

**INSTITUTO
FEDERAL**
Brasília

Instituto Federal de Brasília
Bacharelado em Ciência da Computação
Campus Taguatinga

**COMPARAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA CLASSIFICAÇÃO
DE DENSIDADE DE MAMA EM MAMOGRAFIAS DIGITAIS COMPLETAS**

Por

MATHEUS LOIOLA PINTO CURADO SILVA

Trabalho de Graduação

BRASÍLIA/2025

Matheus Loiola Pinto Curado Silva

**COMPARAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA
CLASSIFICAÇÃO DE DENSIDADE DE MAMA EM MAMOGRAFIAS
DIGITAIS COMPLETAS**

Trabalho apresentado ao Programa de Bacharelado em Ciência da Computação do Departamento de Computação da Instituto Federal de Brasília como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Dr. Fabiano Cavalcanti Fernandes

BRASÍLIA
2025

Trabalho de Graduação apresentado por **Matheus Loiola Pinto Curado Silva** ao programa de Graduação em Ciência da Computação do Instituto Federal de Brasília, sob o título **Comparação de redes neurais convolucionais para classificação de densidade de mama em mamografias digitais completas**, orientado pelo **Prof. Dr. Fabiano Cavalcanti Fernandes** e aprovado pela banca examinadora formada pelos professores:

Prof. Dr. Fabiano Cavalcanti Fernandes
Departamento de Computação/IFB

Prof. Dr. Raimundo C. S. Vasconcelos
Departamento de Computação/IFB

Prof. Me. Thiago Batista Amorim
Departamento de Computação/IFB

BRASÍLIA
2025

Dedico este trabalho a Deus, o qual esteve comigo em todos os momentos. À minha querida família, que esteve ao meu lado nos desafios e conquistas, oferecendo apoio, palavras de incentivo e amor inabalável. Ao meu orientador, pelos valiosos conselhos, oportunidades, amizade e compreensão.

Resumo

Em 2022, o câncer de mama foi o mais prevalente entre as mulheres com 94.728 casos de incidência e 22.189 mortes registradas no Brasil. A densidade mamária é um importante bioindicador de risco para esse tipo de tumor, sendo avaliada qualitativamente por radiologistas em quatro categorias conforme definido pelo BI-RADS®. Porém, existe uma grande variabilidade entre os observadores na avaliação dessas categorias. O uso de aprendizagem profunda tem tido um impacto significativo na área médica, com inúmeras arquiteturas de redes neurais convolucionais surgindo para melhorar a precisão e a consistência dos diagnósticos realizados pelos radiologistas. Assim, este trabalho tem como objetivo adaptar um modelo de rede neural convolucional para a classificação da densidade mamária em imagens de mamografias digitais. Os resultados obtidos demonstram que a abordagem multiclasse alcançou uma acurácia de 79,3%, enquanto a abordagem inédita com modelos binários independentes obteve 76,9%, alinhando-se com o estado da arte na literatura. Esses resultados evidenciam o potencial da aplicação de redes neurais convolucionais para a classificação da densidade mamária.

Palavras-chave: Mamografia Digital Completa, Densidade de Mama, BI-RADS, Rede Neural Convolutacional, Classificação

Abstract

In 2022 breast cancer was the most prevalent cancer among women, with 94,728 incidence cases and 22,189 recorded deaths in Brazil. Breast density is an important risk biomarker for this type of tumor, being qualitatively assessed by radiologists in four categories as defined by BI-RADS®. However, there is significant variability among observers in evaluating these categories. The use of deep learning has had a significant impact in the medical field, with numerous convolutional neural network architectures emerging to improve the accuracy and consistency of diagnoses made by radiologists. Therefore, this work aims to adapt a convolutional neural network model for the classification of breast density in digital mammography images. The obtained results demonstrate that the multiclass approach achieved an accuracy of 79.3%, while the novel approach with independent binary models reached 76.9%, aligning with the state of the art in the literature. These results highlight the potential of applying convolutional neural networks for breast density classification.

Keywords: Full-field Digital Mammography, Breast Density, BI-RADS, Convolutional Neural Network, Classification

Lista de Figuras

2.1	Mamografias de cada categoria de densidade mamária conforme o <i>Breast Imaging and Reporting Data System</i> (BI-RADS®) (5 ^a edição). Nota-se aumento progressivo do tecido fibroglandular da categoria A à D. Imagens de bases de dados públicas e anônimas.	22
2.2	Exemplo de uma camada convolucional. O <i>kernel</i> itera pela matriz aplicando o filtro na região selecionado, que resulta em uma nova matriz. Extraído e adaptado de PODAREANU et al. (2019).	25
2.3	Exemplo da função de ativação <i>Rectified Linear Unit</i> (<i>ReLU</i>), que transforma os valores negativos para zero e não modifica o restante. Extraído de OLANIPE-KUN et al. (2022).	26
2.4	Exemplo de <i>max-pooling</i> , no qual é retornado uma matriz com os valores máximos obtidos por cada <i>kernel</i> . Extraído de Papers With Code (2024).	26
2.5	Representação gráfica da função <i>softmax</i> . A função converte os valores de saída de uma rede neural em uma distribuição de probabilidade, normalizando-os no intervalo [0.0, 1.0]. Elaborado pelo autor.	27
2.6	Arquitetura da <i>AlexNet</i> . A partir dos grupos de convolução, foi possível treinar o modelo em duas <i>Graphical Processing Units</i> (GPUs) diferentes, com os vetores de características sendo concatenados ao final das convoluções. Essa abordagem representou um avanço significativo, dada a limitação computacional da época. Extraído de LI et al. (2020)	28
2.7	Arquitetura da <i>ResNet-18</i> . As setas curvadas demonstram as conexões residuais da rede, possibilitando a propagação dos gradientes para as camadas seguintes. Extraído e adaptado de WU et al. (2021).	29
2.8	Arquitetura da <i>DenseNet</i> . Possui blocos densos, onde os resultados de cada camada são concatenados nas seguintes. Extraído e adaptado de HUANG et al. (2018).	29
2.9	Arquitetura da <i>U-Net</i> . Nota-se que o nome da arquitetura é dado pelo seu formato, que se assemelha a letra U. Extraído e adaptado de RONNERBERGER; FISCHER; BROX (2015).	30
2.10	Exemplo da curva <i>Receiver Operating Characteristic</i> (ROC) e da <i>Area Under the Curve</i> (AUC). Quanto mais perto do canto superior esquerdo, melhor o modelo está desempenhando. A área em cinza é o valor do AUC. Extraído e adaptado de SAIA et al. (2021).	35
2.11	Validação cruzada em 5 dobras. A imagem representa um treinamento com cinco modelos, cada qual com uma dobra (<i>fold</i>) diferente de validação. Elaborado pelo autor.	36

3.1	Fluxograma da condução do Mapeamento Sistemático, que representa o processo de aquisição dos estudos para a revisão da literatura. Elaborado pelo autor.	41
3.2	<i>Framework</i> desenvolvido por DENG et al. (2020). Nota-se que, após cada camada de convolução, os valores são copiados para um bloco de atenção e armazenados em um vetor global. Isso permite ter informações do modelo durante todas as etapas de processamento. Extraído e adaptado de DENG et al. (2020).	43
3.3	Arquitetura da cGAN. A arquitetura de cima é a etapa do <i>encoder</i> , que irá produzir uma máscara binária da região da densidade da imagem de mamografia. A segunda arquitetura compreende a arquitetura do <i>decoder</i> , que irá utilizar camadas de convolução para extrair as características da máscara e retornar a densidade da imagem. Extraído e adaptado de SAFFARI et al. (2020).	47
4.1	Fluxograma da metodologia. As três primeiras etapas focam na infraestrutura e preparação dos dados, enquanto as duas últimas concentram-se na realização e avaliação dos experimentos.	50
4.2	Imagens resultantes após o processamento. Mesmo após a integração, ainda é notável uma diferença nos contrastes das imagens, devido à origem dos dados. . .	54
4.3	Distribuição das classes BI-RADS® após a integração das bases de dados obtidas.	55
4.4	Porcentagem da distribuição de cada classe na base de dados	55
4.5	Quantidade de imagens obtidas em cada base de dados, após aplicar os filtros e processamentos, em escala logarítmica.	56
4.6	Gráfico da contribuição de cada base de dados na quantidade e classe das imagens em escala logarítmica.	57
4.7	Proporção das classes em cada base de dados.	57
4.8	Arquitetura baseada em modelos binários. A imagem de entrada é processada por quatro modelos binários independentes, cada um treinado para identificar uma classe específica (positiva) enquanto agrupa as demais como negativas. As saídas dos modelos são então enviadas para uma camada de classificação, que aplica uma lógica para determinar a classe final da imagem.	58
5.1	Matrizes de confusão dos treinamentos realizados na primeira etapa de experimentos, mostrando a influência das bases de dados no desempenho do modelo.	67
5.2	Matrizes de confusão dos resultados obtidos na segunda etapa de experimentos, utilizando a nova lógica de classificação.	69
5.3	Matriz de confusão obtida ao utilizar apenas o melhor modelo dos <i>k-folds</i> , sem realizar a média das previsões.	70
5.4	Distribuição das classes no conjunto de dados balanceado utilizado na primeira etapa de experimentos multiclasse. O restante dos dados foram alocados para a etapa de teste.	71

5.5	Evolução da métrica F1 ao longo das épocas para diferentes <i>folds</i> de validação. Observa-se que a pontuação ainda apresenta tendência de crescimento, indicando que o treinamento pode ser estendido para um melhor ajuste do modelo.	73
5.6	Distribuição das classes no conjunto de dados reduzido utilizado na segunda etapa de experimentos multiclasse. O restante dos dados foram alocados para a etapa de teste.	74
5.7	Conjunto de dados de treinamento com 20% da quantidade total de imagens. . .	76
5.8	Exemplo de um lote de treinamento com o filtro magma aplicado nas imagens.	77
5.9	Conjunto de dados de treinamento com 90% da quantidade total de imagens. . .	79
5.10	Conjunto de dados de treinamento com 95% da quantidade total de imagens. . .	80
5.11	Matriz de confusão do melhor modelo multiclasse obtido após várias iterações de hiperparâmetros.	81

Lista de Tabelas

2.1	Matriz de Confusão para classificação binária. A partir dos valores, é possível calcular inúmeras métricas de validação, como acurácia, precisão, especificidade, entre outras.	32
2.2	Matriz de Confusão para múltiplas classes. A partir dos valores, é possível calcular inúmeras métricas de validação, como acurácia, precisão, especificidade, entre outras.	33
3.1	Lista das bases de dados eletrônicas nas quais a pesquisa foi realizada, organizadas com os <i>links</i> de acesso (Elaborado pelo autor).	38
3.2	Strings de busca em cada base de dados eletrônica. Foram utilizadas como filtros ou parâmetros para buscas personalizadas (Elaborado pelo autor).	39
3.3	Quantidade de estudos retornados por base de dados eletrônica após as pesquisas com as strings de busca (Elaborado pelo autor)	40
5.1	Configuração inicial dos hiperparâmetros utilizados na primeira etapa de experimentos.	67
5.2	Resultados das métricas dos treinamentos da primeira etapa de experimentos.	68
5.3	Resultados das métricas dos treinamentos realizados na segunda etapa de experimentos.	69
5.4	Resultados das métricas ao utilizar apenas o melhor modelo dos <i>k-folds</i>	70
5.5	Configuração inicial dos hiperparâmetros utilizados na primeira etapa de experimentos multiclasse.	72
5.6	Resultados da primeira etapa de experimentos multiclasse.	72
5.7	Configuração dos hiperparâmetros utilizados na segunda etapa de experimentos.	73
5.8	Resultados da segunda etapa de experimentos com modelos multiclasse.	74
5.9	Configuração dos hiperparâmetros utilizados no primeiro treinamento da terceira etapa de experimentos.	75
5.10	Configuração dos hiperparâmetros utilizados no segundo treinamento da terceira etapa de experimentos.	76
5.11	Configuração dos hiperparâmetros utilizados no terceiro treinamento da terceira etapa de experimentos.	77
5.12	Configuração dos hiperparâmetros utilizados no quarto treinamento da terceira etapa de experimentos.	78
5.13	Configuração dos hiperparâmetros utilizados no quinto treinamento da terceira etapa de experimentos.	79
5.14	Configuração dos hiperparâmetros utilizados no sexto treinamento da terceira etapa de experimentos.	80

