processamento dos melhores resultados foi o #3, originalmente proposto por (62). O mesmo fenômeno ocorreu para todas as arquiteturas CNN, conforme observado na Tabela 6.2. Em relação ao tipo de TL, o ajuste fino completo (1) foi o utilizado para o maior desempenho na maioria das métricas, resultando na acurácia balanceada e micro de 89,7 % e sensibilidade de 82 %. Também destaca-se que o modelo treinado com ajuste fino do classificador e do *encoder* de imagem (4) apresentout sensibilidade 1 % melhor que o com melhor performance geral. Contudo, sua precisão é 3 % inferior e, consequentemente, o F1-*Score* é menor, levando a um desempenho abaixo do ajuste fino completo.

Tabela 6.2. Melhores resultados da classificação #1 entre todos os modelos propostos.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#3	1	Custo	0,862	0,894	**0,873**	0,775	**0,948**	0,821	**0,942**
ResNet50D	1	#3	2	Custo	0,860	0,879	0,809	0,809	0,912	0,809	0,924
EfficientNet-B3	1	#3	4	Último	0,894	0,894	0,817	**0,854**	0,912	0,835	0,925
RegNetY-3.2GF	1	#3	2	*Acurác.	0,882	**0,904**	0,869	0,820	0,943	**0,844**	0,927
MixNet-XL	1	#3	2	Último	0,872	0,883	0,798	0,843	0,902	0,820	0,915
ViT-Base	1	#3	2	Acurác.	**0,897**	0,897	0,849	0,820	0,933	0,834	0,930

*: acurácia e perda com mesmo desempenho

Conforme as métricas apresentadas na Tabela 6.2, o ViT-Base obteve desempenho comparável com os melhores de arquitetura CNN, o RegNetY-3.2GF e o EfficientNet-B3. No qual se ressalta a sensibilidade como a segunda maior entre os seis modelos analisados e o F1-*Score*, como o terceiro. Também observa-se um desempenho comparável aos melhores modelos CNNs quanto à especificidade, esta uma métrica utilizada para medir eficiência do modelo na triagem de pacientes.

Utilizando o AUC como métrica única de ranqueamento, as posições se alteram e o VGG16 se torna o melhor modelo com 94,2 %. Nesse contexto, o ViT-Base é o segundo melhor para classificação de maligno ou não, com ROC médio para comparação dos seis modelos sendo apresentado na Figura 6.1.

A avaliação apenas pelas amostras de teste não indica que um modelo apresentará, necessariamente, o mesmo desempenho para uma nova entrada. Assim, foi realizada a mesma avaliação entre todas as arquiteturas utilizando os bancos de dados BrEaST, OASBUD, e BUSI, conforme Tabela 6.3. Como o intuito é comparar a arquitetura ViT com o restante, não serão repetidas as mesmas avaliações feitas com outros estudos que utilizaram CNN, detalhado na Seção 5.1 do capítulo anterior.

Os resultados para o BrEaST na Tabela 6.3, indicam que o ViT-Base tem desempenho semelhante as outras arquiteturas. Ou seja, comparado ao desempenho no subconjunto de teste (Tabela 6.2), ocorreu uma redução de desempenho para todas as arquiteturas, se destacando uma diminuição de precisão de $\approx 20\,\%$. Outro detalhe foi que o ViT treinado

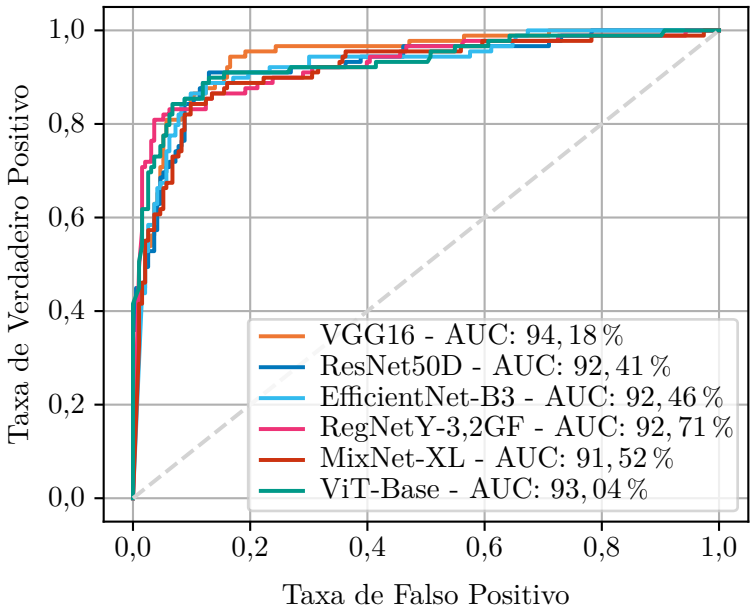


Figura 6.1. Curvas ROC dos melhores modelos para classificação quanto à malignidade (#1).

com amostrador ponderado (Amost.-1) apresentou melhor desempenho comparado ao normal, o oposto do observado nas amostras de teste.

No OASBUD, a redução de desempenho observada na maioria das arquiteturas CNN foi menor no ViT, semelhante ao que ocorre no RegNetY-3.2GF. Como as métricas analisadas são próximas entre esses dois modelos, conclui-se que o ViT apresenta melhora ou se aproxima de resultados do estado da arte (77, 18). Também se ressalta que o pré-processamento com maior desempenho no OASBUD foi o proposto nesse estudo (#2), superando a imagem segmentada por fusão (#3) (melhor geral).

Para as amostras do banco de dados BUSI, espera-se um desempenho superior. De fato, o ViT apresentou os maiores valores de AUC e precisão neste conjunto, atingindo $98,5\,\%$ e $97,8\,\%$, respectivamente. Um detalhe é que o ajuste fino apenas do classificador (TL-2) gerou melhores resultados comparado ao ajuste fino completo do modelo. Este que foi o pior resultado no subconjunto de testes, conforme apresentado na Tabela 6.1. Comparando com estudo de (54), todos os modelos alcançam resultados igual ou melhores na maioria das métricas, com exceção da sensibilidade no VGG16, ResNet50D e ViT. O melhor modelo geral para classificação #1, MixNet-XL, obteve métricas superiores ao HoVer-Trans de (54), com aumentos de $3,4\,\%$ no F1-*Score* e $11,7\,\%$ na AUC.

Considerando os três resultados da Tabela 6.3, o ViT apresenta desempenho próximo aos melhores modelos CNN ao considerar a aplicação para triagem de pacientes. Contudo, ela não é a melhor e, consequentemente, o RegNetY-3.2GF é a arquitetura com melhor desempenho geral para essa aplicação clínica.

Tabela 6.3. Melhores resultados da classificação #1 entre todos os modelos propostos para os bancos de dados BrEaST, OASBUD e BUSI.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
BrEaST											
VGG16	1	#3	1	Acurác.	**0,790**	**0,790**	0,683	0,857	0,747	**0,760**	**0,868**
ResNet50D	1	#3	2	Acurác.	0,786	0,786	**0,686**	0,827	**0,760**	0,750	0,854
EfficientNet-B3	1	#3	4	Último	0,746	0,746	0,631	0,837	0,688	0,719	0,848
RegNetY-3.2GF	1	#3	4	*Acurác.	0,766	0,766	0,661	0,816	0,734	0,731	0,844
MixNet-XL	1	#3	4	Último	0,762	0,762	0,646	0,857	0,701	0,737	0,856
ViT-Base	1	#3	1	Custo	0,758	0,758	0,648	0,827	0,714	0,726	0,828
OASBUD											
VGG16	1	#3	1	Custo	0,700	0,700	0,855	0,510	0,906	0,639	0,788
ResNet50D	1	#3	3	**Último	0,665	0,665	0,718	0,587	0,750	0,646	0,743
EfficientNet-B3	1	#4	3	*Acurác.	0,635	0,635	0,746	0,452	0,833	0,563	0,745
RegNetY-3.2GF	1	#2	1	Custo	**0,775**	**0,775**	**0,915**	0,625	**0,938**	0,743	**0,856**
MixNet-XL	1	#3	2	Último	0,645	0,645	0,714	0,529	0,771	0,608	0,720
ViT-Base	1	#2	1	*Acurác.	0,770	0,770	0,872	**0,654**	0,896	**0,747**	0,848
BUSI											
VGG16	1	#3	1	Último	0,927	0,927	0,971	0,800	0,989	0,877	0,970
ResNet50D	1	#3	4	*Acurác.	0,932	0,932	0,994	0,795	**0,998**	0,884	0,981
EfficientNet-B3	1	#3	4	*Acurác.	0,920	0,920	0,887	0,862	0,947	0,874	0,965
RegNetY-3.2GF	1	#3	1	*Acurác.	0,934	0,934	0,915	**0,876**	0,961	0,895	0,974
MixNet-XL	1	#3	4	Último	**0,941**	**0,941**	0,943	0,871	0,975	**0,906**	0,982
ViT-Base	2	#3	1	***Custo	0,938	0,938	**0,978**	0,829	0,991	0,897	**0,985**

*: Acurácia e Custo com mesmo desempenho. **: Acurácia, Custo e Último com mesmo desempenho. ***: Custo e Último com mesmo desempenho.

6.2 Classificação BI-RADS Multiclasse (#2)

Na classificação #2, o modelo é treinado para identificar uma imagem com lesão entre quatro categorias BI-RADS: 2, 3, 4 e 5. Mesmo o BI-RADS sendo um sistema de padronização para classificação de imagens mamárias, as categorias definidas nele são sujeitas a subjetividade e experiência do profissional responsável. Portanto, esse modelos contêm um viés de erro associado com a experiência desse profissional, diferente da classificação #1 que é definida pelo diagnóstico histológico por biópsia. De qualquer forma, um sistema automatizado de avaliação de imagens mamárias precisaria incluir a classificação BI-RADS para uma melhor categorização das lesões identificadas.

Conforme visto anteriormente, é esperado que as métricas sejam mais baixas quando comparadas às obtidas na classificação quanto à malignidade (#1). Dessa forma, a Tabela 6.4 apresenta os melhores desempenhos para cada tipo de TL aplicado no ViT-Base, onde elas são ranqueadas em ordem decrescente de performance.

Tabela 6.4. Melhores resultados da classificação #2 entre as variações de TL no ViT-Base.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Prec. (macro)	Sensib. (macro)	Especif. (macro)	F1 (macro)	Acurác. (micro)	AUC
ViT-Base	1	#3	2	Acurác.	**0,874**	**0,709**	**0,712**	**0,911**	**0,710**	**0,748**	**0,881**
ViT-Base	4	#3	1	Custo	0,819	0,608	0,674	0,877	0,620	0,638	0,872
ViT-Base	2	#3	1	Custo	0,794	0,564	0,628	0,861	0,571	0,589	0,847
ViT-Base	3	#3	2	Custo	0,821	0,475	0,497	0,867	0,483	0,642	0,796

Semelhante ao resultado encontrado nos modelos CNN, visualizado na Tabela 6.5, o pré-processamento #3 produziu o modelo com melhor desempenho para o subconjunto de teste do BUS-BRA. Também se observa que o melhor modelo para todas as métricas foi aquele no qual se realizou o ajuste fino completo dos parâmetros treináveis (TL-1), alcançando uma sensibilidade média por classe de $91,1\,\%$ e F1-*Score* de $75,8\,\%$.

Ao analisar outros resuldados do modelo ViT-Base com melhor desempenho, destaca-se a matriz de confusão apresentada na Figura 6.2. Nesta, nota-se uma maior dificuldade no modelo na distinção entre a categoria 5 da 4 e entre 2 e 3, semelhante a matriz de confusão do MixNet-XL no capítulo anterior (Figura 5.2).