5.15	Configuração dos hiperparâmetros utilizados no sétimo treinamento da terceira etapa de experimentos.	81
5.16	Resultados consolidados da terceira etapa de experimentos.	82
5.17	Comparação de diferentes modelos na classificação da densidade mamária nas quatro classes do BI-RADS®.	82

Lista de Acrônimos

ACAM	<i>Adaptive Channel Attention Module</i>	48
ACC	Acurácia	33
AdamW	<i>Adaptive Moment Estimation Weighted</i>	64
ASAM	<i>Adaptive Spatial Attention Module</i>	48
AUC	<i>Area Under the Curve</i>	34
BCE LOSS	<i>Binary Cross Entropy Loss</i>	
BI-RADS®	<i>Breast Imaging and Reporting Data System</i>	17
BMCD	<i>Breast Micro-Calcifications Dataset</i>	51
CAD	<i>Computer-Aided Diagnosis</i>	40
CAM	<i>Class Activation Maps</i>	47
CBIS-DDSM	<i>Curated Breast Imaging Subset of DDSM</i>	46
CC	Crânio-Caudal	43
CE	Critérios de Exclusão	38
CE LOSS	<i>Cross-Entropy Loss</i>	41
CI	Critérios de Inclusão	38
CLAHE	<i>Contrast Limited Adaptive Histogram Equalization</i>	45
CM-CNN	<i>Confusion Matrix Convolutional Neural Network</i>	42
CNN	<i>Convolutional Neural Network</i>	17
Cohen's κ	<i>Cohen's Kappa</i>	35
ConvNeXt	<i>Convolutional Neural Networks inspired by ViT</i>	60
DCNN	<i>Deep Convolutional Neural Network</i>	47
DDSM	<i>Digital Database for Screening Mammography</i>	43
DenseNet	<i>Densely Connected Convolutional Network</i>	60
DFS	<i>Depth First Search</i>	45
DICOM	<i>Digital Imaging and Communications in Medicine</i>	20
DL	<i>Deep Learning</i>	19
EfficientNet	<i>Efficient Convolutional Neural Network</i>	60
F1	<i>F1 Score</i>	33
FDA	<i>Food and Drug Administration</i>	18

FFDM	<i>Full-Field Digital Mammography</i>	19
FL LOSS	<i>Focal Loss</i>	44
FN	Falso Negativo	32
FP	Falso Positivo	32
GAP	<i>Global Average Pooling</i>	44
GAN	<i>Generative Adversarial Networks</i>	30
GPU	<i>Graphical Processing Unit</i>	28
IA	Inteligência Artificial	22
ILSVRC	<i>ImageNet Large Scale Visual Recognition Challenge</i>	27
ImageNet	<i>ImageNet</i>	27
InBreast	<i>Integrated Breast Dataset</i>	43
Kaggle	<i>Kaggle largest AI and ML community</i>	52
Linear κ	<i>Linear Kappa</i>	41
M-AUC	<i>Macro AUC</i>	46
MC	Matriz de Confusão	32
Mini-DDSM	<i>Mini Digital Database for Screening Mammography</i>	51
ML	<i>Machine Learning</i>	23
MLO	Médio-Oblíquo Lateral	43
MS	Mapeamento Sistemático	37
OAUC	<i>Overall AUC</i>	45
OR LOSS	<i>Ordinal Regression Loss</i>	41
PLN	Processamento de Linguagem Natural	24
PNG	<i>Portable Network Graphics</i>	46
QP	Questões de Pesquisa	37
ReLu	<i>Rectified Linear Unit</i>	25
ResNet	<i>Residual Network</i>	60
ROC	<i>Receiver Operating Characteristic</i>	34
RMSProp	<i>Root Mean Square Propagation</i>	46
RSNA	<i>Radiological Society of North America</i>	51
SE-Attention	<i>Squeeze and Excitation attention</i>	42
SGD	<i>Stochastic Gradient Descend</i>	45

SSE LOSS	<i>Sum of Squared Errors</i>	44
SVM	<i>Support Vector Machine</i>	23
VinDr	<i>VinDr-Mammo</i>	44
ViT	<i>Vision Transformer</i>	60
VN	Verdadeiro Negativo	32
VP	Verdadeiro Positivo	32
WCE LOSS	<i>Weighted Cross-Entropy Loss</i>	42
YOLOv5	<i>You Only Look Once Version 5</i>	44

Sumário

1	Introdução	17
1.1	Formulação do Problema	17
1.2	Motivação	18
1.3	Objetivos	18
1.3.1	Objetivo Geral	18
1.3.2	Objetivos Específicos	18
1.4	Descrição dos Capítulos	19
2	Referencial Teórico	20
2.1	Mamografia	20
2.1.1	DICOM	20
2.2	Densidade de Mama	21
2.3	Inteligência Artificial	22
2.4	Inteligência Artificial na Medicina	23
2.5	<i>Machine Learning</i>	23
2.6	<i>Deep Learning</i>	24
2.6.1	Redes Neurais Convolucionais	24
2.6.2	Arquitetura da Rede	25
2.6.3	Camada Convolucional	25
2.6.4	Camada de <i>Pooling</i>	26
2.6.5	Camada Totalmente Conectada	26
2.6.6	<i>Softmax</i>	27
2.7	Arquitetura de Redes Neurais Convolucionais	27
2.7.1	<i>AlexNet</i>	27
2.7.2	<i>EfficientNets</i>	28
2.7.3	<i>ResNets</i>	28
2.7.4	<i>DenseNets</i>	29
2.7.5	<i>U-Nets</i>	29
2.8	Redes Adversárias Generativas	30
2.9	<i>Data Augmentation</i>	31
2.10	Aprendizado por Transferência e Ajuste Fino	31
2.11	Integração dos Dados	31
2.12	Problemas de Classificação	32
2.13	Métricas de Avaliação de Desempenho	32
2.13.1	Matriz de Confusão	32
2.13.2	Acurácia	33

2.13.3	Precisão	33
2.13.4	<i>Recall</i>	33
2.13.5	Escore F1	33
2.13.6	Sensibilidade e Especificidade	34
2.13.7	Curva ROC e AUROC	34
2.13.8	<i>Cohen's Kappa</i>	35
2.14	Validação Cruzada	36
3	Revisão da Literatura	37
3.1	Questões de pesquisa	37
3.2	Estratégias de Busca	37
3.3	Critérios de Inclusão e Exclusão	38
3.4	Condução do Mapeamento Sistemático	40
3.5	Análise e Síntese dos Resultados	41
3.6	Conclusão dos resultados	49
4	Metodologia	50
4.1	Organização do trabalho	50
4.2	Ambiente de desenvolvimento	50
4.3	Coleta dos bancos de dados	51
4.3.1	BMCD	51
4.3.2	InBreast	52
4.3.3	MiniDDSM	52
4.3.4	RSNA	52
4.3.5	VinDr	52
4.4	Integração dos dados	53
4.4.1	Análise estatística dos dados	54
4.5	Treinamento das redes	57
4.5.1	Heurística de treinamento	59
4.5.2	Arquiteturas das redes neurais	60
4.5.3	Processamento de Imagens	61
4.5.4	Funções de Perda	62
4.5.5	Preparação dos Dados	62
4.5.6	Otimizadores	63
4.5.7	Agendadores de Taxa de Aprendizado	64
4.5.8	Amostragem de dados	64
4.6	Avaliação das Métricas	65

5 Resultados e Discussão	66
5.1 Modelos Binários	66
5.1.1 Primeira Etapa de Experimentos	66
5.1.2 Segunda Etapa de Experimentos	68
5.1.3 Análise da Abordagem Binária	70
5.2 Modelos Multiclasse	71
5.2.1 Primeira Etapa de Experimentos	71
5.2.2 Segunda Etapa de Experimentos	73
5.2.3 Terceira Etapa de Experimentos	75
5.2.3.1 Primeiro Treinamento	75
5.2.3.2 Segundo Treinamento	75
5.2.3.3 Terceiro Treinamento	76
5.2.3.4 Quarto Treinamento	77
5.2.3.5 Quinto Treinamento	78
5.2.3.6 Sexto Treinamento	79
5.2.3.7 Sétimo Treinamento	80
5.2.3.8 Resultados da Etapa	81
5.3 Resultados Finais	82
6 Conclusões e Trabalhos Futuros	84
6.1 Conclusões	84
6.2 Trabalhos Futuros	85
REFERÊNCIAS	86

1

Introdução

O câncer de mama foi o mais prevalente entre as mulheres, com 94.728 casos de incidência e 22.189 mortes registradas no Brasil em 2022 (IARC, 2024). A densidade de mama é um dos fatores de biorisco na detecção do câncer, visto que a grande quantidade de tecido fibro-glandular, encontrada em mamas densas, pode dificultar ou mascarar a identificação de achados malignos. O *Breast Imaging and Reporting Data System* (BI-RADS®) é um conjunto de regras e definições para exames de mama (SICKLES et al., 2013) e classifica a densidade em quatro categorias: (I) Mama lipo-substituída; (II) Mama fibro-glandular; (III) Mama heterogeneamente densa e (IV) Mama acentuadamente densa. Com base nessas classificações, o radiologista analisa as imagens e determina a densidade mamária por meio de sua observação.

Entretanto, a variabilidade de classificação entre os radiologistas pode gerar inconsistências ou erros, e prejudicar a análise do exame (BERG et al., 2000). Assim, vê-se a necessidade de métodos automáticos e padronizados. A utilização de *Convolutional Neural Network* (CNN) tem se mostrado promissora na área médica, a ser utilizada como uma excelente ferramenta de análise (PAWAR et al., 2022). A partir disso, o presente trabalho busca adaptar uma CNN capaz de classificar a densidade mamária de forma automática, utilizando o estado-da-arte em processamentos de imagens e arquiteturas de redes neurais vistos na literatura, a fim de proporcionar uma ferramenta que auxilie o radiologista na classificação da densidade.

1.1 Formulação do Problema

A falta de um método objetivo e confiável para avaliar a densidade de mama pode levar a uma incerteza e inconsistência entre os médicos radiologistas, e pode comprometer o tratamento e diagnóstico do paciente (DONTCHOS et al., 2021). Por conseguinte, o uso de CNNs tem demonstrado um potencial significativo na área médica, contribuindo para diversas aplicações, como processos de decisão clínica, análise de exames de imagem (mamografias, raio-x, ultrassom, entre outros), avaliação eficiente de resultados com base nos dados do paciente, monitoramento em tempo real e alertas sobre condições críticas, além de auxiliar no treinamento e capacitação de profissionais da saúde (HALEEM; JAVAID; KHAN, 2019).

Assim, o uso de um modelo de rede neural para a classificação da densidade mamária

apresenta-se como uma ferramenta viável e promissora para auxiliar os especialistas na área. Essa abordagem não apenas fornece diagnósticos mais precisos, mas também garante uma padronização na classificação, resultando em maior consistência nos resultados. Além disso, ao automatizar essa tarefa, o modelo reduz significativamente o tempo de entrega dos laudos, permitindo que os radiologistas concentrem-se em análises mais complexas e subjetivas. Isso contribui para um aumento na qualidade dos exames e na eficácia do diagnóstico, beneficiando tanto os profissionais da saúde quanto os pacientes.

1.2 Motivação

Destaca-se a necessidade da detecção precoce do câncer, visando possibilitar o tratamento antes do avanço da doença. Como um dos principais fatores de biorisco do câncer de mama, a densidade mamária é fundamental para o diagnóstico e tratamento do tumor, uma vez que mamas mais densas apresentam maior probabilidade de ocultar lesões malignas em exames de mamografia. A classificação precisa da densidade mamária, realizada por meio de modelos baseados em CNNs, pode auxiliar os médicos radiologistas na identificação de achados suspeitos, melhorando a eficácia do diagnóstico e, consequentemente, aumentando as chances de detecção precoce do tumor. Isso não apenas permite intervenções médicas mais rápidas e eficazes, mas também contribui para melhorar a qualidade de vida das pacientes, reduzindo a necessidade de tratamentos mais invasivos em estágios avançados da doença.

Além disso, a agência federal estadunidense *Food and Drug Administration* (FDA) tornou obrigatória a apresentação da densidade mamária para os pacientes nas instituições de oncologia e mamografia nos Estados Unidos (FDA, 2024), uma regulamentação que pode se tornar referência mundial. No futuro, é essencial implementar essa regulação no Brasil, proporcionando, assim, mais informações sobre a mama para as pacientes e aumentando as chances de detecção precoce do câncer. A adoção de modelos de CNNs para a classificação da densidade mamária pode ser um passo crucial nessa direção, garantindo diagnósticos mais precisos e acessíveis em larga escala.

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho tem como objetivo investigar, adaptar, treinar e comparar os principais modelos de CNNs para a classificação da densidade mamária em imagens de mamografias digitais, seguindo o padrão estabelecido pelo BI-RADS®.

1.3.2 Objetivos Específicos

- Realizar experimentos em conjuntos de dados públicos para avaliar a generalização e robustez dos modelos de CNNs na classificação da densidade mamária, considerando

variações na qualidade e na distribuição das imagens.

- Avaliar modelos que realizem tanto a classificação binária quanto a classificação multiclasse, utilizando uma metodologia de experimentos consistente para garantir a comparabilidade dos resultados.
- Aplicar técnicas de otimização para ajustar os hiperparâmetros dos modelos, visando maximizar o desempenho e a eficiência computacional.

1.4 Descrição dos Capítulos

Este trabalho está organizado em seis capítulos. No Capítulo 2, é apresentado o referencial teórico, que aborda os fundamentos de redes neurais, com ênfase em CNNs. No Capítulo 3, é realizada uma revisão bibliográfica sistemática, compilando estudos recentes que aplicam técnicas de *Deep Learning* (DL) para a classificação da densidade mamária em imagens de mamografia *Full-Field Digital Mammography* (FFDM). O Capítulo 4 detalha a metodologia adotada, descrevendo as tecnologias, ferramentas e procedimentos utilizados para o desenvolvimento do projeto proposto. No Capítulo 5, são apresentados os resultados dos experimentos. Por fim, o Capítulo 6 traz as conclusões do estudo, destacando as contribuições do trabalho, suas limitações e sugestões para pesquisas futuras.

2

Referencial Teórico

2.1 Mamografia

A mamografia é um exame de imagem que utiliza raios-X para visualizar o tecido mamário. Este exame é essencial para a detecção precoce do câncer de mama, permitindo a identificação de alterações no tecido mamário, como nódulos ou calcificações, antes mesmo de serem palpáveis. A mamografia pode ser de rastreamento, para mulheres sem sintomas, ou diagnóstica, quando há sintomas como nódulos ou dor (INCA, 2022).

Diferentemente da mamografia tradicional em filme, onde a imagem é capturada, exibida e armazenada no filme, a mamografia digital separa essas tarefas. A imagem é capturada pelo detector digital, mas é exibida em um monitor ou filme (LEWIN, 2008). O detector digital registra os raios-X detectados como sinais elétricos que são convertidos na imagem final. Em comparação com o filme, o detector digital possui maior resolução de contraste e latidez mais ampla. A imagem final é salva em um arquivo *Digital Imaging and Communications in Medicine* (DICOM), que é definido como um protocolo de intercâmbio de dados, um formato de imagem digital e uma estrutura de arquivos padronizada para o armazenamento e transferência de imagens biomédicas e metadados associados (BIDGOOD et al., 1997).

2.1.1 DICOM

O padrão DICOM tem o foco em metadados, que são essenciais para o uso completo das imagens médicas em contextos clínicos, estabelecendo a indivisibilidade dos dados de pixel em relação aos metadados. Cada imagem DICOM consiste em metadados e dados de pixel incorporados em um único arquivo, garantindo que a imagem nunca seja separada dessas informações por engano. Essa integração é fundamental para preservar o contexto clínico das imagens, como informações do paciente, parâmetros de aquisição e detalhes do equipamento utilizado, assegurando a rastreabilidade e a confiabilidade dos dados ao longo de todo o processo de diagnóstico e tratamento (LAROBINA, 2023).

2.2 Densidade de Mama

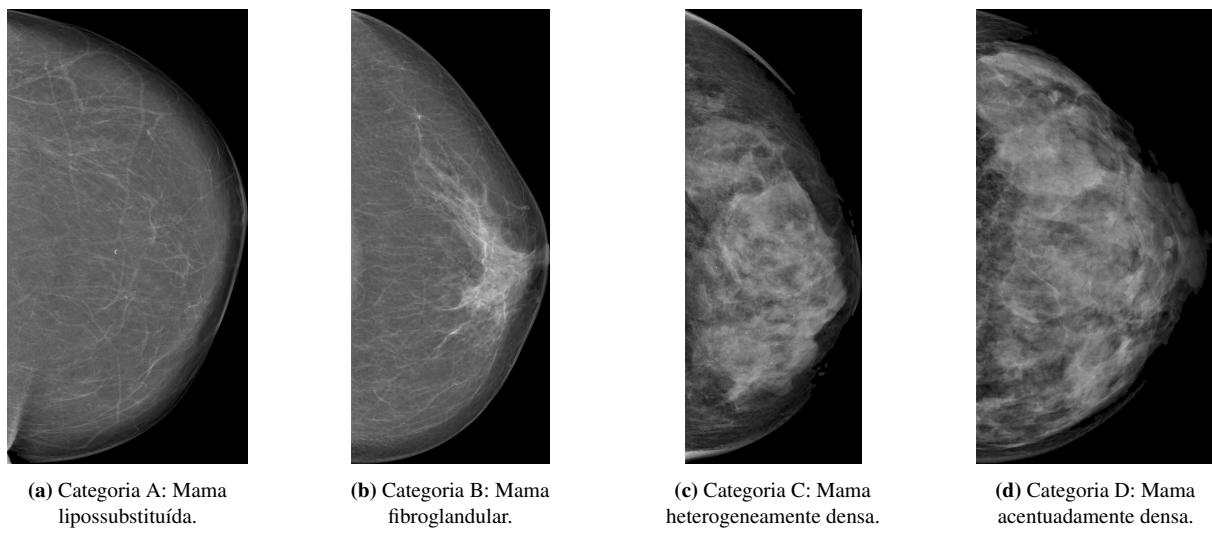
A densidade de mama refere-se à quantidade de tecido fibro-glandular, radiologicamente denso, presente na mama. A presença de tecido mamário denso está associada a uma sensibilidade mamográfica marcadamente reduzida e a uma maior taxa de câncer detectado entre os exames de rotina. Além disso, sugere-se que ter tecido mamário denso está relacionado a um aumento no risco de câncer de mama em comparação com mulheres que possuem tecido mamário adiposo (BODEWES et al., 2022). A densidade mamária é classificada pelo BI-RADS® em quatro categorias distintas (SICKLES et al., 2013):

- **Categoria 1/A:** Mama liposubstituída, caracterizada por uma maior quantidade de gordura em relação ao tecido fibro-glandular.
- **Categoria 2/B:** Mama fibroglandular, apresentando pequenas quantidades de tecido fibro-glandular distribuídas de forma dispersa.
- **Categoria 3/C:** Mama heterogeneamente densa, com uma quantidade significativa de tecido fibro-glandular, o que pode dificultar a detecção de lesões.
- **Categoria 4/D:** Mama acentuadamente densa, composta quase inteiramente por tecido fibro-glandular, o que reduz a sensibilidade da mamografia na identificação de anormalidades

Por consenso de especialistas, mamas heterogeneamente densas e mamas acentuadamente densas são categorizadas como "densas", enquanto mamas lipo-substituídas e mamas fibro-glandulares são consideradas "não densas". Além disso, o tecido denso está associado a um risco elevado de câncer de mama, embora o nível de risco seja amplamente desconhecido, pois muitos fatores influenciam a densidade do tecido, como idade, hormônios endógenos e exógenos, quimioterapia e radioterapia, e lactação (WINKLER et al., 2015).

Antes da 5^a edição do BI-RADS®, a densidade mamária era classificada de forma quantitativa, com base na porcentagem de tecido fibroglandular presente na mama. A partir da 5^a edição, essa abordagem foi substituída por uma análise qualitativa, eliminando a necessidade de uma avaliação numérica (GEMICI et al., 2020). Exemplos de imagens que ilustram cada categoria de densidade mamária, conforme a nova edição do BI-RADS®, são apresentados na Figura 2.1.

Figura 2.1: Mamografias de cada categoria de densidade mamária conforme o BI-RADS® (5^a edição). Nota-se aumento progressivo do tecido fibroglandular da categoria A à D. Imagens de bases de dados públicas e anônimas.



(a) Categoria A: Mama liposubstituída.

(b) Categoria B: Mama fibroglandular.

(c) Categoria C: Mama heterogeneamente densa.

(d) Categoria D: Mama acentuadamente densa.

A variabilidade interobservador e intraobservador na interpretação mamográfica refere-se às diferenças na classificação da densidade mamária entre diferentes profissionais (interobservador) e pelo mesmo profissional em momentos distintos (intraobservador). Essa variabilidade é um desafio reconhecido no uso do sistema BI-RADS®, pois a avaliação visual da densidade mamária pode ser subjetiva e influenciada por fatores como experiência, percepção individual e condições de leitura (BERG et al., 2000).

Estudos mostram que, sem treinamento adequado, há uma variabilidade significativa tanto interobservador quanto intraobservador, o que pode impactar a consistência dos resultados. No entanto, após capacitação específica nas diretrizes do BI-RADS®, a concordância interobservador melhora significativamente, atingindo níveis substanciais de acordo. Apesar de a variabilidade persistir em certa medida, o sistema BI-RADS® demonstra ser uma ferramenta útil e confiável para a padronização da classificação da densidade mamária. Essa padronização é essencial para garantir resultados consistentes e comparáveis, fundamentais para o rastreamento e diagnóstico do câncer de mama (OOMS et al., 2007).