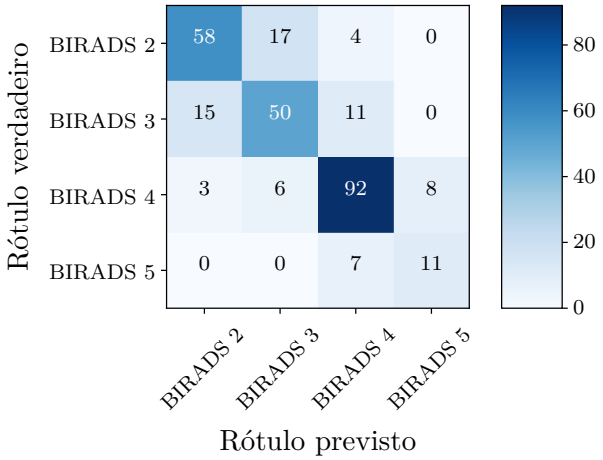


Figura 6.2. Matriz de confusão para o melhor modelo ViT-Base na classificação BI-RADS multiclasse com total de 282 amostras.

Entre as amostras cujo rótulo verdadeiro era BI-RADS 4, verificou-se que aquelas nas quais o modelo atribuiu a classificação BI-RADS 2 ou 3 representavam a $64\,\%$ (7 amostras) com resultado de benigno de biópsia. Enquanto, nas que predição foi BI-RADS 5, $87,5\,\%$ (7 amostras) são lesões malignas confirmadas por biópsia. Contudo, uma das amostras identificadas como BI-RADS 2 foi classificada como maligna. Esse erro é considerado grave, em que a amostra corresponde a um caso de linfoma. Este tipo de patologia tem baixa representatividade no banco de dados de treinamento ($0,3\,\%$ do total de amostras) e impõe aos modelos ViT-Base e MixNet-XL dificuldade em classificá-la como suspeita de malignidade (BI-RADS 4) ou altamente sugestiva de malignidade (BI-RADS 5).

O AUC apresentado na Tabela 6.2 é a média das quatro classes BI-RADS. Essa métrica, em conjunto com as curvas ROC para cada classe do ViT-Base, é apresentada na Figura 6.3. Os rótulos relacionados a suspeita de câncer, BI-RADS 4 e 5, têm o maior desempenho semelhante ao melhor modelo CNN, o MixNet-XL. Porém, ao compará-los, o MixNet-XL apresenta melhor desempenho para identificação dos BI-RADS 4 e 5.

Quando se avalia o ViT com as demais arquiteturas CNN pelos resultados da Tabela 6.5, ele tem o melhor desempenho geral. O ViT é o único modelo que alcança

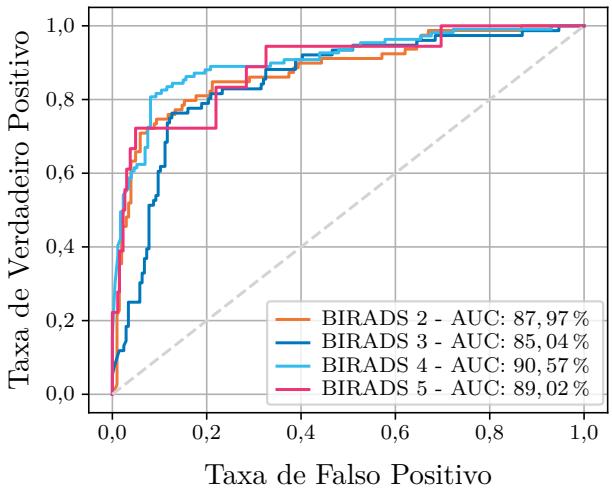


Figura 6.3. Curvas ROC por classe para o melhor modelo ViT-Base na classificação BI-RADS multiclasse.

desempenho superior a 70 % para todas as métricas avaliadas. Comparando a curva ROC média de cada modelo, apresentada na Figura 6.4, se observa um comportamento muito entre as curvas. Consequentemente, a diferença do AUC médio para cada modelo não varia mais que $\approx 2\,\%$.

Tabela 6.5. Melhores resultados da classificação #2 entre os modelos propostos.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Prec. (macro)	Sensib. (macro)	Especif. (macro)	F1 (macro)	Acurác. (micro)	AUC
VGG16	1	#3	1	Custo	0,620	0,661	0,620	0,892	0,634	0,699	0,887
ResNet50D	1	#3	3	Custo	0,846	0,660	0,646	0,891	0,652	0,691	0,887
EfficientNet-B3	1	#3	4	*Acurác.	0,849	0,726	0,601	0,889	0,626	0,699	0,871
RegNetY-3.2GF	1	#2	3	**Último	0,849	0,690	0,670	0,892	0,678	0,699	0,880
MixNet-XL	1	#3	3	*Acurác.	0,856	**0,744**	0,645	0,894	0,675	0,713	0,887
ViT-Base	1	#3	2	Acurác.	**0,874**	0,709	**0,712**	**0,911**	**0,710**	**0,748**	0,881

*: Acurácia e Custo com mesmo desempenho. **: Acurácia, Custo e Último com mesmo desempenho.

Na aplicação clínica de triagem, o ViT-Base é o modelo mais indicado, visto que sua especificidade ($91,1\,\%$) é a maior entre os melhores resultados de cada arquitetura.

Para avaliar o desempenho dos modelos implementados em outros bancos de dados, todos foram testados nas categorias BI-RADS 2 a 5 no BrEaST e de 3 a 5 no OASBUD, nas quais as subdivisões da categoria 4 são agrupadas em uma única classe. Os melhores modelos, incluindo o ViT-Base, estão presentes na Tabela 6.6.

Comparando os resultados do ViT, Tabela 6.6, com os de teste, observa-se uma redução de desempenho para todas as métricas. Isso sugere uma dificuldade do ViT para classificação multiclasse BI-RADS e uma possível necessidade de mais dados de treinamento para melhorar os modelos, conforme observado em uma aplicação mais generalista usando o ImageNet1K (48).

Para as amostras do BrEaST, o melhor modelo geral se mantém o EfficientNet-B3, que alcança AUC, F1-*Score* (macro), e acurácia (macro) de $69,9\,\%$, $43,5\,\%$ e $75,6\,\%$, res-

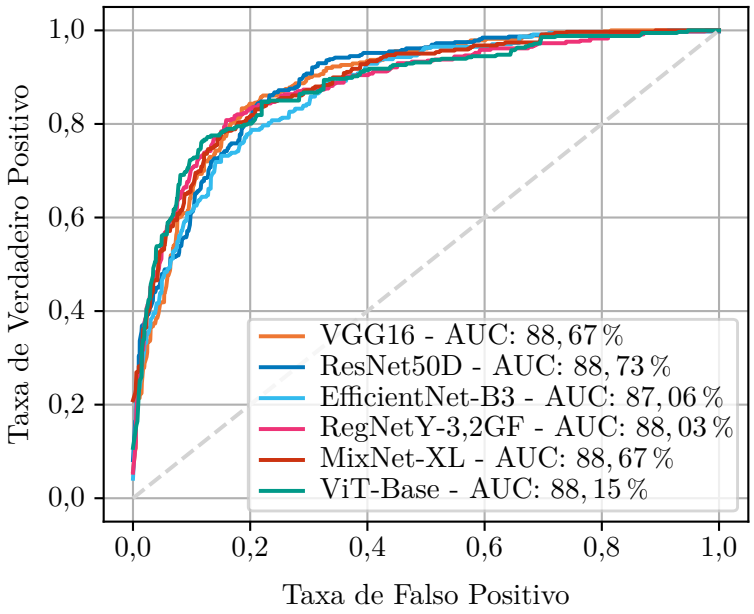


Figura 6.4. Curvas ROC dos melhores modelos para classificação BI-RADS multiclasse (#2).

pectivamente. Embora o RegNetY-3.2GF evidencia desempenho superior no OASBUD, o ViT é o imediatamente seguinte, apresentando AUC de 59,2 %, F1-*Score* de 40,5 % e acurácia (macro) de 67,2 % para classificação #2. Os resultados no OASBUD são semelhantes aos modelos que utilizam um único pré-processamento em (99). Porém, são inferiores quando comparados ao uso de múltiplos modelos com votação suave (99).

Considerando as três avaliações, o ViT é o com melhor desempenho geral para classificação multiclasse do BI-RADS, a mais difícil entre as avaliadas. Contudo, no caso a aplicação em triagem, o MixNet-XL é o modelo mais indicado, pois apresenta as maiores especificidades em duas das três avaliações.

6.3 Classificação BI-RADS (#3)

A classificação #3 é o experimento onde as métricas tendem a alcançar os maiores valores possíveis. Devido aos modelos serem treinados e testados apenas com amostras de BI-RADS 2 e 5, estas que são classes extremas uma das outras entre as categorias de uma lesão mamária.

Para a arquitetura ViT-Base, primeiramente são apresentados os melhores modelos treinados para as diferentes técnicas de TL implementadas na Tabela 6.7. O modelo com maior desempenho geral foi aquele que utilizou o ajuste fino completo durante o treinamento (TL-1) e o pré-processamento #3, resultado análogo às classificações anteriores.

Os resultados da Tabela 6.7 indicam que o ajuste fino completo (TL-1) se aproximou

Tabela 6.6. Melhores resultados da classificação #2 entre os modelos propostos para os bancos de dados BrEaST e OASBUD.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Prec. (macro)	Sensib. (macro)	Especif. (macro)	F1 (macro)	Acurác. (micro)	AUC
BrEaST											
VGG16	1	#3	1	Acurác.	0,728	0,427	**0,445**	0,794	0,418	0,456	**0,737**
ResNet50D	1	#3	2	Acurác.	0,756	**0,462**	0,419	0,798	0,411	0,512	0,714
EfficientNet-B3	1	#3	4	Último	0,756	0,438	0,436	0,797	**0,435**	0,512	0,699
RegNetY-3.2GF	1	#3	4	Custo	0,756	0,432	0,432	0,801	0,426	0,512	0,688
MixNet-XL	1	#2	3	Último	**0,774**	0,455	0,414	**0,802**	0,424	0,548	0,663
ViT-Base	1	#3	2	Custo	0,756	0,433	0,415	0,803	0,416	0,512	0,673
OASBUD											
VGG16	1	#1	1	*Acurác.	0,668	0,462	0,378	0,771	0,393	0,425	0,631
ResNet50D	1	#3	3	Acurác.	0,685	**0,727**	0,319	0,813	0,355	0,375	0,667
EfficientNet-B3	1	#4	3	*Acurác.	0,650	0,471	0,383	0,747	0,410	0,415	0,630
RegNetY-3.2GF	1	#2	3	Último	0,667	0,510	**0,424**	0,721	**0,414**	**0,490**	0,603
MixNet-XL	1	#3	4	**Custo	**0,688**	0,567	0,365	**0,870**	0,412	0,335	**0,681**
ViT-Base	1	#2	2	Último	0,672	0,482	0,420	0,730	0,405	0,490	0,592

*: Acurácia e Custo com mesmo desempenho. **: Custo e Último com mesmo desempenho.

Tabela 6.7. Melhores resultados da classificação #3 entre as variações de TL no ViT-Base.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
ViT-Base	1	#3	1	Último	**0,991**	**0,991**	**1,000**	**0,952**	**1,000**	**0,976**	**0,999**
ViT-Base	2	#3	1	Custo	0,982	0,982	1,000	0,905	1,000	0,950	0,998
ViT-Base	4	#3	2	Custo	0,972	0,972	0,950	0,905	0,989	0,927	0,996
ViT-Base	3	#3	2	Custo	0,963	0,963	0,947	0,857	0,989	0,900	0,995

do resultado "perfeito", onde os erros ocorreram na identificação de algumas imagens do BI-RADS 5. A segunda melhor técnica (TL-2) ficou próxima da primeira na maioria das métricas, com exceção da sensibilidade, que diminuiu $\approx 5\,\%$. Este modelo se destaca pelo menor custo computacional durante o treinamento, pois o TL-2 se refere ao modelo onde apenas o classificador MLP é refinado nessa etapa.