2.3 Inteligência Artificial

A Inteligência Artificial (IA) é o processo de criar máquinas tão inteligentes quanto o cérebro humano, e em Ciência da Computação, é o estudo de "agentes inteligentes", que são dispositivos que percebem seu ambiente e tomam ações que maximizam sua probabilidade de alcançar com sucesso os objetivos definidos (SHINDE; SHAH, 2018).

2.4 Inteligência Artificial na Medicina

O uso de inteligência artificial e aprendizado de máquina (ML) na medicina pode oferecer melhores indicações de riscos e implicações nos diagnósticos e terapias. Além disso, essas tecnologias podem servir como uma segunda avaliação ou confirmação do diagnóstico, aumentando a precisão do profissional de saúde, reduzindo erros e aliviando tensões. Isso permite que o profissional tenha mais tempo para obter ou avaliar outros dados, ou mesmo entregar um resultado mais rapidamente ao paciente (LOBO, 2018).

A IA tem tido um impacto significativo na medicina, com aplicações que abrangem várias áreas, desde o diagnóstico até ao tratamento e gestão. No diagnóstico, a IA auxilia na análise de dados de exames radiológicos e ultrassonográficos, permitindo um diagnóstico mais rápido, eficiente e preciso. No tratamento, a IA revolucionou a cirurgia com os sistemas cirúrgicos, que oferece operações minimamente invasivas e maior precisão (LIU et al., 2021). Abordagens práticas de DL alcançam desempenho superior ao nível humano (HOLZINGER, 2020)

2.5 *Machine Learning*

Machine Learning (ML) é uma subárea de IA com o objetivo de criar técnicas computacionais sobre o aprendizado e construir sistemas que aprendem e adquirem conhecimento automaticamente (MONARD; BARANAUSKAS, 2003). Essa subárea pode ser encontrada nas tecnologias de informação, estatística, probabilidade, inteligência artificial, psicologia, medicina e muitas outras áreas. A partir dele, problemas podem ser resolvidos ao construir um modelo que tenha uma boa representação da base de dados selecionada. Ou seja, é o processo de criar algoritmos que permitem o computador "aprender", de forma a achar padrões ou estruturas de forma estatística (MUHAMEDYEV, 2015).

Técnicas de ML envolvem uma vasta classe de algoritmos, como métodos estatísticos, técnicas de métricas, *Support Vector Machine* (SVM), e redes neurais artificiais. Todas essas técnicas são otimizadas para resolver o problema central de avaliar as situações definidas (SHINDE; SHAH, 2018).

O ML é o processo de programar computadores para otimizar um critério de desempenho utilizando dados de exemplo ou experiências passadas. O modelo pode ser preditivo, visando fazer previsões sobre o futuro, descritivo, com o objetivo de adquirir conhecimento a partir dos dados, ou ainda combinar ambos os aspectos (ALPAYDIN, 2020).

Ademais, existem três principais abordagens de aprendizado de máquina: supervisionado, semi-supervisionado e não supervisionado. No aprendizado supervisionado, o modelo é treinado a partir de iterações em um conjunto de dados que possui rótulos das classes para cada objeto. O objetivo é obter um classificador capaz de generalizar e determinar corretamente as classes de dados não vistos anteriormente. Essa abordagem é amplamente utilizada em tarefas de classificação e regressão, onde a relação entre as entradas e as saídas é bem definida (LUDERMIR,

2021).

No aprendizado semi-supervisionado, o modelo utiliza tanto dados rotulados quanto não rotulados durante o treinamento. Essa abordagem é particularmente útil quando a obtenção de rótulos é custosa ou trabalhosa, mas há uma grande quantidade de dados não rotulados disponíveis. O modelo aprende a partir dos dados rotulados e, em seguida, utiliza os dados não rotulados para refinar e melhorar sua capacidade de generalização (YANG et al., 2023).

por último, no aprendizado não supervisionado, o algoritmo analisa dados sem rótulos e tenta agrupá-los em conjuntos, ou *clusters*, de forma a rotular os agrupamentos ao final do treinamento (LUDERMIR, 2021). Cada uma dessas abordagens tem suas aplicações específicas e pode ser escolhida com base na natureza do problema e na disponibilidade de dados rotulados. O foco deste trabalho serão modelos de aprendizagem supervisionado.

2.6 Deep Learning

Aprendizagem profunda, ou *Deep Learning* (DL), é uma subárea que produz inúmeros métodos para descobrir estruturas e padrões em dados de alta dimensão, e demonstra um ótimo potencial para as áreas de saúde, negócios e ciências (LECUN; BENGIO; HINTON, 2015). O DL utiliza uma cascata de múltiplas camadas de unidades de processamento não lineares para extração e transformação de características, ou seja, essa abordagem é viável para extrair informações úteis tanto de grandes quantidades de dados de uma mesma fonte ou de diferentes fontes (SHARIFANI; AMINI, 2023) (ZHANG; WANG; LIU, 2018).

o DL realizou avanços significativos e produziu resultados de ponta em diversos domínios de aplicação, como visão computacional, reconhecimento de fala e Processamento de Linguagem Natural (PLN). A disponibilidade de poder computacional devido aos avanços tecnológicos nos *hardwares*, a grande quantidade de dados de treinamento disponível e o poder da própria estrutura de DL permitiram que ela se destacasse nessas tarefas, tornando-se muito popular (ZHANG; WANG; LIU, 2018).

2.6.1 Redes Neurais Convolucionais

CNN são estruturas capazes de extrair características de dados a partir de convoluções matemáticas. Elas se assemelham a neurônios biológicos, e é possível dizer que seus neurônios artificiais são semelhantes. Camadas da CNN representam diversos receptores que reagem a diversas informações dos dados (LI et al., 2020), de forma a aprender padrões simples nas primeiras camadas, e nas camadas mais profundas aprendem detalhes e padrões difíceis ou implícitos nos dados (HAYKIN, 2001). Essas redes neurais são aplicadas a problemas de classificação, sendo esse um dos problemas mais comuns e importantes (ESCOVEDO; KOSHIYAMA, 2020).

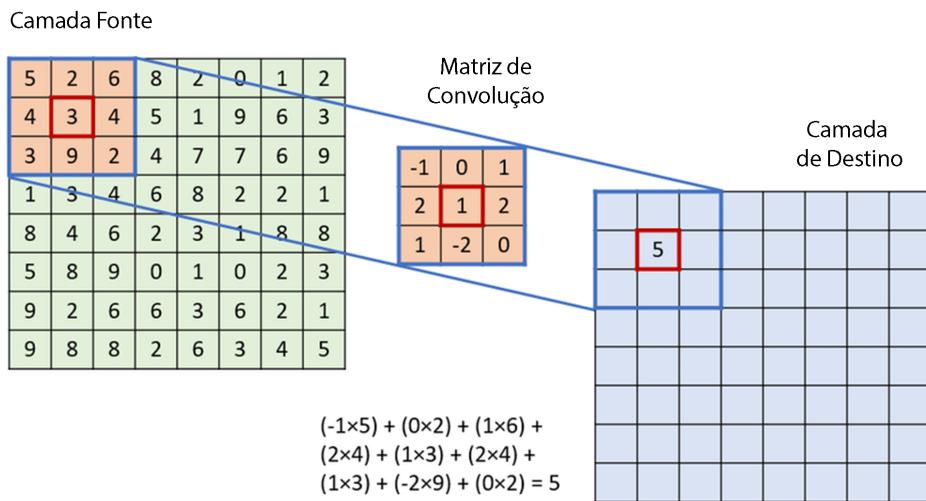
2.6.2 Arquitetura da Rede

As CNNs são feitas para processar informações com múltiplos vetores de dados. Sua arquitetura é constituída por estágios, que envolvem camadas convolucionais, camadas de *pooling*, camadas totalmente conectadas (LECUN; BENGIO; HINTON, 2015), funções de ativação e *softmax*.

2.6.3 Camada Convolutional

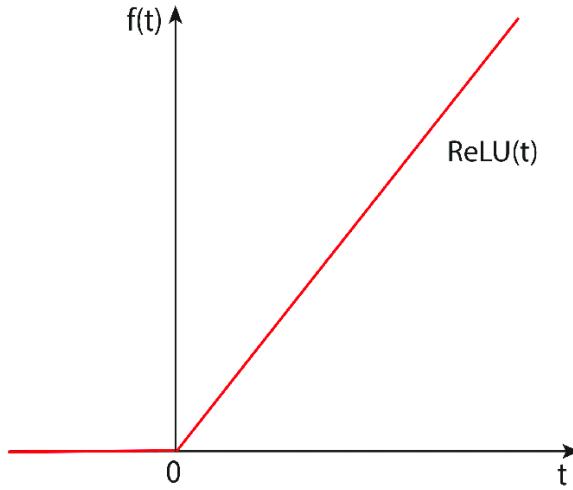
A camada de convolução, representada na Figura 2.2, é o componente mais importante de uma CNN. Ela funciona a partir de um conjunto de filtros, que recebe uma entrada de dados para gerar um mapa de características. Os filtros são uma matriz, ou *kernel*, com números discretos, que são definidos durante o aprendizado do modelo. Esses filtros modificam o valor de saída da camada de convolução, a gerar as características dos dados, e isso é feito a partir de uma operação de convolução. Ao final, os valores obtidos são enviados a uma função de ativação, que comumente são funções não-lineares, que transformam o valor real em um intervalo menor, como o *Rectified Linear Unit (ReLU)*. Isso possibilita a rede a aprender mapas não-lineares (KHAN; RAHMANI; SHAH, 2018).

Figura 2.2: Exemplo de uma camada convolucional. O *kernel* itera pela matriz aplicando o filtro na região selecionado, que resulta em uma nova matriz. Extraído e adaptado de PODAREANU et al. (2019).



As camadas convolucionais realizam múltiplas operações de convolução matemática, porém, ao final desse processo, todas essas operações podem ser combinadas em uma única convolução matemática, a partir da função de ativação. Elas desempenham um papel crucial ao introduzir não-linearidade a essas convoluções, o que permite a resolução de problemas complexos (GÉRON, 2022). Uma das funções de ativação mais utilizadas é a *ReLU*, que zera os valores negativos e mantém os valores positivos, como demonstrado na Figura 2.3.

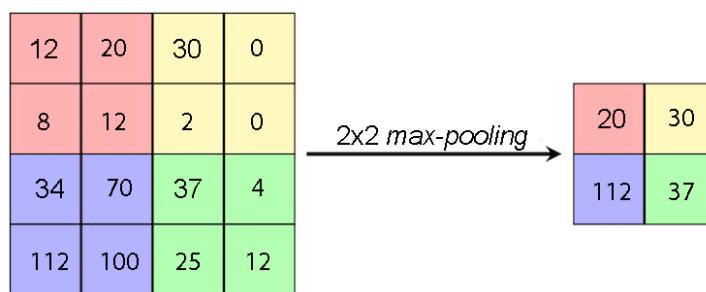
Figura 2.3: Exemplo da função de ativação *ReLU*, que transforma os valores negativos para zero e não modifica o restante. Extraído de OLANIPEKUN et al. (2022).



2.6.4 Camada de *Pooling*

A função da camada de *pooling* é juntar informações semânticas similares em apenas um valor (LECUN; BENGIO; HINTON, 2015). A camada de *pooling* reduz o tamanho do mapa de características de entrada. Esse processo é útil para obter uma representação que é invariável a mudanças em uma imagem (KHAN; RAHMANI; SHAH, 2018). Uma técnica muito utilizada na camada de *pooling* é o *max-pooling*, que envolve selecionar o maior valor obtido pelo *kernel* para a saída da camada (LI et al., 2014), como visto na Figura 2.4 .