Na Tabela 6.8 estão listados os melhores modelos de cada arquitetura para a classificação #3. Semelhante aos casos anteriores, o pré-processamento #3 se mantém o melhor para classificação das imagens de lesões mamárias, como maligno ou não, BI-RADS 2 e 5, e para BI-RADS multiclasse (2, 3, 4 e 5).

Tabela 6.8. Melhores resultados da classificação #3 entre os modelos propostos.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#3	*1	**Custo	1,000	1,000	1,000	1,000	1,000	1,000	1,000
ResNet50D	1	#3	4	Último	1,000	1,000	1,000	1,000	1,000	1,000	1,000
EfficientNet-B3	1	#3	*1	**Acurác.	1,000	1,000	1,000	1,000	1,000	1,000	1,000
RegNetY-3.2GF	1	#3	1	***Custo	1,000	1,000	1,000	1,000	1,000	1,000	1,000
MixNet-XL	1	#3	1	Último	0,994	0,991	0,955	1,000	0,989	0,977	0,999
ViT-Base	1	#3	1	Último	0,991	0,991	1,000	0,952	1,000	0,976	0,999

*: Amostrador 1 e 2 com mesmo desempenho. **: Custo, Acurácia e Último com mesmo desempenho. ***: Custo e Acurácia com mesmo desempenho. ****: Custo e Último com mesmo desempenho.

Outro aspecto que se destaca é a possível necessidade de melhor refinamento dos hiperparâmetros para o ViT-Base atingir o resultado "perfeito", igual aos observados

para o VGG16, ResNet50D, EfficientNet-B3 e RegNetY-3.2GF. Adicionalmente, levanta-se a hipótese de que uma avaliação mais profunda, como uso 5-Fold ou 10-Fold, seja necessária para validar esses modelos na classificação #3. Devido a uma variação entre os subconjuntos de treinamento, validação e teste possa alterar os resultados finais.

A Tabela 6.9 apresenta os modelos com melhor desempenho aplicadas aos conjuntos de dados BrEaST e OASBUD. Semelhante ao que ocorreu nas amostras de teste, a classificação #3 apresenta os maiores valores para as métricas avaliadas. Outro detalhe é que, no BrEaST, ocorreu maior performance geral na acurácia, sensibilidade, F1-*Score* e AUC. Por outro lado, no OASBUD, todos os modelos apresentaram maiores valores na precisão e especificidade, estas não utilizadas durante o ranqueamento dos modelos.

Tabela 6.9. Melhores resultados da classificação #3 entre os modelos propostos para os bancos de dados BrEaST e OASBUD.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
BrEaST											
VGG16	1	#3	1	Acurác.	0,868	0,868	0,860	0,935	0,767	0,896	0,927
ResNet50D	1	#3	2	*Acurác.	0,895	0,895	0,896	0,935	0,833	0,915	**0,964**
EfficientNet-B3	1	#3	4	**Custo	0,868	0,868	0,875	0,913	0,800	0,894	0,936
RegNetY-3.2GF	1	#3	4	Acurác.	0,895	0,895	0,880	**0,957**	0,800	**0,917**	0,946
MixNet-XL	1	#3	3	Último	0,868	0,868	0,875	0,913	0,800	0,894	0,909
ViT-Base	1	#3	1	**Custo	0,882	0,882	**0,911**	0,891	**0,867**	0,901	0,940
OASBUD											
VGG16	1	#3	1	Acurác.	0,848	0,848	0,923	0,720	0,952	0,809	0,881
ResNet50D	1	#3	3	Custo	0,821	0,821	0,826	0,760	0,871	0,792	0,859
EfficientNet-B3	1	#3	4	**Custo	0,813	0,813	**1,000**	0,580	**1,000**	0,734	0,880
RegNetY-3.2GF	1	#3	1	Custo	**0,857**	**0,857**	0,972	0,700	0,984	**0,814**	0,895
MixNet-XL	1	#3	2	Custo	0,830	0,830	0,897	0,700	0,935	0,787	0,850
ViT-Base	3	#3	2	Acurác.	0,830	0,830	0,844	0,760	0,887	0,800	**0,902**

*: Acurác., Custo e Último com mesmo desempenho. **: Custo e Último com mesmo desempenho.

Comparando as arquiteturas CNN com o ViT, Tabela 6.9, observa-se valores semelhantes para maioria das métricas, comprovando a eficácia do ViT em relação a CNN na classificação #3. Especificamente no OASBUD, o ViT que utilizou o ajuste fino parcial do *encoder* de imagem e do classificador (*Transfer Learning* (TL)-3) foi que apresentou o melhor desempenho. Alcançando AUC de $90,2\,\%$, F1-*Score* de $80,0\,\%$, sensibilidade de $76,0\,\%$ e acurácias de $83,0\,\%$.

No BrEaST, o RegNetY-3.2GF se manteve como melhor arquitetura, enquanto o ViT foi o terceiro, atrás do ResNet50D. Contudo, o ViT alcançou a melhor precisão ($91,1\,\%$) e especificidade ($89,1\,\%$) comparado as outras arquiteturas. Isso indica para uma maior eficiência na identificação da BI-RADS 2 e precisão em acertar o BI-RADS 5. Considerando todos os testes realizados, a melhor arquitetura para classificação #3 é o RegNetY-3.2GF, sendo necessário a utilização da imagem segmentada por fusão da ROI (Pré-proc. #3). Por outro lado, tanto o EfficientNet-B3 quanto o ViT-Base são arquiteturas que apresentaram desempenho destacado quanto à especificidade na classificação #3. Consequentemente, esse modelo é uma boa escolha para uso em triagem.

6.4 Classificação BI-RADS (#4)

A classificação #4 se refere tipo que as quatro categorias BI-RADS são agrupadas em duas classes para verificar se ocorre melhoria de desempenho ao simplificar a classificação para uma variante binária. Nesse caso, as categorias BI-RADS são separadas em classe A, com as categorias 2 e 3, e B, com as categorias 4 e 5. A classe A representa as amostras com maior probabilidade de serem benignas, enquanto a classe B representa aquelas com suspeita de câncer (diagnóstico positivo de malignidade).

A Tabela 6.10 apresenta os modelos com melhor desempenho da arquitetura ViT-Base para as quatro variantes de *transfer learning*. Todos os modelos alcançaram esse resultado utilizando o pré-processamento #3, semelhante ao ocorrido nas classificações #1 a #3.

Tabela 6.10. Melhores resultados da classificação #4 entre as variações de TL no ViT-Base.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
ViT-Base	1	#3	1	Último	**0,922**	**0,922**	**0,920**	**0,906**	**0,935**	**0,913**	**0,941**
ViT-Base	3	#3	2	Acurác.	0,890	0,890	0,907	0,843	0,929	0,873	0,935
ViT-Base	2	#3	1	Último	0,883	0,883	0,873	0,866	0,897	0,870	0,924
ViT-Base	4	#3	1	Custo	0,872	0,872	0,847	0,874	0,871	0,860	0,923

Como o modelo da primeira linha da Tabela 6.10 se refere ao modelo com a maior performance, observa-se que o ajuste fino completo da arquitetura ViT apresentou o melhor resultado para o subconjunto de teste do BUS-BRA. Neste modelo ocorreu um aumento substancial de precisão e sensibilidade ao compará-lo com a classificação BI-RADS multiclasse da Tabela 6.4. Desconsiderando esse modelo, os que utilizaram o ajuste fino do classificador e parte do *encoder* de imagem (TL-3) e o ajuste fino isolado do classificador (TL-2) foram os melhores, em ordem decrescente. A redução na sensibilidade se destaca como a mais importante entre as métricas, o que explica a maior ocorrência de erros na identificação da classe B (categorias BI-RADS 4 e 5).

A Tabela 6.11 agrupa os melhores resultados para todas as seis arquiteturas propostas. O principal destaque foi o ViT-Base ser o único modelo que alcança valores de 90 % ou mais para todas as métricas.

Tabela 6.11. Melhores resultados da classificação #4 entre os modelos propostos.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#3	2	Acurác.	0,894	0,897	0,902	0,866	0,923	0,884	0,939
ResNet50D	1	#3	3	Custo	0,887	0,887	0,874	0,874	0,897	0,874	0,946
EfficientNet-B3	1	#3	1	Acurác.	0,867	0,865	0,830	0,882	0,852	0,855	0,926
RegNetY-3.2GF	1	#3	3	Último	0,908	0,908	0,886	0,913	0,903	0,899	**0,949**
MixNet-XL	1	#3	4	Último	0,897	0,897	0,866	0,913	0,884	0,889	0,948
ViT-Base	1	#3	1	Último	**0,922**	**0,922**	**0,920**	0,906	**0,935**	**0,913**	0,941

Quando comparado aos outros modelos, o ViT-Base se equipara em desempenho do RegNetY-3.2GF e do MixNet-XL, estes que foram os melhores com arquitetura CNN. O ViT obteve acurácia macro e binária entre 1–2 % melhor, especificidade 3–5 % maior, precisão 2,5 % maior e, consequentemente, F1-*Score* ≈ 1 % maior. Contudo, existe uma piora quanto à AUC média, sendo ≈ 1 % pior que o RegNetY-3.2GF. Essa diferença torna-se mais evidente na comparação das curvas ROC por classe do ViT-Base (Figura 6.5b) com as do RegNetY-3.2GF (Figura 6.5a).

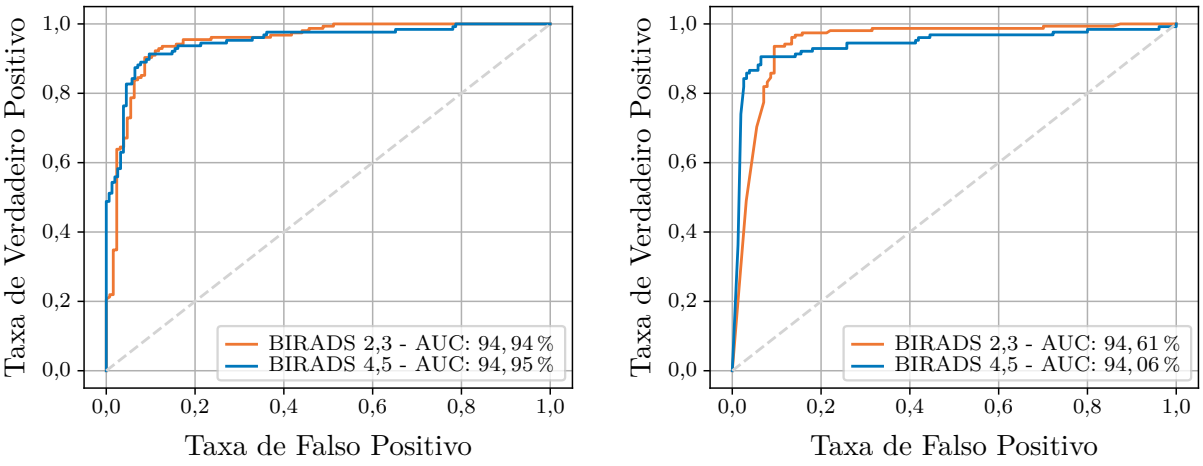


Figura 6.5. Curvas ROC por classe para classificação BI-RADS #4. Em que (a) é para o melhor modelo RegNetY-3.2GF e (b) para o melhor ViT-Base.