Figura 2.4: Exemplo de *max-pooling*, no qual é retornado uma matriz com os valores máximos obtidos por cada *kernel*. Extraído de Papers With Code (2024).



2.6.5 Camada Totalmente Conectada

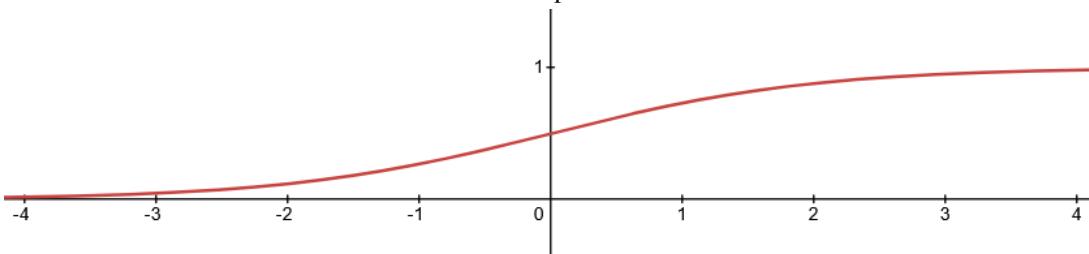
Em uma camada totalmente conectada, os neurônios possuem conexões com todas as ativações da camada anterior. Ela passa a saída para uma camada final, onde é possível utilizar uma função *softmax* ou *sigmoid* para prever o rótulo da classe dos dados de entrada (SAKIB et al., 2019). Ela funciona de forma semelhante a camada convolucional, porém o *kernel* utilizado é menor (KHAN; RAHMANI; SHAH, 2018).

2.6.6 Softmax

A função *softmax* é comumente aplicada nas últimas camadas das CNNs, de forma a converter os valores de saída em uma distribuição de probabilidade. Esse processo normaliza os valores para o intervalo [0.0, 1.0], o qual é útil para problemas de classificação multiclasse, onde o maior valor corresponde à classe com maior probabilidade (GÉRON, 2022). O gráfico da função *softmax* pode visto na Figura 2.5.

Figura 2.5: Representação gráfica da função *softmax*. A função converte os valores de saída de uma rede neural em uma distribuição de probabilidade, normalizando-os no intervalo [0.0, 1.0].

Elaborado pelo autor.



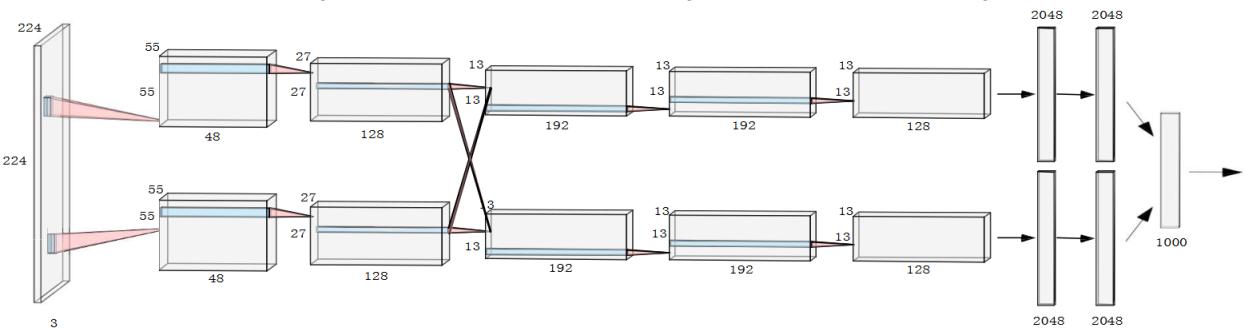
2.7 Arquitetura de Redes Neurais Convolucionais

Existem inúmeras arquiteturas de CNNs. Desde o primeiro modelo proposto por Yann LeCun, conhecido como LeNet (LECUN et al., 1998), até a arquitetura que revelou o potencial das CNNs, o *AlexNet*, muitas outras foram desenvolvidas com maior profundidade, largura ou leveza (LI et al., 2020). Cada modelo possui sua própria arquitetura, com o objetivo de solucionar problemas específicos ou melhorar a estrutura de redes anteriores para solucionar o mesmo problema.

2.7.1 AlexNet

(KRIZHEVSKY; SUTSKEVER; HINTON, 2012) criaram uma CNN para treinar em uma grande quantidade de dados, como na base de imagens *ImageNet* (*ImageNet*), que contém mais de 15 milhões de imagens rotuladas com mais de 22 mil categorias. A base de dados escolhida foi um subconjunto do *Imagenet*, o *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC), e a resolução das imagens utilizada foi de 256x256. Os autores utilizam técnicas de *dropout* e *data augmentation* para evitar *overfitting* do modelo. A arquitetura consiste em oito camadas: cinco de convolução e três camadas totalmente conectadas, que pode ser visto na Figura 2.6. Os resultados da CNN treinada superaram o estado-da-arte e quebraram recordes, e o modelo se tornou referência para a tarefa de classificação de imagens. O *AlexNet* demonstrou o potencial das CNNs, e foi considerada a pioneira na popularização de DL (ALOM et al., 2018).

Figura 2.6: Arquitetura da *AlexNet*. A partir dos grupos de convolução, foi possível treinar o modelo em duas *Graphical Processing Units* (GPUs) diferentes, com os vetores de características sendo concatenados ao final das convoluções. Essa abordagem representou um avanço significativo, dada a limitação computacional da época. Extraído de LI et al. (2020)



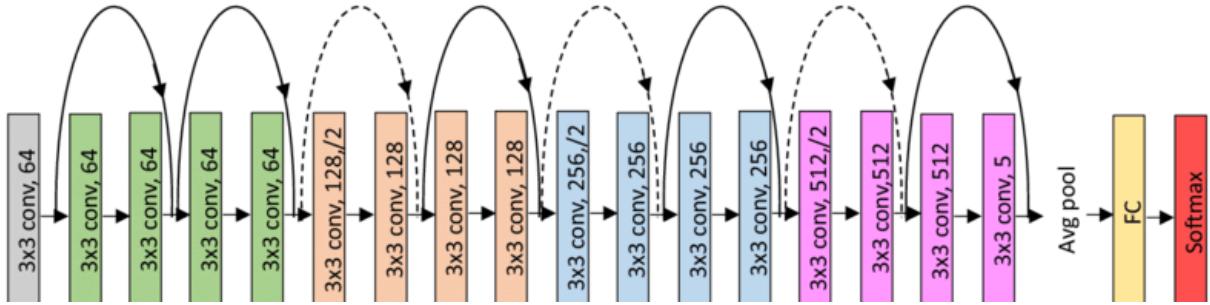
2.7.2 EfficientNets

Os autores TAN; LE (2019) buscavam uma rede precisa e com baixo custo computacional simultaneamente. Para isso, desenvolveram o método chamado de escalonamento composto (*compound scaling method*), no qual as resoluções da imagem são escaladas de maneira igual e balanceada. Esse método é justificado pela intuição de que se a imagem de entrada for maior, são necessárias mais camadas para aumentar o campo receptivo e mais canais para identificar os detalhes da imagem com maior precisão.

2.7.3 ResNets

Com o gradual aumento de camadas de convolução, as CNNs sofrem com problemas de *vanishing/exploding gradients*, ou seja, gradientes se tornando muito pequenos ou muito grandes, resultando no lento aprendizado do modelo LI et al. (2020). Dessa forma, HE et al. (2015) criaram as *ResNets*, CNNs capazes de realizar conexões residuais, que são atalhos de uma camada para a saída de outra camada mais profunda, de forma que os gradientes sejam propagados diminuindo o problema encontrado nos gradientes. As *ResNets* são divididas pela sua quantidade de camadas, como *ResNet-18*, *ResNet-34* e *Resnet-50*. A arquitetura da *ResNet-18* se encontra na Figura 2.7

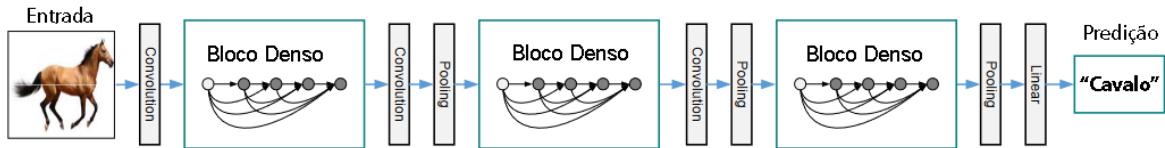
Figura 2.7: Arquitetura da *ResNet-18*. As setas curvadas demonstram as conexões residuais da rede, possibilitando a propagação dos gradientes para as camadas seguintes. Extraído e adaptado de WU et al. (2021).



2.7.4 DenseNets

DenseNets são semelhantes a *ResNets*, pois ambos utilizam conexões entre as camadas, a diferença é que os gradientes são concatenados na *DenseNet*, e não somados. A concatenação permite que cada camada acesse diretamente os mapas de características das camadas anteriores, o que evita o aprendizado duplicado das características (HUANG et al., 2018). A arquitetura está presente na Figura 2.8

Figura 2.8: Arquitetura da *DenseNet*. Possui blocos densos, onde os resultados de cada camada são concatenados nas seguintes. Extraído e adaptado de HUANG et al. (2018).

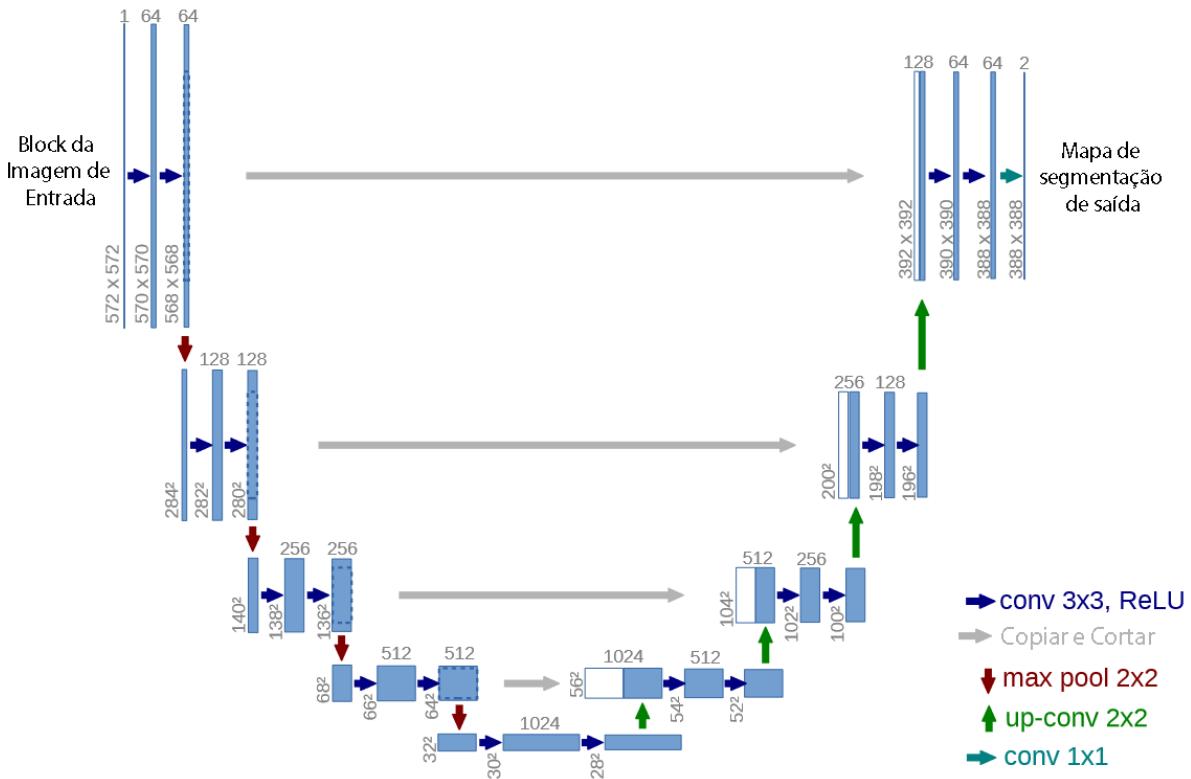


2.7.5 U-Nets

Na medicina, muitos problemas de classificação envolvem também a localização do achado, ou seja, cada pixel da imagem deve possuir uma classificação. Assim, para resolver essa tarefa, RONNERBERGER; FISCHER; BROX (2015) desenvolveram uma arquitetura denominada de *U-Net*. A arquitetura é composta pelas seguintes etapas: a de contração e a de expansão. A primeira etapa funciona como uma CNN normal, na qual cada bloco possui uma camada de convolução, uma *ReLU* e uma camada de *pooling* para duplicar o vetor de características. Na segunda etapa, de expansão, os blocos tem como objetivo reduzir o vetor de característica do modelo ao fazer *upsampling* do mapa de características e concatenar as imagens, de forma a produzir uma única camada de convolução capaz de inferir as classes. O modelo foi treinado com poucos dados, por isso, os autores optaram por utilizar *Data Augmentation*

excessivamente, de forma a fazer o modelo aprender invariâncias e deformações. A arquitetura se encontra na Figura 2.9.

Figura 2.9: Arquitetura da *U-Net*. Nota-se que o nome da arquitetura é dado pelo seu formato, que se assemelha a letra U. Extraído e adaptado de RONNERBERGER; FISCHER; BROX (2015).



2.8 Redes Adversárias Generativas

As Redes Adversárias Generativas *Generative Adversarial Networks* (GAN) fornecem uma maneira de aprender representações profundas sem a necessidade de dados de treinamento extensivamente anotados. Esse modelo gerativo aprende a capturar a distribuição estatística dos dados de treinamento, permitindo-nos sintetizar amostras a partir da distribuição aprendida (GOODFELLOW et al., 2020). As GAN são utilizadas para tarefas de edição de imagens, *data augmentation*, classificação e recuperação de imagens (CRESWELL et al., 2018).

Elas podem ser caracterizadas pelo treinamento de um par de redes em competição entre si. Uma analogia comum sobre dados visuais é pensar na rede como um falsificador de arte e um especialista de arte. O falsificador cria as falsificações com o objetivo de produzir imagens realistas, e o especialista recebe tanto falsificações quanto imagens reais, e tenta diferenciá-las (CRESWELL et al., 2018).

2.9 Data Augmentation

Data Augmentation é uma estratégia de DL para prevenir um superajuste (*overfitting*) dos dados por meio da regularização. Esse processo gera dados automaticamente durante o treinamento, compensando uma baixa quantidade do conjunto ou diversificar os dados utilizados. No contexto de imagens como dados, algumas técnicas de *data augmentation* incluem o uso de rotações, inversões, adição de ruídos, deslocamentos, recortes, e a utilização de GANs para a geração de novos dados (MAHARANA; MONDAL; NEMADE, 2022). Essas técnicas enriquecem o conjunto de dados, melhorando a robustez e a generalização dos modelos de DL.

2.10 Aprendizado por Transferência e Ajuste Fino

As técnicas de aprendizado por transferência (*transfer learning*) e ajuste fino (*fine-tuning*) são estratégias fundamentais para o treinamento de modelos de DL, especialmente quando há limitações na quantidade de dados disponíveis.

No aprendizado por transferência, o objetivo é aproveitar os pesos de um modelo pré-treinado em uma tarefa semelhante, transferindo o conhecimento adquirido para um novo problema. O modelo pode ter parte de suas camadas congeladas, para que mantenha os pesos originais e reduza o custo computacional, além de focar o treinamento nas camadas ativadas, que são geralmente as últimas, por aprenderem características mais abstratas e específicas da tarefa (VRBANCIC; PODGORELEC, 2020). Enquanto isso, na técnica de ajuste fino, o modelo ajusta todas as suas camadas para a nova tarefa, o que requer mais poder computacional e dados de treinamento.

Ambas as técnicas são amplamente utilizadas em problemas de classificação de imagens, como a classificação de densidade mamária. Neste projeto, quando nos referirmos ao treinamento de um modelo, estaremos tratando especificamente do ajuste fino do modelo.

2.11 Integração dos Dados

A integração de dados é um processo fundamental na ciência de dados e em pesquisas que envolvem múltiplas fontes de informação. Consiste na combinação de dados provenientes de diferentes origens, formatos e estruturas em um repositório único e coerente, permitindo uma visão unificada e consistente das informações. Essa etapa é crucial para garantir que os dados estejam prontos para análise, eliminando redundâncias, inconsistências e lacunas que possam comprometer os resultados (MAHARANA; MONDAL; NEMADE, 2022).

A heterogeneidade das fontes de dados é um dos principais desafios enfrentados nesse processo. Dados podem variar em formato, protocolos de acesso, padrões de codificação e semântica (DOAN; HALEVY; IVES, 2012). Por exemplo, um banco de dados hospitalar pode usar diferentes terminologias para descrever o mesmo conceito clínico, enquanto um sistema de imagens médicas pode armazenar informações distintas em um arquivo DICOM.

2.12 Problemas de Classificação

Problemas de classificação podem ser definidos como aqueles em que algoritmos correlacionam uma entrada X a uma saída Y . O objetivo do algoritmo é mapear essas relações de tal maneira que, ao receber novos dados, ele seja capaz de realizar essa função com a maior precisão possível (ESCOVEDO; KOSHIYAMA, 2020). Existem dois tipos principais de problemas de classificação: a classificação binária e a classificação multi-classe. Na classificação binária, o modelo determina se os dados pertencem ou não a uma determinada classe. Na classificação multi-classe, o modelo atribui probabilidades a cada classe possível, sendo a classe com a maior probabilidade a escolhida como a classe inferida (KOLO, 2011).

2.13 Métricas de Avaliação de Desempenho

Existe uma necessidade de validar a qualidade dos métodos e modelos utilizados em ML. Assim, a avaliação de um modelo de classificação é feita a partir de métricas, que comparam as classes inferidas pelos modelos com as classes reais. As métricas tem como objetivo medir o quanto distante o modelo está da classificação perfeita (BOTCHKAREV, 2019).

2.13.1 Matriz de Confusão

As métricas da qualidade de uma classificação são construídas a partir de uma Matriz de Confusão (MC), que registra as classificações corretas e incorretas feitas pelo modelo para cada classe do conjunto de dados. Para isso, são utilizados quatro termos: Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) (SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006). Os VPs ocorrem quando o valor real é positivo e foi classificado corretamente como positivo. Os VNs acontecem quando o valor real é negativo e foi classificado corretamente como negativo. Os FPs indicam que o valor real é negativo, porém foi classificado incorretamente como positivo. Por fim, os FNs ocorrem quando o valor real é positivo, mas foi classificado incorretamente como negativo. Um exemplo da matriz de confusão se encontra na Tabela 2.1.

Tabela 2.1: Matriz de Confusão para classificação binária. A partir dos valores, é possível calcular inúmeras métricas de validação, como acurácia, precisão, especificidade, entre outras.

		Real	
		Classe	A
Predito	A	VP	FN
	B	FP	VN

Em outras palavras, a MC registra o número de ocorrências entre duas visualizações: a classificação verdadeira e a classificação predita pelo modelo (GRANDINI; BAGLI; VISANI, 2020). Um exemplo com múltiplas classes se encontra na Tabela 2.2.

Tabela 2.2: Matriz de Confusão para múltiplas classes. A partir dos valores, é possível calcular inúmeras métricas de validação, como acurácia, precisão, especificidade, entre outras.

		Real			
		Classes	Classe A	Classe B	Classe C
Predito	Classe A	50	2	1	
	Classe B	10	45	5	
	Classe C	0	8	40	

2.13.2 Acurácia

A Acurácia (ACC) é a probabilidade das previsões estarem corretas. Em outras palavras, a acurácia retorna uma medida de quão bem o modelo está predizendo as classes de todo o conjunto de dados, com cada elemento tendo o mesmo peso e contribuindo de forma igual para o cálculo (GRANDINI; BAGLI; VISANI, 2020). É uma métrica menos útil para classes desbalanceadas. A equação da ACC está disposta na Equação (2.1).

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \quad (2.1)$$

2.13.3 Precisão

A métrica de precisão, ou PPV, expressa a proporção de unidades que o modelo diz ser positivo e realmente são positivos, ou seja, nos diz quanto devemos confiar no modelo quando prediz uma classificação positiva (GRANDINI; BAGLI; VISANI, 2020). A métrica é boa para classes desbalanceadas e quando o custo de FP é alto, além de não considerar casos de FN. A equação para calcular a precisão está disposta em Equação (2.2).

$$Precisão = \frac{VP}{VP + FP} \quad (2.2)$$

2.13.4 Recall

A métrica de *Recall*, também chamada de sensibilidade ou taxa de VP, mede a capacidade do modelo de encontrar todos os casos positivos no conjunto (GRANDINI; BAGLI; VISANI, 2020). A métrica é boa para classes desbalanceadas. e para quando o custo de FN é alto, além de não considerar casos de FP. A equação para calcular o *recall* está presente na Equação (2.3).

$$Recall = \frac{VP}{VP + FN} \quad (2.3)$$

2.13.5 Escore F1

O *F1 Score* (F1) agrupa a precisão e o *recall* ao utilizar conceitos de média harmônica, e pode ser interpretada como uma média ponderada entre as duas métricas (SOKOLOVA;

JAPKOWICZ; SZPAKOWICZ, 2006). Quanto maior o resultado de F1, melhor a classificação do modelo. O cálculo para o F1 se encontra em Equação (2.4).

$$F1 = 2 \cdot \left(\frac{\text{precisão} \cdot \text{recall}}{\text{precisão} + \text{recall}} \right) \quad (2.4)$$

Existe também o *macro* F1, que é calculado a partir da média da precisão e do *recall* para todas as classes (GRANDINI; BAGLI; VISANI, 2020).

2.13.6 Sensibilidade e Especificidade

A sensibilidade mede a proporção de casos positivos corretamente classificados, e é o mesmo que *recall*. A diferença existe devido a métrica ser muito utilizada em contextos médicos, pois indica o quanto bom é a rede para classificar doenças. A especificidade mede a proporção de casos negativos corretamente classificados (TING, 2010), ou seja, o quanto bom a rede classifica casos negativos de doenças. As fórmulas da sensibilidade e especificidade estão dispostas em Equação (2.5) e Equação (2.6), respectivamente.