Na Figura 6.5, observa-se que o RegNetY-3.2GF realiza a classificação para ambos os rótulos de maneira aproximadamente igual para toda a faixa da taxa de FP. Enquanto o ViT-Base tende a "acertar" com baixa taxa de FP até um limite de 90 % da taxa de VP, se estabilizando em valores maiores. Quando calculado a área para ambos os modelos, se observa uma melhor probabilidade de classificação do rótulo maligno (BI-RADS 4 e 5) no RegNetY-3.2GF em comparação com o ViT-Base.

Quando comparado o ROC binário entre os modelos, na Figura 6.6, o ViT tem características semelhantes aos melhores modelos CNN, mas com o 4º maior AUC. O RegNetY-3.2GF é o melhor modelo para classificação #4, mesmo não alcançando os maiores valores de acurácia, precisão e F1-*Score* comparado ao ViT. Considerando a aplicação por triagem, o ViT é o modelo com maior eficácia pois apresenta a maior especificidade entre as arquiteturas ao obter valor de 93,5 %.

Aplicando a mesma avaliação feita anteriormente para os dados do BrEaST e OAS-BUD, são apresentados os modelos de melhor desempenho na Tabela 6.12. No geral, o desempenho para todas as métricas diminuiu comparado ao observado nas amostras de teste (Tabela 6.11). Principalmente para a especificidade que, nos testes, variava em torno de 90 %. Enquanto, nesses dados, ocorre uma diminuição entre 45–65 %.

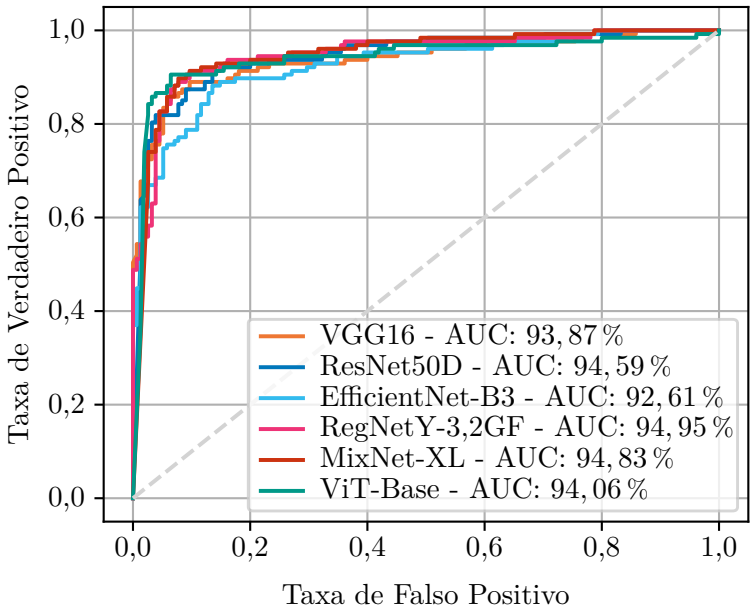


Figura 6.6. Curvas ROC dos melhores modelos para classificação BI-RADS #4.

Comparando os resultados da Tabela 6.12, evidencia-se que o ViT obteve a maior acurácia, sensibilidade e F1-*Score*. Isso sugere um bom desempenho para classificação #4 devido a também ter alcançado as maiores métricas para o subconjunto de teste do BUS-BRA, conforme Tabela 6.11. Em termos do pré-processamento utilizado, o ViT com melhor performance geral utilizou o #2, esta composta de 3 canais e proposta nesse estudo. Também se observa um desempenho superior com o uso do amostrador ponderado (Amost.-1) durante o treinamento, este presente em todos os melhores modelos ViT treinados.

Considerando que o ViT-Base foi a melhor arquitetura em dois dos três testes, concluise que sua aplicação é a mais indicada para classificação #4. Observando a necessidade do uso de um amostrador ponderado durante o treinamento, realizado o TL por ajuste fino completo, e utilizado uma entrada de 3 canais que segmenta a lesão na imagem.

Um destaque importante é o caso do RegNetY-3.2GF para triagem, pois a arquitetura apresentou especificidade constante e elevada nos três testes, ainda que não tenha alcançado o resultado geral mais alto em nenhum deles.

6.5 Classificação BI-RADS (#5)

A classificação #5 é uma variante da BI-RADS que a categoria 5 é isolada na classe B, enquanto o restante é agrupado na A. Os modelos treinados nessa classificação têm como objetivo identificar a classe em que a lesão contida na imagem é altamente sugestiva de câncer. Essa que foi uma das deficiências encontradas na matriz de confusão do melhor

Tabela 6.12. Melhores resultados da classificação #4 entre os modelos propostos para os bancos de dados BrEaST e OASBUD.

BrEaST											
Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#3	2	Último	0,734	0,734	0,839	0,789	0,582	0,813	0,769
ResNet50D	1	#3	2	Custo	0,738	0,738	**0,870**	0,757	**0,687**	0,809	0,790
EfficientNet-B3	1	#3	2	Acurác.	0,746	0,746	0,807	0,859	0,433	0,832	0,798
RegNetY-3.2GF	1	#3	4	*Acurác.	0,750	0,750	0,851	0,800	0,612	0,825	**0,820**
MixNet-XL	1	#3	1	Acurác.	0,766	0,766	0,842	0,838	0,567	0,840	0,784
ViT-Base	1	#2	1	Último	**0,778**	**0,778**	0,831	**0,876**	0,507	**0,853**	0,761
OASBUD											
VGG16	1	#1	1	Acurác.	0,750	0,750	0,782	0,884	0,452	0,830	0,712
ResNet50D	1	#2	1	Último	0,725	0,725	0,817	0,775	0,613	0,796	0,777
EfficientNet-B3	1	#3	4	Último	0,760	0,760	**0,888**	0,746	**0,790**	0,811	0,829
RegNetY-3.2GF	1	#3	4	Acurác.	0,790	0,790	0,853	0,841	0,677	0,847	0,823
MixNet-XL	1	#2	3	Custo	0,720	0,720	0,770	0,848	0,435	0,807	0,753
ViT-Base	1	#2	1	**Custo	**0,835**	**0,835**	0,839	**0,942**	0,597	**0,887**	**0,858**

*: Acurác. e Último com mesmo desempenho. **: Acurác. e Custo com mesmo desempenho.

modelo da classificação #2.

No caso da arquitetura ViT-Base, os modelos com maior desempenho para cada variação de TL são apresentados na Tabela 6.13. O principal resultado observado foi que o melhor modelo utilizou um pré-processamento diferente do #3, quando comparado às classificações anteriores. Nesse caso, o pré-processamento proposto nessa pesquisa com 3 canais (#2) apresentou o melhor desempenho.

Tabela 6.13. Melhores resultados da classificação #5 entre as variações de TL no ViT-Base.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
ViT-Base	1	#2	1	Último	**0,957**	**0,957**	**0,750**	0,500	**0,989**	**0,600**	0,903
ViT-Base	4	#3	2	Último	0,780	0,780	0,225	1,000	0,765	0,367	0,924
ViT-Base	3	#3	2	Último	0,759	0,759	0,209	1,000	0,742	0,346	**0,927**
ViT-Base	2	#3	2	Custo	0,787	0,787	0,216	0,889	0,780	0,348	0,910

Ao analisar a Tabela 6.13, observa-se uma dificuldade nos modelos em identificar corretamente o BI-RADS 5 com alta precisão. O ViT com as melhores métricas utilizou o ajuste fino completo como técnica de TL e alcançou 75 % de precisão com apenas 50 % de sensibilidade. Enquanto, as outras variantes de TL, apresentaram baixa precisão ($\approx$ 21 %) e alta sensibilidade (entre 90-100 %). Devido a essa grande variação, o F1-*Score* é uma das métricas mais importantes para definir o melhor modelo na classificação #5, esta que é uma média harmônica entre precisão e sensibilidade.

Quando se compara o ViT com os outros modelos, conforme resultado da Tabela 6.14, o fenômeno discutido anteriormente se repete em todos os casos. Ou seja, as amostras BI-RADS 5 (B) são difíceis de distinguir da classe A, esta que agrupa as categorias 2, 3 e 4. Considerando a matriz de confusão do ViT para classificação BI-RADS multiclasse (#2), Figura 6.2, a dificuldade está relacionada em identificar diferenças entre as categorias 4

e 5.

Tabela 6.14. Melhores resultados da classificação #5 entre os modelos propostos.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#1	1	Último	0,742	0,954	0,692	0,500	0,985	0,581	0,872
ResNet50D	1	#3	3	Custo	0,947	0,947	0,615	0,444	0,981	0,516	0,947
EfficientNet-B3	1	#1	3	Acurác.	0,957	0,957	0,714	0,556	0,985	0,625	0,943
RegNetY-3.2GF	1	#2	1	Último	0,772	**0,961**	**0,769**	0,556	0,989	**0,645**	**0,953**
MixNet-XL	1	#4	3	Custo	0,947	0,947	0,600	0,500	0,977	0,545	0,884
ViT-Base	1	#2	1	Último	0,957	0,957	0,750	0,500	0,989	0,600	0,903

Comparando o resultado de todos os modelos da Tabela 6.14, o melhor é o RegNetY-3.2GF. Ele alcança o maior valor de sensibilidade (55,6 %), F1-*Score* (64,5 %) e AUC médio (95,3 %). Cabe mensionar que o pré-processamento #2 foi o utilizado tanto no melhor modelo quanto no ViT-Base.

Na Figura 6.7 são exibidas as curvas ROC de cada modelo proposto para a classificação #5. Observa-se que os piores modelos apresentam uma maior taxa de falso positivo (FP) em baixos valores da taxa de verdadeiro positivo (VP), indicando um desempenho inferior. O RegNetY-3.2GF é o modelo com a maior taxa de VP antes do aumento significato da probabilidade de FP. O EfficientNet-B3 é um modelo que se destaca por atingir uma alta taxa de verdadeiro positivo em uma faixa com baixa taxa de falso positivo. Esse fenômeno ocorre devido à sua elevada acurácia e sensibilidade, em comparação com o RegNetY.

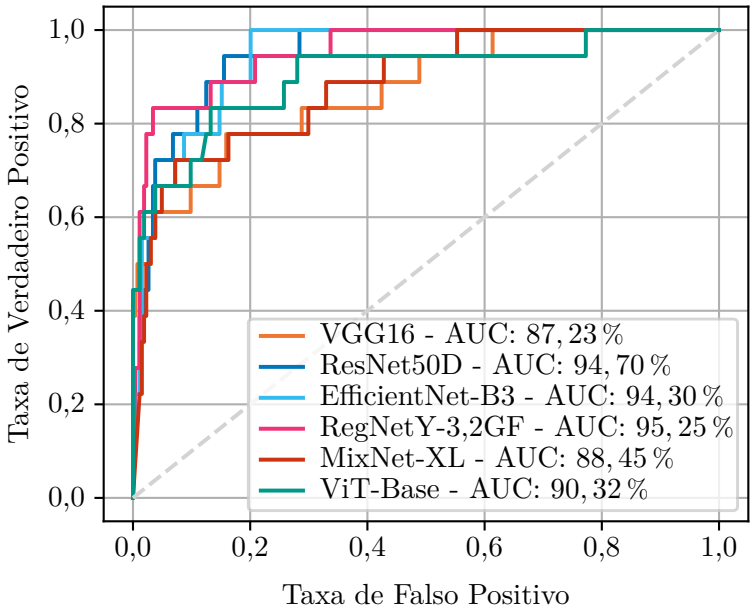


Figura 6.7. Curvas ROC dos melhores modelos para classificação BI-RADS #5.