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (2.5)$$

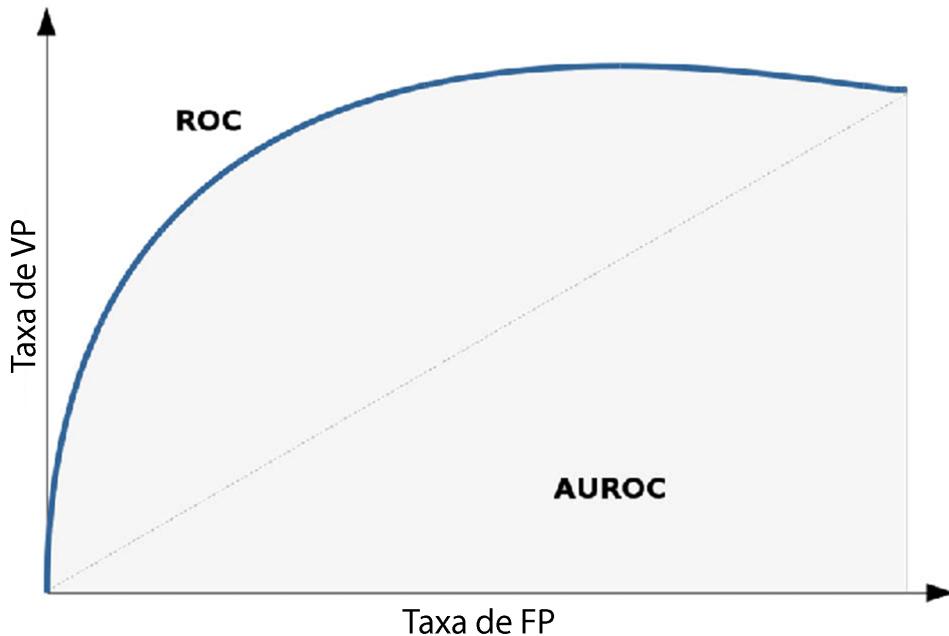
$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} \quad (2.6)$$

2.13.7 Curva ROC e AUROC

A curva *Receiver Operating Characteristic* (ROC) é um método analítico, representado como um gráfico, que serve para avaliar o desempenho de um método de classificação binário. Ele conecta a especificidade com a sensitividade nos eixos x e y, respectivamente. Quanto mais rigorosos os critérios de avaliação, mais pontos da curva se deslocam para a esquerda e para baixo. Em contrapartida, quanto menos rigorosos os critérios de avaliação, mais a curva expande para cima e para a direita (NAHM, 2022). Ou seja, quanto mais próxima a curva estiver no canto superior esquerdo, melhor é o desempenho do modelo, pois a sensibilidade é alta e a taxa de falsos positivos é baixa.

A métrica *Area Under the Curve* (AUC) é utilizada para medir a acurácia dos modelos, e é calculada integrando a função que descreve a Curva ROC ao longo de todos os possíveis valores de pontos de corte. A métrica varia entre 0 e 1, e valores abaixo de 0.6 são considerados falhas. Valores entre 0.6 e 0.7 são considerados ruins. Valores entre 0.7 e 0.8 são considerados justos. Valores entre 0.8 e 0.9 são bons e valores acima de 0.9 são excelentes (NAHM, 2022). Um exemplo das métricas ROC e AUC se encontram na Figura 2.10.

Figura 2.10: Exemplo da curva ROC e da AUC. Quanto mais perto do canto superior esquerdo, melhor o modelo está desempenhando. A área em cinza é o valor do AUC. Extraído e adaptado de SAIA et al. (2021).



2.13.8 Cohen's Kappa

A métrica *Cohen's Kappa* (κ), criada por COHEN (1960), avalia o grau de concordância entre dois avaliadores, como as avaliações entre um modelo de classificação e um radiologista. Essa métrica é uma estatística que pode ser obtida a partir da seguinte fórmula:

$$\text{Cohen's } \kappa = \frac{P_0 - P_e}{1 - P_e} \quad (2.7)$$

P_0 é importante para compreender a qualidade de acordo entre os avaliadores disponíveis, ou seja, é a proporção de concordância dos avaliadores em todas as observações. O P_e é a proporção de concordância esperada ao acaso, e é calculada com base nas distribuições total das avaliações de cada avaliador. O cálculo de cada um se encontra a seguir

$$P_0 = \frac{\text{VP} + \text{VN}}{N} \quad (2.8)$$

$$P_e = P_{\text{positive}} + P_{\text{negative}} \quad (2.9)$$

$$P_{\text{positive}} = \frac{\text{VP} + \text{FN}}{N} \cdot \frac{\text{VP} + \text{FP}}{N} \quad (2.10)$$

$$P_{\text{negative}} = \frac{\text{VN} + \text{FP}}{N} \cdot \frac{\text{VN} + \text{FN}}{N} \quad (2.11)$$

Os valores de Cohen's κ podem variar entre -1 e 1. O valor 0 indica concordância por chance. Um valor entre 0.10 e 0.20 é uma concordância ligeira. Um valor entre 0.21 e 0.40 é uma concordância justa. Um valor entre 0.41 e 0.60 é uma concordância moderada. Um valor entre 0.61 e 0.80 é uma concordância substancial. Um valor entre 0.81 e 0.99 é uma concordância quase perfeita, e o valor 1 indica a concordância perfeita entre os avaliadores (GRANDINI; BAGLI; VISANI, 2020).

2.14 Validação Cruzada

Treinar um algoritmo e avaliar sua performance estatística no mesmo conjunto de dados em que foi treinado resulta em uma avaliação superestimada, uma vez que o modelo pode ter se ajustado muito aos dados vistos, causando o *overfitting* e gerando um modelo que não generalize bem em novos dados. Dessa forma, a validação cruzada foi desenvolvida para mitigar esse problema. Na maioria das aplicações, a quantidade de dados é limitada, o que leva a ideia de separar os dados em uma parte de treinamento e uma de validação, a fornecer, assim, uma estimativa mais realista da performance do modelo em dados não vistos (ARLOT; CELISSE, 2010).

Uma das variantes mais comuns da validação cruzada é a *K-Fold*, ou K-dobras. Nesse método, os dados são divididos em K subconjuntos (ou "dobras") de forma aleatória e com uma distribuição aproximadamente igual entre as classes em cada subconjunto. Em cada iteração, um subconjunto é definido como conjunto de validação, e o modelo é treinado utilizando os K - 1 subconjuntos restantes. O processo é repetido K vezes, cada turno trocando o conjunto de validação. Esse método garante que cada amostra do conjunto de dados seja usada tanto para treinamento quanto para validação, resultando em uma avaliação mais robusta da performance do modelo (CROSS-VALIDATION, 2009). A Figura 2.11 contém um exemplo de validação cruzada.

Figura 2.11: Validação cruzada em 5 dobras. A imagem representa um treinamento com cinco modelos, cada qual com uma dobra (*fold*) diferente de validação. Elaborado pelo autor.

	Dobra 1	Dobra 2	Dobra 3	Dobra 4	Dobra 5
Iteração 1	Validação	Treinamento	Treinamento	Treinamento	Treinamento
Iteração 2	Treinamento	Validação	Treinamento	Treinamento	Treinamento
Iteração 3	Treinamento	Treinamento	Validação	Treinamento	Treinamento
Iteração 4	Treinamento	Treinamento	Treinamento	Validação	Treinamento
Iteração 5	Treinamento	Treinamento	Treinamento	Treinamento	Validação

3

Revisão da Literatura

A revisão da literatura é a base fundamental para embasar o trabalho, pois define artigos atuais e relevantes ao tema proposto. Dessa forma, para produzir um conteúdo de qualidade, a técnica de Mapeamento Sistemático (MS) (*surveys*) foi escolhida, em conformidade com as diretrizes propostas por KITCHENHAM et al. (2010) e visto em PEGORINI et al. (2019), para auxiliar no levantamento de trabalhos, estudos e dados.

3.1 Questões de pesquisa

A primeira etapa do MS é a definição de Questões de Pesquisa (QP) sobre a temática escolhida. As QP visam compreender a arquitetura selecionada para a tarefa de classificação de densidade mamária, abordando as bases de imagens utilizadas, o pré-processamento das imagens e os modelos de redes neurais aplicados no treinamento e classificação dos dados. Assim, as seguintes QP foram propostas:

1. Quais pré-processamentos estão sendo aplicados nas imagens antes do treinamento?
2. Quais modelos de CNN estão demonstrando os melhores desempenhos na tarefa de classificação de densidade mamária?
3. Os modelos de CNN classificam a densidade mamária de acordo com as definições do BI-RADS®?

3.2 Estratégias de Busca

Os estudos selecionados foram buscados nas bases de dados eletrônicas *Google Scholar*, *MDPI*, *Science Direct*, *RSNA Journals*, *Frontiers*, *IEEX Xplore*, *SCIRP*, *Springer* e *ArXiv*. Essas bases foram escolhidas devido a quantidade disponível em seus acervos, da relevância acadêmica e da disponibilidade dos artigos. O link para o acesso desses acervos está disponível na tabela Tabela 3.1.

Para cada base de dados, foi empregado um método de busca avançada utilizando *strings* de busca, disponíveis na tabela Tabela 3.2. Esse método permite uma pesquisa precisa e focada

Tabela 3.1: Lista das bases de dados eletrônicas nas quais a pesquisa foi realizada, organizadas com os *links* de acesso (Elaborado pelo autor).

Base de Dados	Link para Acesso
Google Scholar	< https://scholar.google.com/ >
MDPI	< https://www.mdpi.com/ >
Science Direct	< https://www.sciencedirect.com/ >
RSNA Journals	< https://pubs.rsna.org/ >
Frontiers in Radiology	< https://www.frontiersin.org/journals/radiology >
IEEE Xplore	< https://ieeexplore.ieee.org/Xplore/home.jsp >
Scientific Research Publishing	< https://www.scirp.org/ >
Springer Link	< https://link.springer.com/ >
ArXiv	< https://arxiv.org/ >

em artigos relacionados ao tema, ampliando o escopo dos estudos incluídos no trabalho. As *strings* de busca foram geradas com base nas palavras-chave que seguem as QP definidas anteriormente.

3.3 Critérios de Inclusão e Exclusão

Com base nas QP da Seção 3.1, foram definidas três Critérios de Inclusão (CI) e quatro Critérios de Exclusão (CE), com o objetivo de auxiliar na seleção de artigos e estudos pertinentes à pesquisa. Os CIs visam incluir estudos relevantes para o trabalho, como visto a seguir:

1. Estudos primários que citam as bases de imagens utilizadas para a classificação de densidade mamária.
2. Estudos primários que apresentam abordagens no processamento das imagens para treinar o modelo.
3. Estudos primários que apresentam modelos de redes neurais convolucionais treinados para classificação de densidade mamária.
4. Estudos primários que foram publicados após 2018, com o foco na busca por técnicas de processamento de imagens.

Da mesma forma, os CE ajudam a excluir artigos não relevantes à pesquisa, dado que auxiliam na escolha dos trabalhos e estudos encontrados. Eles estão definidos a seguir:

1. Estudos primários que apresentem apenas uma revisão teórica baseada em pesquisa bibliográfica.
2. Estudos primários que não utilizem imagens de mamografias no treinamento dos modelos.
3. Estudos primários que sejam versões defasadas ou anteriores a estudos mais recentes.

Tabela 3.2: Strings de busca em cada base de dados eletrônica. Foram utilizadas como filtros ou parâmetros para buscas personalizadas (Elaborado pelo autor).