Avaliando os modelos propostos nas imagens do BrEaST e OASBUD, são obtidos os resultados da Tabela 6.15. Na qual a classe A inclui as categorias BI-RADS 3 e 4

no OASBUD, enquanto o BrEaST é igual ao feito para o subconjunto de teste do BUS-BRA. Para ambas as imagens, o ViT apresentou comportamento semelhante ao teste no subconjunto de teste, onde ocorre uma variabilidade da precisão e sensibilidade. No BrEaST, a sensibilidade do ViT foi 21,8 % maior que o segundo melhor modelo, contudo existiu uma diminuição considerável na precisão e na especificidade. Outro detalhe é que essa arquitetura foi treinada com ajuste fino do *encoder* de imagem e o classificador (TL-4), diferente das classificações anteriores.

Tabela 6.15. Melhores resultados da classificação #5 entre os modelos propostos para os bancos de dados BrEaST e OASBUD.

BrEaST											
Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#1	1	*Custo	0,688	0,821	0,512	0,478	0,898	0,494	0,764
ResNet50D	1	#4	1	Acurác.	0,727	**0,829**	**0,531**	0,565	0,888	**0,547**	0,807
EfficientNet-B3	1	#3	4	Último	0,806	0,806	0,463	0,413	0,893	0,437	0,790
RegNetY-3.2GF	1	#1	3	Último	0,770	0,770	0,417	0,652	0,796	0,508	0,776
MixNet-XL	1	#1	4	Último	**0,821**	0,821	0,514	0,413	**0,913**	0,458	0,749
ViT-Base	4	#3	1	Custo	0,726	0,726	0,388	**0,870**	0,694	0,537	**0,809**
OASBUD											
VGG16	1	#2	2	**Custo	0,695	0,695	0,430	**0,680**	0,700	0,527	0,735
ResNet50D	1	#3	2	Acurác.	0,775	0,775	0,593	0,320	0,927	0,416	0,739
EfficientNet-B3	1	#3	4	Último	**0,785**	**0,785**	**0,733**	0,220	0,973	0,338	0,714
RegNetY-3.2GF	1	#3	1	Último	0,760	0,760	0,600	0,120	0,973	0,200	0,699
MixNet-XL	1	#3	4	Último	0,750	0,750	0,500	0,360	0,880	0,419	0,740
ViT-Base	1	#3	1	Último	0,770	0,770	0,534	0,620	0,820	**0,574**	**0,797**

*: Acurác. e Custo com mesmo desempenho. **: Custo e Último com mesmo desempenho.

Para os resultados do BrEaST na Tabela 6.15, o ResNet50D se mantém como arquitetura com o melhor desempenho geral. O ViT é o terceiro colocado, mas tem a maior sensibilidade (87,0 %) e AUC (80,9 %). No OASBUD, o ViT apresenta melhores resultados que as arquiteturas CNN, como o ResNet50D. Observam-se valores de 79,7 % para AUC, 57,4 % para F1, 62,0 % para sensibilidade e 77 % em ambas as acurácias. O melhor modelo geral, ao considerar os três testes, fica empatado entre o ResNet50D e o ViT. Adicionando o AUC como critério de desempate, considerando que é uma métrica utilizada para ranqueamento de classificadores, o ViT-Base é a melhor arquitetura para identificação do BI-RADS 5 isoladamente (classificação #5).

6.6 Classificação BI-RADS (#6)

Na classificação BI-RADS #6, a abordagem é o oposto da anterior: as amostras BI-RADS 2 são isoladas na classe A, enquanto o restante é agrupado na B. Devido ao tipo de categoria que é isolada, são aplicadas duas avaliações distintas para ranqueamento dos melhores modelos e das respectivas técnicas de TL.

A primeira, apresentada na primeira parte da Tabela 6.16 para os modelos ViT, foi a mesma utilizada nas classificações anteriores, realizando o ranqueando pela acurácia

binária, sensibilidade, F1-*Score* e AUC. Na segunda, os modelos são ranqueados com a substituição da sensibilidade pela especificidade no ranqueamento anterior. Esta é uma métrica que avalia melhor os modelos, seguindo o objetivo de verificar os melhores modelos para classificação das amostras BI-RADS #2 das outras.

Tabela 6.16. Melhores resultados da classificação #6 entre as variações de TL no ViT-Base. A avaliação de sensibilidade mantém o mesmo ranqueamento aplicado nas classificações anteriores, enquanto, na segunda avaliação, a especificidade substitui a sensibilidade.

Aval.	Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
Sensib.	ViT-Base	1	#2	2	Último	**0,855**	**0,855**	**0,875**	0,931	**0,658**	**0,902**	**0,854**
	ViT-Base	4	#2	2	Acurác.	0,794	0,794	0,811	0,931	0,443	0,867	0,729
	ViT-Base	2	#2	2	Acurác.	0,780	0,780	0,795	0,936	0,380	0,860	0,769
	ViT-Base	3	#2	2	Acurác.	0,798	0,798	0,817	0,926	0,468	0,868	0,708
Especif.	ViT-Base	1	#2	1	Acurác.	**0,855**	**0,855**	0,886	**0,916**	0,696	**0,901**	**0,866**
	ViT-Base	2	#3	1	Acurác.	0,617	0,617	1,000	0,468	1,000	0,638	0,824
	ViT-Base	4	#3	2	Custo	0,589	0,589	1,000	0,429	1,000	0,600	0,807
	ViT-Base	3	#3	2	Acurác.	0,589	0,589	1,000	0,429	1,000	0,600	0,802

Comparando os resultados das duas avaliações da Tabela 6.16, observa-se que, em ambas, o modelo com melhor desempenho utilizou o TL de ajuste fino completo. Em relação ao pré-processamento, na primeira avaliação (sensib.), todos os melhores modelos utilizaram o proposto nesse estudo, a variante #2 que adota imagens de 3 canais. Na avaliação da especificidade, o modelo de melhor desempenho também utiliza esse pré-processamento, porém os demais modelos utilizam a variante #3.

Em relação aos modelos com arquitetura ViT, a avaliação de sensibilidade aponta para um comportamento semelhante a classificação #5, mas aplicada aos casos benignos (BI-RADS 2). Ou seja, nesses modelos se observam uma alta precisão com relativo baixo desempenho para identificar amostras BI-RADS 2 sem error. Para os modelos ranqueados na segunda forma (especificidade), o melhor modelo apresenta comportamento semelhante ao anterior. Porém, as outras três variantes de TL atingem especificidade e precisão "perfeita", com custo elevado na sensibilidade (média de $44,2\,\%$).

Na Tabela 6.17 são apresentados todos os melhores modelos para cada arquitetura com base no ranqueamento por sensibilidade, mesmo aplicado nas classificações anteriores.

Tabela 6.17. Melhores resultados da classificação #6 entre os modelos propostos utilizando o ranqueamento por acurácia binária, sensibilidade, F1-*Score* e AUC.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#3	2	Acurác.	0,730	0,823	0,834	0,941	0,519	0,884	0,893
ResNet50D	1	#3	4	Acurác.	0,844	0,844	0,860	0,936	0,608	0,896	0,871
EfficientNet-B3	1	#2	4	Acurác.	0,848	0,848	0,864	0,936	0,620	0,898	0,846
RegNetY-3.2GF	1	#3	1	Acurác.	0,812	**0,869**	**0,884**	0,941	**0,684**	**0,912**	**0,901**
MixNet-XL	1	#2	4	Acurác.	0,851	0,851	0,864	0,941	0,620	0,901	0,859
ViT-Base	1	#2	2	Último	**0,855**	0,855	0,875	0,931	0,658	0,902	0,854

De acordo com os resultados observados na Tabela 6.17, não existe uma predominância em relação ao melhor pré-processamento, pois tanto o #3 como #2 foram utilizados para o melhor desempenho em determinadas arquiteturas. Considerando o ranqueamento aplicado, o RegNetY-3.2GF obteve o melhor desempenho para classificação #2. Na qual, a especificidade foi 3 % superior e apresentou maior acurácia binária e precisão em relação segundo colocado, o ViT-Base.

Analisando os melhores modelos quando substituído a sensibilidade pela especificidade no ranqueamento, Tabela 6.18, a arquitetura RegNetY se mantém com o melhor desempenho geral. Nesse caso, ocorre a modificação do pré-processamento utilizado, do #3 para o #2. A principal diferença observada é a especificidade ser ≈ 6 % superior ao segundo colocado, o ViT-Base, e apresentar o maior valor de F1-*Score*.

Tabela 6.18. Melhores resultados da classificação #6 entre os modelos propostos utilizando o ranqueamento por acurácia binária, especificidade, F1-*Score* e AUC.

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
VGG16	1	#1	1	Último	0,778	0,837	0,869	0,911	0,646	0,889	0,848
ResNet50D	1	#2	1	Acurác.	0,789	0,840	0,876	0,906	0,671	0,891	0,858
EfficientNet-B3	1	#2	3	Acurác.	0,826	0,826	0,877	0,882	0,684	0,880	0,878
RegNetY-3.2GF	1	#2	1	Último	0,846	**0,872**	**0,915**	0,906	**0,785**	**0,911**	0,877
MixNet-XL	1	#3	1	Acurác.	0,789	0,830	0,882	0,882	0,696	0,882	**0,882**
ViT-Base	1	#2	1	Acurác.	**0,855**	0,855	0,886	**0,916**	0,696	0,901	0,866

Ambas avaliações são comparadas quanto as respectivas curvas ROC para cada modelo, conforme Figuras 6.8a e 6.8b. Nestas curvas, observa-se uma maior taxa de verdadeiro positivo para os melhores modelos em faixas com menor valor de falso positivo na avaliação de sensibilidade (Figura 6.8a). Outro detalhe é o maior valor de AUC estar presente no melhor modelo, o RegNetY-3.2GF.

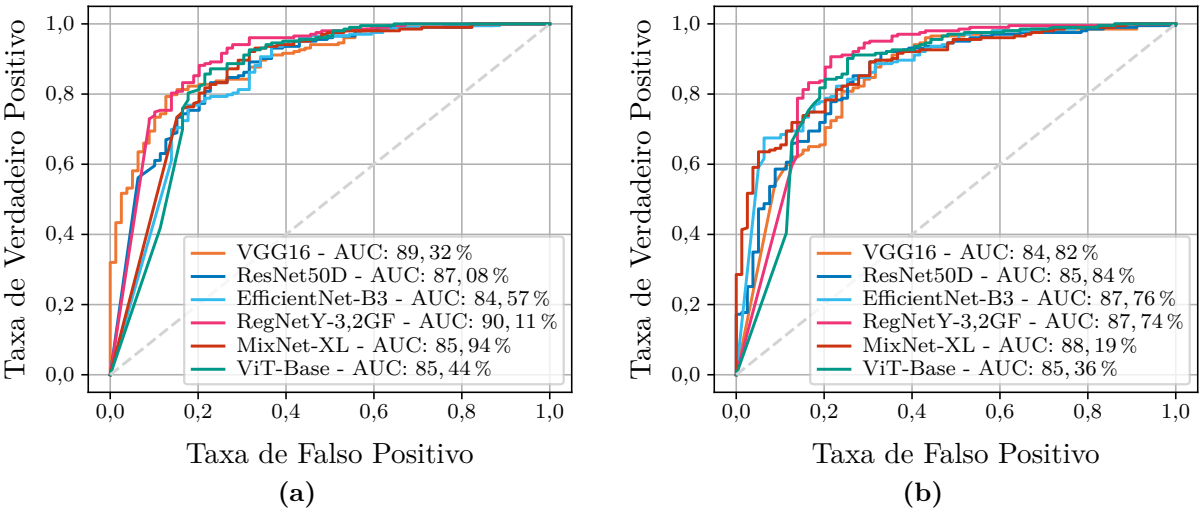


Figura 6.8. Curvas ROC dos melhores modelos para classificação BI-RADS #6. Em que (a) são os modelos para avaliação com sensibilidade e (b) para avaliação com especificidade.

Para verificar o desempenho dos modelos para dados não relacionados com os de treinamento, utiliza-se os resultados da Tabela 6.19. Na qual, a classificação #6 pôde apenas ser testada no banco de dados BrEaST. Em geral, o ViT apresenta métricas semelhantes as arquiteturas CNN, estas que tiveram dificuldade na identificação do BI-RADS 2 do restante. Destaca-se que, para os três testes analisados, o amostrador aleatório normal (Amost.-2 ou 4) treinou modelos melhores que o ponderado para a arquitetura ViT. Quanto ao tipo de TL, o ajuste fino completo se manteve como o melhor; contudo, o tipo em que apenas o classificador (TL-2) é refinado foi melhor para as imagens do BrEaST.

Tabela 6.19. Melhores resultados da classificação #6 entre os modelos propostos para o banco de dados BrEaST.

					BrEaST							
Aval.	Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
Sensib.	VGG16	1	#3	2	Acurác.	0,881	0,881	0,893	0,982	0,133	0,936	**0,795**
	ResNet50D	1	#3	2	Acurác.	0,869	0,869	**0,906**	0,950	**0,267**	0,927	0,774
	EfficientNet-B3	1	#3	4	*Acurác.	0,853	0,853	0,897	0,941	0,200	0,919	0,735
	RegNetY-3.2GF	1	#2	3	Último	0,881	0,881	0,903	0,968	0,233	0,935	0,711
	MixNet-XL	1	#1	3	Último	**0,893**	**0,893**	0,895	**0,996**	0,133	**0,942**	0,603
	ViT-Base	2	#2	2	**Custo	0,869	0,869	0,886	0,977	0,067	0,929	0,627
Especif.	VGG16	1	#3	2	Acurác.	0,881	0,881	0,893	0,982	0,133	0,936	**0,795**
	ResNet50D	1	#3	2	Acurác.	0,869	0,869	0,906	0,950	0,267	0,927	0,774
	EfficientNet-B3	1	#3	4	*Acurác.	0,853	0,853	0,897	0,941	0,200	0,919	0,735
	RegNetY-3.2GF	1	#3	2	*Acurác.	0,857	0,857	**0,919**	0,919	**0,400**	0,919	0,786
	MixNet-XL	1	#1	3	*Acurác.	**0,889**	**0,889**	0,894	**0,991**	0,133	**0,940**	0,611
	ViT-Base	1	#2	2	*Acurác.	0,821	0,821	0,908	0,887	0,333	0,897	0,708

*: Acurác. e Custo com mesmo desempenho. **: Custo, Acurácia e Último com mesmo desempenho.

Comparando as arquiteturas no ranqueamento com a sensibilidade (Tabela 6.19), observa-se uma dificuldade maior do ViT em identificar BI-RADS 2 do restante, apresentado metade do desempenho que o pior caso entre as arquiteturas CNN. No geral, o ViT-Base foi a pior arquitetura em ambas as avaliações. Dessa forma, se mantém a recomendação do capítulo anterior em utilizar a VGG16 para identificação do BI-RADS 2 das categorias 3, 4 e 5.

6.7 Avaliação do uso do pré-processamento #4 no ViT

O pré-processamento #4 se refere àquele em que a entrada do modelo é a imagem original. Nesse capítulo, os melhores modelos para cada classificação não utilizaram esse pré-processamento, pois o desempenho deste foi muito inferior aos melhores resultados dos modelos ViT treinados. Tal achado difere do observado anteriormente para os modelos CNN.

A Tabela 6.20 apresenta os modelos com melhor desempenho obtidos com o pré-processamento #4, bem como os melhores de cada classificação binária. Foram descartados os resultados cuja sensibilidade ou especificidade foi igual ou inferior a 10 %.

A partir da análise dos resultados da Tabela 6.20, verifica-se uma grande perda de

Tabela 6.20. Comparação entre melhores resultados de cada classificação binária com os que utilizaram o pré-processamento #4 (imagem original).

Classif.	Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Acurác. (binária)	Prec.	Sensib.	Especif.	F1	AUC
#1	ViT-Base	1	#3	2	Acurác.	**0,897**	**0,897**	**0,849**	**0,820**	**0,933**	**0,834**	**0,930**
	ViT-Base	3	#4	1	Último	0,589	0,589	0,400	0,607	0,580	0,482	0,584
#3	ViT-Base	1	#3	1	Último	**0,991**	**0,991**	**1,000**	**0,952**	**1,000**	**0,976**	**0,999**
	ViT-Base	1	#4	1	Acurác.	0,844	0,844	0,591	0,619	0,898	0,605	0,731
#4	ViT-Base	1	#3	1	Último	**0,922**	**0,922**	**0,920**	**0,906**	**0,935**	**0,913**	**0,941**
	ViT-Base	1	#4	1	Último	0,610	0,610	0,581	0,480	0,716	0,526	0,647
#5	ViT-Base	1	#2	1	Último	**0,957**	**0,957**	**0,750**	**0,500**	0,989	**0,600**	**0,903**
	ViT-Base	1	#4	1	Acurác.	0,940	0,940	0,667	0,111	**0,996**	0,190	0,623
#6	ViT-Base	1	#2	2	Último	**0,855**	**0,855**	**0,875**	**0,931**	**0,658**	**0,902**	**0,854**
	ViT-Base	1	#4	2	Custo	0,741	0,741	0,778	0,897	0,342	0,833	0,635

desempenho entre os modelos na classificação #1, #4, e #6. Na qual, o modelo da classificação #6 foi decidido após descarte de quatro modelos onde a especificidade era entre 0-3 %. Na classificação #3 é esperado que as métricas apresentem valores mais altos, por se tratar de uma classificação mais simples. Portanto, o resultado obtido nesse caso é considerado muito baixo em comparação ao melhor modelo. Na classificação #5, algumas métricas aproximam-se das do melhor modelo; porém, as mais importantes a serem avaliadas são a sensibilidade e o F1-*Score*, ambas com valores consideravelmente inferiores em relação ao melhor resultado.

A Tabela 6.21 apresenta a comparação entre o melhor modelo geral e aquele que utilizou o pré-processamento #4 para classificação multiclasse do BI-RADS 2, 3, 4, 5. Na qual, o mesmo baixo desempenho visto anteriormente está presente nessa classificação.

Tabela 6.21. Comparação entre melhor resultado da classificação BI-RADS multiclasse (#2) com aquele que utilizou o pré-processamento #4 (imagem original).

Modelo	TL	Pré-proc.	Amost.	Melhor modelo	Acurác. (macro)	Prec. (macro)	Sensib. (macro)	Especif. (macro)	F1 (macro)	Acurác. (micro)	AUC
ViT-Base	1	#3	2	Acurác.	**0,874**	**0,709**	**0,712**	**0,911**	**0,710**	**0,748**	**0,881**
ViT-Base	1	#4	1	Último	0,695	0,359	0,356	0,779	0,357	0,390	0,637

Como essa diferença entre o pré-processamento #4 do restante foi observado apenas no ViT-Base, certas hipóteses foram levantadas. A principal é que, como a arquitetura ViT realiza uma análise global da imagem, as relações encontradas pelo *encoder transformer* entre as partes tem resultado distinto do mecanismo presente nas CNNs. Dado que os hiperparâmetros de treinamento foram refinados em experimentos com os pré-processamentos #2 e #3, um refinamento específico é necessário para que o uso da imagem original como entrada apresente melhores resultados.

7 Conclusão

O desenvolvimento de CADx é essencial para melhorar o desempenho dos exames de rastreamento com intuito de diagnóstico precoce de câncer de mama. Na população de mulheres mais jovens ($\leqslant$ 50 anos) é recomendado que esses exames sejam realizados por ultrassonografia, porém sua eficácia é ainda muito dependente da experiência dos profissionais que os realizam. Para melhorar esse cenário, foi realizada uma avaliação de diferentes modelos de redes neurais com objetivo de classificação de imagens da mama com lesões, tanto quanto à malignidade como por categorias BI-RADS.

A aplicação de redes neurais para classificação de imagens de ultrassom mamário é, contudo, um desafio. As arquiteturas recentes são formadas por várias camadas de extração de características e, respectiva, classificação. Essa abordagem necessita de volume considerável de dados, todavia, os conjuntos publicamente disponíveis são, em geral, de tamanho reduzido. Esse estudo avaliou cinco arquiteturas CNN e uma ViT utilizando *transfer learning* e aumento de dados para seis classificações, tanto para maligno ou não quanto categorias definidas no BI-RADS. Além das variações de arquitetura, também foram implementadas quatro estratégias de pré-processamento e dois amostradores aleatórios, normal e ponderado. Para os modelos ViT foram comparados quatro níveis de *transfer learning*.

Para classificação quanto à malignidade, foi observado maior desempenho nas arquiteturas CNN desenvolvidas por NAS. A melhor foi a MixNet-XL, obtendo $88,3\,\%$ de acurácia, $82,0\,\%$ de F1-*Score*, e $91,5\,\%$ de AUC. No caso da classificação BI-RADS multiclasse, categorias 2 a 5 (total de 4), a ViT-Base apresentou maior desempenho do que as arquiteturas CNN, apresentando uma acurácia balanceada de $87,4\,\%$, F1-*Score* médio por classe de $71,0\,\%$, e AUC de $88,1\,\%$. Pode-se destacar que em ambos os casos, o pré-processamento que gerava uma imagem de 3 canais ao redor da lesão, original de (62), foi o melhor. No caso das CNNs, o uso da imagem original como entrada apresentou resultados promissores, como uma perda de $6,0\,\%$ de F1-*Score* para classificação quanto à malignidade utilizando a RegNetY-3.2GF. Esta que foi a segunda melhor arquitetura, diferença de $\approx 1\,\%$ da MixNet-XL nessa classificação.

Com intuito de melhorar o desempenho da classificação BI-RADS aplicando simplificações, quatro variações foram avaliadas. Na primeira apenas amostras de BI-RADS 2 e

5 foram utilizadas para treinamento, validação e teste, o que propocionou métricas "perfeitas" para várias das arquiteturas CNN. Na segunda, os rótulos benignos do BI-RADS foram agrupados em uma classe distinta dos rótulos malignos, (2, 3) e (4, 5) respectivamente, resultando na ViT como melhor desempenho em dois dos três testes avaliados. As duas últimas variações isolaram a categoria com maior probabilidade de maligno (5) e benigno (2) em uma classe enquanto o restante foi agrupado em outra, respectivamente classificações #5 e #6. Nessas classificações se observou a dificuldade de diferenciação entre categorias semelhantes, como 5 e 4, mesmo após junção com categorias bem diferenciáveis entre si. Esse comportamento foi encontrado após análise da matriz de confusão do melhor modelo para identificação do BI-RADS multiclasse.

O uso da ViT, quando utilizada para classificação de lesões, evidenciou melhor desempenho no caso mais complexo (multiclasse) e em algumas variações binárias do BI-RADS, como quando se isola o BI-RADS 5. Porém, isso não foi observado quando se utiliza a imagem original como entrada sem processamentos adicionais. Nesse tipo de imagem ocorreu uma perda expressiva de desempenho quando comparado ao melhor caso, este que utilizou a lesão segmentada para classificação. A principal hipótese levantada é a necessidade de um refinamento específico dos hiperparâmetros por terem sido treinados com eles para todos os casos da ViT.

7.1 Futuros trabalhos

Para trabalhos futuros é proposta uma análise das imagens de ultrassom original aplicando hiperparâmetros refinados para arquiteturas que utilizam *transformers*, como a ViT. No mesmo tipo de arquitetura, este estudo realizou apenas a avaliação do *transfer learning* aplicado as camadas que não eram o *encoder transformer*. Assim, um tópico de avaliação é o uso do ajuste fino do bloco *transformer* no ViT.

Os modelos mais complexos de estudos recentes mostraram maior desempenho na classificação de imagens de domínio generalista, como o EVA-02 no ImageNet1K (19, 100). Considerando que a aplicação clínica busca sempre o melhor desempenho e a limitação de *hardware* não seja um obstáculo, sua implementação apresenta potencial de melhorar o desempenho na classificação de imagens de ultrassom mamário, principalmente no caso do BI-RADS multiclasse.

Adicionalmente, o uso de redes neurais em ambientes clínicos precisa justificar ao radiologista a classificação gerada durante a utilização do sistema. Algumas formas a serem estudadas são: a apresentação do contorno da lesão identificada ou o emprego de IA Explicável (xAI), como mapas de calor sobrepostos à imagem do exame.

Lista de Referências

1 Institute for Health Metrics and Evaluation (IHME). *GBD Compare Data Visualization*. Seattle, WA, 2024. Disponível em: ⟨http://vizhub.healthdata.org/gbd-compare⟩. Acesso em: 12 dez. 2024.

2 INSTITUTO NACIONAL DE CÂNCER (INCA). *Estimativa 2023: incidência de câncer no Brasil*. Rio de Janeiro, RJ: Instituto Nacional De Câncer, 2023. ISBN 978-65-88517-09-3.

3 Instituto Nacional de Câncer (INCA). *Estatísticas de câncer*. 2023. Disponível em: ⟨https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/⟩. Acesso em: 02 jan. 2025.

4 Instituto Nacional de Câncer (INCA). *Atlas Online de mortalidade por câncer no Brasil 1979-2023*. Ministério da Saúde, Instituto Nacional do Câncer, 2023. Disponível em: ⟨https://www.inca.gov.br/MortalidadeWeb⟩. Acesso em: 08 maio 2025.

5 BRASIL. INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA (INCA). *Diretrizes para a detecção precoce do câncer de mama no Brasil*. Rio de Janeiro, RJ. 166 p.

6 Instituto Nacional de Câncer. *Controle do câncer de mama no Brasil*. Rio de Janeiro, RJ: Instituto Nacional De Câncer, 2024. ISBN 978-65-88517-69-7.

7 WILD, C.; WEIDERPASS, E.; STEWART, B. W. *World cancer report: cancer research for cancer prevention*. [S.l.]: IARC Press, 2020. 25-27 p.

8 CÂNCER, I. N. d. *Parâmetros técnicos para detecção precoce do câncer de mama*. Rio de Janeiro, RJ: Instituto Nacional De Câncer, 2023. ISBN 978-65-88517-07-9.

9 COOP, P.; COWLING, C.; LAWSON, C. Tomosynthesis as a screening tool for breast cancer: A systematic review. *Radiography*, v. 22, n. 3, p. e190–e195, 2016. ISSN 1078-8174.

10 QI, X. et al. Automated diagnosis of breast ultrasonography images using deep neural networks. *Medical image analysis*, Elsevier, v. 52, p. 185–198, 2019.

11 ILESANMI, A. E.; CHAUMRATTANAKUL, U.; MAKHANOV, S. S. Methods for the segmentation and classification of breast ultrasound images: a review. *Journal of Ultrasound*, Springer, v. 24, n. 4, p. 367–382, 2021.

12 The Royal College of Radiologists. *AI Deployment Fundamentals for Medical Imaging*. London, UK, 2024. 26 p. Disponível em: ⟨https://www.rcr.ac.uk/our-services/all-our-publications/clinical-radiology-publications/ai-deployment-fundamentals-for-medical-imaging⟩. Acesso em: 31 ago. 2024.

13 The Royal College of Radiologists. *Guidance on Auto-Contouring in Radiotherapy*. London, UK. 36 p. Disponível em: ⟨https://www.rcr.ac.uk/our-services/all-our-publications/clinical-oncology-publications/auto-contouring-in-radiotherapy⟩. Acesso em: 31 ago. 2024.

14 NAJJAR, R. Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*, v. 13, n. 17, p. 2760, 2023. ISSN 2075-4418. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

15 SPAK, D. A. et al. BI-RADS® fifth edition: A summary of changes. *Diagnostic and Interventional Imaging*, v. 98, n. 3, p. 179–190, 2017. ISSN 2211-5684.

16 QIAN, X. et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nature Biomedical Engineering*, v. 5, n. 6, p. 522–532, jun. 2021. ISSN 2157-846X. Publisher: Nature Publishing Group.

17 KIM, H. E. et al. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, v. 22, n. 1, p. 69, abr. 2022. ISSN 1471-2342.

18 BYRA, M. et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Medical Physics*, v. 46, n. 2, p. 746–755, fev. 2019. ISSN 0094-2405.

19 FANG, Y. et al. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, v. 149, p. 105171, set. 2024. ISSN 0262-8856.

20 LUCENA, C. *Mastologia: Do Diagnóstico ao Tratamento*. Rio de Janeiro, RJ: MedBook, 2023. ISBN 978-65-5783-095-6.

21 J, F. et al. *Global Cancer Observatory: Cancer Today*. Lyon, France: International Agency for Research on Cancer, 2022. Disponível em: ⟨https://gco.iarc.who.int/today⟩. Acesso em: 07 jan. 2025.

22 IBGE. *Censo 2022: População por idade e sexo*. Brasil: [s.n.], 2023. Disponível em: ⟨https://censo2022.ibge.gov.br/panorama/⟩. Acesso em: 07 jan. 2025.

23 SIB/ANS/MS. *Informações de Saúde Suplementar*. Brasil: [s.n.]. Disponível em: ⟨https://www.ans.gov.br/anstabnet/cgi-bin/dh?dados/tabnet_tx.def⟩. Acesso em: 07 jan. 2025.

24 STANISLAVSKY, A. Intraductal papilloma. In: *Radiopaedia.org*. Radiopaedia.org, 2014. Disponível em: ⟨http://radiopaedia.org/cases/intraductal-papilloma-3⟩. Acesso em: 10 fev. 2025.

25 ASHRAF, A. Ductal carcinoma in situ (DCIS). In: *Radiopaedia.org*. Radiopaedia.org, 2023. Disponível em: ⟨https://radiopaedia.org/cases/ductal-carcinoma-in-situ-dcis-2⟩. Acesso em: 10 fev. 2025.

26 KNIPE, H.; RADSWIKI, T. Papillary carcinoma of the breast. In: *Radiopaedia.org*. Radiopaedia.org, 2011. Disponível em: ⟨http://radiopaedia.org/articles/12794⟩. Acesso em: 02 mar. 2025.

27 AWAL, S. Fibroadenoma of the breast. In: *Radiopaedia.org*. Radiopaedia.org, 2020. Disponível em: ⟨http://radiopaedia.org/cases/fibroadenoma-of-the-breast-2⟩. Acesso em: 10 fev. 2025.

28 HACKING, C.; FAROUK, A. Phyllodes tumour. In: *Radiopaedia.org*. Radiopaedia.org, 2017. Disponível em: ⟨http://radiopaedia.org/cases/phyllodes-tumour-3⟩. Acesso em: 02 mar. 2025.

29 SIEGEL, R. L. et al. Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, v. 71, n. 1, p. 7–33, 2021.

30 OBAID, A. A. Invasive ductal carcinoma of the breast. In: *Radiopaedia.org*. Radiopaedia.org, 2025. Disponível em: ⟨https://radiopaedia.org/cases/203579⟩. Acesso em: 02 mar. 2025.

31 JIN, T.; BABU, V. Invasive lobular breast carcinoma. In: *Radiopaedia.org*. Radiopaedia.org, 2012. Disponível em: ⟨http://radiopaedia.org/cases/invasive-lobular-breast-carcinoma⟩. Acesso em: 02 mar. 2025.

32 BELL, D.; AHIR, N. Breast carcinoma (multicentric multifocal in mammary Paget disease). In: *Radiopaedia.org*. Radiopaedia.org, 2017. Disponível em: ⟨https://radiopaedia.org/cases/50966⟩. Acesso em: 10 fev. 2025.

33 SOOD, R. et al. Ultrasound for Breast Cancer Detection Globally: A Systematic Review and Meta-Analysis. *Journal of Global Oncology*, n. 5, p. 1–17, ago. 2019. Publisher: Wolters Kluwer.

34 THIGPEN, D.; KAPPLER, A.; BREM, R. The Role of Ultrasound in Screening Dense Breasts—A Review of the Literature and Practical Solutions for Implementation. *Diagnostics*, v. 8, n. 1, p. 20, mar. 2018. ISSN 2075-4418. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

35 REBOLJ, M. et al. Addition of ultrasound to mammography in the case of dense breast tissue: systematic review and meta-analysis. *British Journal of Cancer*, v. 118, n. 12, p. 1559–1570, jun. 2018. ISSN 1532-1827. Publisher: Nature Publishing Group.

36 RADIOLOGY, A. C. of et al. Breast imaging reporting and data system. *BI-RADS*, American College of Radiology, 2003.

37 TURING, A. M. Computing Machinery and Intelligence. In: EPSTEIN, R.; ROBERTS, G.; BEBER, G. (Ed.). *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Dordrecht: Springer Netherlands, 2009. p. 23–65. ISBN 978-1-4020-6710-5.

38 ROSENBLATT, F. Perceptron Simulation Experiments. *Proceedings of the IRE*, v. 48, n. 3, p. 301–309, mar. 1960. ISSN 0096-8390.

39 WEIZENBAUM, J. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco, USA: W. H. Freeman & Co., 1976. ISBN 978-0716704645.

40 MINSKY, M.; PAPERT, S. *Perceptrons: an introduction to computational geometry*. 2. print. with corr. ed. Cambridge/Mass.: The MIT Press, 1972. ISBN 978-0-262-13043-1 978-0-262-63022-1.

41 Computer-Based Medical Consultations: Mycin. [S.l.]: Elsevier, 1976. ISBN 978-0-444-00179-5.

42 RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, n. 6088, p. 533–536, out. 1986. ISSN 1476-4687. Publisher: Nature Publishing Group.

43 LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, nov. 1998. ISSN 00189219.

44 HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, v. 18, n. 7, p. 1527–1554, jul. 2006. ISSN 0899-7667, 1530-888X.

45 HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science*, v. 313, n. 5786, p. 504–507, jul. 2006. ISSN 0036-8075, 1095-9203.

46 KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: PEREIRA, F. et al. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2012. v. 25.

47 VASWANI, A. et al. Attention is All you Need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2017. v. 30.

48 DOSOVITSKIY, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv, 2020. Version Number: 2. Disponível em: ⟨https://arxiv.org/abs/2010.11929⟩. Acesso em: 25 fev. 2025.

49 LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, maio 2015. ISSN 1476-4687. Publisher: Nature Publishing Group.

50 CARRIERO, A. et al. Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements in Early 2024. *Diagnostics*, v. 14, n. 8, p. 848, jan. 2024. ISSN 2075-4418. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

51 LUO, L. et al. Deep Learning in Breast Cancer Imaging: A Decade of Progress and Future Directions. *IEEE Reviews in Biomedical Engineering*, v. 18, p. 130–151, 2025. ISSN 1941-1189. Conference Name: IEEE Reviews in Biomedical Engineering.

52 PI, Y. et al. Automated assessment of BI-RADS categories for ultrasound images using multi-scale neural networks with an order-constrained loss function. *Applied Intelligence*, v. 52, n. 11, p. 12943–12956, set. 2022. ISSN 1573-7497.

53 XING, J. et al. Using BI-RADS Stratifications as Auxiliary Information for Breast Masses Classification in Ultrasound Images. *IEEE Journal of Biomedical and Health Informatics*, v. 25, n. 6, p. 2058–2070, jun. 2021. ISSN 2168-2208. Conference Name: IEEE Journal of Biomedical and Health Informatics.

54 MO, Y. et al. HoVer-Trans: Anatomy-Aware HoVer-Transformer for ROI-Free Breast Cancer Diagnosis in Ultrasound Images. *IEEE Transactions on Medical Imaging*, v. 42, n. 6, p. 1696–1706, jun. 2023. ISSN 1558-254X. Conference Name: IEEE Transactions on Medical Imaging.

55 AL-DHABYANI, W. et al. Dataset of breast ultrasound images. *Data in Brief*, v. 28, p. 104863, 2020. ISSN 2352-3409.

56 FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. ISSN 0167-8655.

57 SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, jul. 2009. ISSN 0306-4573. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0306457309000259⟩.

58 CHINCHOR, N. Muc-4 evaluation metrics. In: *Proceedings of the 4th conference on Message understanding - MUC4 '92*. McLean, Virginia: Association for Computational Linguistics, 1992. p. 22. ISBN 978-1-55860-273-1.

59 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. MIT Press, 2016. ISBN 9780262035613. Disponível em: ⟨http://www.deeplearningbook.org⟩.

60 PIOTRZKOWSKA, H. et al. Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. *Medical Physics*, v. 44, n. 11, p. 6105–6109, nov. 2017. ISSN 0094-2405, 2473-4209.

61 PAWŁOWSKA, A. et al. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, Nature Publishing Group, v. 11, n. 1, p. 148, jan. 2024. ISSN 2052-4463.

62 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. ISSN 2473-4209.

63 PAWŁOWSKA, A.; KARWAT, P.; ŻOŁEK, N. Letter to the editor. re: "[dataset of breast ultrasound images by w. al-dhabyani, m. gomaa, h. khaled & a. fahmy, data in brief, 2020, 28, 104863]". *Data in Brief*, v. 48, p. 109247, jun. 2023. ISSN 23523409.

64 WIGHTMAN, R. et al. *PyTorch Image Models (TIMM)*. [S.l.]: GitHub, 2023. ⟨https://github.com/huggingface/pytorch-image-models⟩. Acesso em: 17 mar. 2023.

65 WIGHTMAN, R.; TOUVRON, H.; JEGOU, H. Resnet strikes back: An improved training procedure in timm. In: *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*. [S.l.: s.n.], 2021.

66 RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, Springer, v. 115, p. 211–252, 2015.

67 WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. *Journal of Big Data*, v. 3, n. 1, p. 9, maio 2016. ISSN 2196-1115.

68 GONZALEZ, R. C. Deep convolutional neural networks [lecture notes]. *IEEE Signal Processing Magazine*, v. 35, n. 6, p. 79–87, nov. 2018. ISSN 1053-5888, 1558-0792.

69 LENAIL, A. Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, v. 4, n. 33, p. 747, jan. 2019. ISSN 2475-9066.

70 CYBENKO, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, v. 2, n. 4, p. 303–314, dez. 1989. ISSN 0932-4194, 1435-568X.

71 SNYMAN, J. A.; WILKE, D. N. *Practical Mathematical Optimization*. Cham: Springer International Publishing, 2018. v. 133. (Springer Optimization and Its Applications, v. 133). ISBN 978-3-319-77585-2.

72 BYRA, M. et al. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomedical Signal Processing and Control*, v. 61, p. 102027, ago. 2020. ISSN 17468094.

73 WANG, C. et al. Convolutional embedding makes hierarchical vision transformer stronger. In: _____. *Computer Vision – ECCV 2022*. Cham: Springer Nature Switzerland, 2022. (Lecture Notes in Computer Science, v. 13680), p. 739–756. ISBN 978-3-031-20043-4.

74 SCHUHMANN, C. et al. LAION-5b: An open large-scale dataset for training next generation image-text models. In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. [S.l.: s.n.], 2022.

75 JESUS, G. *Explainable Artificial Intelligence Model For Mammogram Breast Cancer Classifiers*. Dissertação (Master's Thesis in Biomedical Engineering;) — University of Brasilia;, Gama, Brazil, nov. 2023.

76 MOON, W. K. et al. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer Methods and Programs in Biomedicine*, v. 190, p. 105361, jul. 2020. ISSN 0169-2607.

77 PODDA, A. S. et al. Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images. *Journal of Computational Science*, v. 63, p. 101816, set. 2022. ISSN 1877-7503.

78 ZHANG, G. et al. Sha-mtl: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification. *International Journal of Computer Assisted Radiology and Surgery*, v. 16, 07 2021.

79 LIAO, W.-X. et al. Automatic Identification of Breast Ultrasound Image Based on Supervised Block-Based Region Segmentation Algorithm and Features Combination Migration Deep Learning Model. *IEEE Journal of Biomedical and Health Informatics*, v. 24, n. 4, p. 984–993, abr. 2020. ISSN 2168-2208. Conference Name: IEEE Journal of Biomedical and Health Informatics. Disponível em: ⟨https://ieeexplore.ieee.org/document/8936948⟩.

80 WIGHTMAN, R. et al. *TIMM ImageNet-1k validation results*. [S.l.]: Hugging Face, 2023. ⟨https://huggingface.co/docs/timm/v0.9.16/en/results⟩. Acesso em: 28 maio 2023.

81 SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

82 HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

83 HE, T. et al. Bag of tricks for image classification with convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 558–567, 2018.

84 TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *International conference on machine learning*. [S.l.], 2019. p. 6105–6114.

85 TAN, M. et al. Mnasnet: Platform-aware neural architecture search for mobile. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 2820–2828.

86 RADOSAVOVIC, I. et al. Designing network design spaces. In: *CVPR*. [S.l.: s.n.], 2020.

87 HU, J.; SHEN, L.; SUN, G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 7132–7141.

88 TAN, M.; LE, Q. V. *MixConv: Mixed Depthwise Convolutional Kernels*. 2019. Disponível em: ⟨https://arxiv.org/abs/1907.09595⟩. Acesso em: 22 nov. 2023.

89 RADFORD, A. et al. Learning transferable visual models from natural language supervision. In: *ICML*. [S.l.: s.n.], 2021.

90 ILHARCO, G. et al. *OpenCLIP*. Zenodo, 2021. If you use this software, please cite it as below. Disponível em: ⟨https://doi.org/10.5281/zenodo.5143773⟩. Acesso em: 13 nov. 2024.

91 ZHANG, Y. et al. Contrastive learning of medical visual representations from paired images and text. In: LIPTON, Z. et al. (Ed.). *Proceedings of the 7th Machine Learning for Healthcare Conference*. [S.l.]: PMLR, 2022. (Proceedings of Machine Learning Research, v. 182), p. 2–25.

92 CHERTI, M. et al. Reproducible scaling laws for contrastive language-image learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 2818–2829.

93 SUTSKEVER, I. et al. On the importance of initialization and momentum in deep learning. In: DASGUPTA, S.; MCALLESTER, D. (Ed.). *Proceedings of the 30th International Conference on Machine Learning*. Atlanta, Georgia, USA: PMLR, 2013. (Proceedings of Machine Learning Research, 3), p. 1139–1147.

94 KINGMA, D. P.; BA, J. *Adam: A Method for Stochastic Optimization*. 2017. Disponível em: ⟨https://arxiv.org/abs/1412.6980⟩. Acesso em: 09 dez. 2023.

95 LOSHCHILOV, I.; HUTTER, F. *Decoupled Weight Decay Regularization*. 2019. Disponível em: ⟨https://arxiv.org/abs/1711.05101⟩. Acesso em: 18 abr. 2024.

96 GRAVES, A. *Generating Sequences With Recurrent Neural Networks*. arXiv, 2014. ArXiv:1308.0850 [cs]. Disponível em: ⟨http://arxiv.org/abs/1308.0850⟩. Acesso em: 18 abr. 2024.

97 DETTMERS, T. et al. 8-bit optimizers via block-wise quantization. *CoRR*, abs/2110.02861, 2021. Disponível em: ⟨https://arxiv.org/abs/2110.02861⟩. Acesso em: 15 maio 2024.

98 DUDA, R. O.; STORK, D. G.; HART, P. E. *Pattern classification*. 2nd ed. ed. New York: Wiley, 2001. ISBN 978-0-471-05669-0.

99 BOBOWICZ, M. et al. Segmentation-based bi-rads ensemble classification of breast tumours in ultrasound images. *International Journal of Medical Informatics*, v. 189, p. 105522, set. 2024. ISSN 1386-5056.

100 SUN, Q. et al. *EVA-CLIP: Improved Training Techniques for CLIP at Scale*. 2023. Disponível em: ⟨https://arxiv.org/abs/2303.15389⟩.

101 CÂNCER, I. N. d. *Parâmetros técnicos para detecção precoce do câncer de mama*. Rio de Janeiro, RJ: Instituto Nacional De Câncer, 2021. ISBN 978-65-88517-25-3.

Anexo A

Parâmetros técnicos para detecção precoce do câncer de mama

Esta seção apresenta a memória do cálculo utilizado para definição dos parâmetros técnicos definidos pelo INCA (8). No Quadro 7.1, estão sintetizados os cálculos realizados para os parâmetros de detecção precoce na população de mulheres sintomáticas com câncer de mama no Brasil.

Quadro 7.1. Resumo do cálculo para construção dos parâmetros para estimação de procedimentos de ultrassonografia mamária na população de mulheres sintomáticas do Brasil. Fonte: (8).

Faixa etária	Parâmetro	Memória de cálculo
< 30 anos	$278,35\,\%$	$100\,\%$ das mulheres com sinais e sintomas suspeitos + $71,85\,\%$ do acompanhamento de mulheres com ultrassonografia sem alterações ou benigna + $106,52\,\%$ do acompanhamento de mulheres com nódulo e histopatológico benigno
> 30 anos ou mais	$207,03\,\%$	$80,18\,\%$ das mulheres com nódulo e mamografia sem alterações ou alteração benigna + $11,70\,\%$ das mulheres com outros sintomas + $4,66\,\%$ de acompanhamento de mulheres com descarga papilar e imagem benigna + $5,67\,\%$ de acompanhamento de mulheres com pele alterada + $104,82\,\%$ de acompanhamento de mulheres com nódulo e histopatológico benigno

Para o rastreamento de mulheres na faixa recomendada, 50 a 69 anos, o INCA utilizou o mesmo método definido para o triênio 2021–2023. Ou seja, o parâmetro $3,5\,\%$ é calculado pela necessidade total estimada de ultrassonografia mamária ($7\,\%$) multiplicado a $50\,\%$ da população na faixa recomendada a ser rastreada pelo SUS, definido devido a recomendação ser bienal (101).

A necessidade total estimada de ultrassonografia na população de 50 a 69 anos ($7\,\%$) é calculada da seguinte forma (101): $6,7\,\%$, avaliação do BI-RADS 0 a partir dos resultados da mamografia de rastreamento, somado a $0,284\,\%$, resultado de BI-RADS 0 no acompanhamento do BI-RADS 3 pós-mamografia diagnóstica $[15,8\,\% * 1,8\,\% = 0,284\,\%]$.