Bases de dados	Strings de busca
Google Scholar	"breast density classification" OR "breast density classification model" OR "deep learning breast density"
MDPI	"All fields": "breast density" AND "All fields": "deep learning" OR "All fields": "breast density" AND "All fields": "artificial intelligence"
MDPI	"All fields": "breast density" AND "All fields": "deep learning" AND "Abstract": "density"
MDPI	"All fields": "breast density" AND "All fields": "machine learning" AND "Abstract": "density"
ScienceDirect	"Find articles with these terms": "ffdm" AND "Title, abstract or author-specified keywords": breast density deep learning
ScienceDirect	"Title, abstract or author-specified keywords": breast density deep learning
RSNA Journals	"Abstract": "density" AND "Anywhere": "mammogram" AND "Anywhere": "deep learning" AND "Anywhere": "breast density"
RSNA Journals	"Abstract": "density" AND "Anywhere": "mammogram" AND "Anywhere": "machine learning" AND "Anywhere": "breast density"
Frontiers	"Health" AND "Frontiers in Radiology" AND "Artificial Intelligence in Radiology" AND "breast density classification"
IEEE Xplore	("All Metadata":breast density) AND ("Abstract":density) AND ("All Metadata":deep learning) AND ("All Metadata":mammography)
Springer Link	breast density deep learning mammography
ACM Digital Library	[All: breast density] AND [Abstract: density] AND [All: deep learning] AND [All: mammography] AND [E-Publication Date: (01/01/2018 TO *)]
ACM Digital Library	[All: breast density] AND [Abstract: density] AND [All: machine learning] AND [All: mammography] AND [E-Publication Date: (01/01/2018 TO *)]
ACM Digital Library	[All: breast density] AND [Abstract: density] AND [All: classification] AND [All: mammography] AND [E-Publication Date: (01/01/2018 TO *)]
ArXiv	order: -announced_date_first; size: 50; date_range: from 2018-01-01 ; classification: Computer Science (cs); include_cross_list: True; terms: AND all=breast density; AND abstract=density; AND all=deep learning; AND all=mammography"

4. Estudos primários que não sigam o BI-RADS® para classificação da densidade.
5. Estudos primários que utilizem sistemas *Computer-Aided Diagnosis* (CADs) privados.
6. Estudos primários que não utilizem DL.
7. Estudos primários cuja versão completa não esteja disponível.
8. Estudos primários publicados antes de 2018.

Serão considerados os trabalhos que se encaixam em ao menos um dos CI e desconsiderados aqueles que atendem a pelo menos um dos CE, de acordo com as diretrizes de DERMEVAL; COELHO; BITTENCOURT (2020).

3.4 Condução do Mapeamento Sistemático

Para selecionar os estudos relevantes, seguiu-se as etapas: *i*) leitura do título e *abstract* de cada estudo; *ii*) leitura da introdução e conclusão de cada estudo; *iii*) leitura integral dos estudos restantes. Em todas as etapas, aplicaram-se os critérios da Seção 3.3, a reduzir o número de estudos não relevantes.

Após a consulta nas bases de dados eletrônicas, disponíveis em Tabela 3.1, foram obtidos 1816 estudos, conforme a lista da Tabela 3.3.

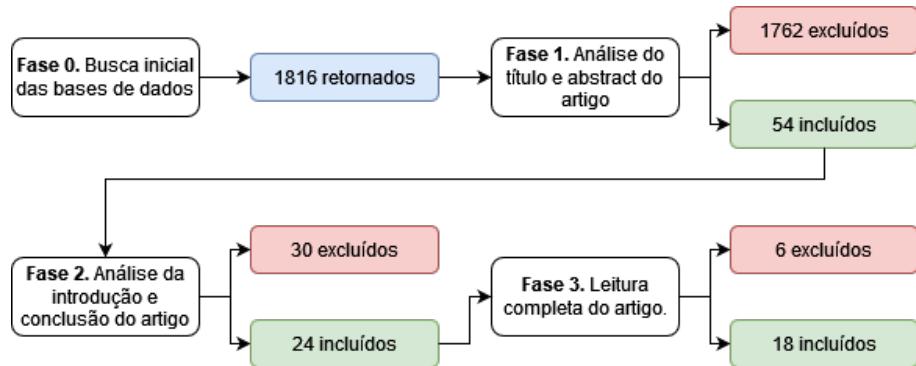
Tabela 3.3: Quantidade de estudos retornados por base de dados eletrônica após as pesquisas com as strings de busca (Elaborado pelo autor)

Base de dados	Quantidade
ACM Digital Library	10
Arxiv	17
Frontiers	15
Google Scholar	1.230
IEEE Xplore	34
MDPI	163
RSNA Journals	27
Science Direct	52
Scirp	60
Springer Link	208
Total	1816

A partir dos estudos retornados, realizou-se um filtro com base na leitura dos títulos. Nos casos em que o título indicava potencial relevância para o trabalho, procedeu-se à leitura do resumo (*abstract*). Dessa forma, dos artigos identificados, 1762 foram excluídos após essa etapa, pois não apresentavam relevância para o trabalho. Dos 54 artigos restantes, 30 foram removidos devido a introdução e conclusão não serem adequadas para o contexto do estudo. Os 24 artigos

restantes foram lidos na íntegra, resultando na exclusão de mais 6 estudos. Ao final, 18 estudos relevantes foram incluídos na análise. A condução do mapeamento sistemático está ilustrado na Figura 3.1.

Figura 3.1: Fluxograma da condução do Mapeamento Sistemático, que representa o processo de aquisição dos estudos para a revisão da literatura. Elaborado pelo autor.



3.5 Análise e Síntese dos Resultados

DONTCHOS et al. (2021) treinam um modelo de *ResNet18* a partir de uma base de dados privada, com 58.894 imagens de mamografias digitais, nas quatro categorias do BI-RADS®. com o objetivo de comparar o resultado da CNN com o diagnóstico do radiologista. O teste foi realizado em uma clínica parceira e em clínicas externas, e na primeira, o resultado da rede neural foi aceita em 94% dos exames, enquanto em clínicas externas a aceitação alcançou 92,1%. Além disso, o modelo reduziu a proporção de mamografias definidas como densas, pois obteve menos classificações nesta categoria do que nas classificações dos radiologistas.

CHANG et al. (2020) utilizam um conjunto de 108.230 imagens de mamografia digital de 21.759 pacientes, oriundas de 33 clínicas, obtidas a partir de PISANO et al. (2005). O estudo investiga como a distribuição dos dados, a escolha do modelo, seus parâmetros e sua performance influenciam na tarefa de classificação de densidade de mama. Para isso, foram escolhidos as quatro categorias de classificação do BI-RADS® e o *Linear Kappa* (*Linear κ*) foi utilizado para avaliar a métrica. O treinamento do modelo foi realizado tanto com 2% da base de dados quanto com 100%, a fim de avaliar a diferença de desempenho. Observou-se que o *Linear κ* apresentou melhores resultados quando todas as imagens de treinamento foram utilizadas. Houve também aumento artificial dos dados por meio de técnicas de inversão e rotação aleatória das imagens. Quatro arquiteturas de redes neurais foram avaliadas: *ResNet50*, *DenseNet121*, *InceptionV3* e *VGG16*. Entre elas, a *ResNet50* desempenhou melhor nos dados. Além disso, os autores avaliaram a performance do uso de um conjunto de dois a quatro modelos da mesma arquitetura, a combinar os resultados de cada um para conseguir o resultado final. Esse método mostrou uma melhoria na métrica *Linear κ* de 0,660 para 0,667. Duas funções de perda foram utilizadas, *Cross-Entropy Loss* (CE LOSS) e *Ordinal Regression Loss* (OR LOSS),

as quais tiveram desempenhos semelhantes. O modelo final alcançou Linear κ de 0,667. Foi notável também que a utilização de redes pré-treinadas resultou em melhorias na métrica. O estudo destacou a importância do balanceamento das classes, observando que, ao utilizar um conjunto com distribuição igual, o modelo acaba prevendo mais casos minoritários e menos casos majoritários, aumentando a sensibilidade, mas reduzindo a especificidade. Outra observação importante é a de que modelos treinados em um certo formato de FFDM não generalizam bem para casos com formatos de dados diferentes, como filtros, exposições, e focos do tubo de *x-ray* diferentes.

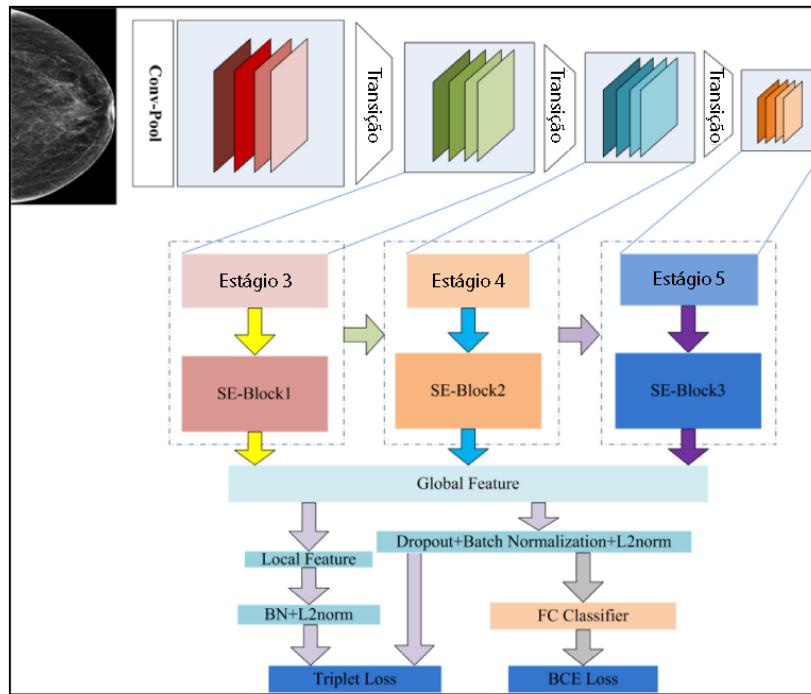
LOPEZ-ALMAZAN et al. (2022) desenvolvem um modelo que utiliza a avaliação de radiologistas para gerar MC como *ground-truth* para o treinamento, além de incluir na arquitetura uma etapa de pré-processamento para imagens com alto nível de ruído. Os autores reuniram um conjunto de 1.395 imagens FFDM de três centros clínicos para estimar a densidade mamária. A arquitetura desenvolvida, denominada *RegL*, é dividida em duas etapas: processamento de imagens e classificação da densidade de mama por meio de uma *Confusion Matrix Convolutional Neural Network* (CM-CNN). Na primeira etapa, os autores realizaram remoção de ruído do fundo das imagens, segmentação da mama, ajuste de intensidade e normalização dos pixels, resultando em uma imagem limpa e facilitando o aprendizado do modelo ao eliminar variáveis externas. Na segunda etapa, o modelo CM-CNN criado permite aprender as classificações por meios das MCs. Este modelo utiliza como base a rede *DenseNet121* para treinamento de imagens com resolução de 224x224 pixels, normalizadas no intervalo [0,1], e emprega a função de perda *Weighted Cross-Entropy Loss* (WCE LOSS).

LEHMAN et al. (2019) avalia a implementação clínica de um modelo de DL treinado com dados de concordância entre os radiologistas. Para isso, ele utiliza 58.894 imagens FFDM oriundas de 39.272 pacientes, extraídas de um conjunto de dados privado. O pré-processamento das imagens envolveu a conversão para o tamanho 256x256 pixels e a normalização da base de dados. A arquitetura CNN empregada foi a *ResNet18*, pré-treinada no ImageNet, que demonstrou a mesma acurácia e maior rapidez em comparação com outros modelos testados pelos autores, como *ResNet-34* e *ResNet50*. O processo de aumento de dados foi realizado, e incluiu rotações entre 10 e -10 graus, além de rotações da imagem em 90, 180 e 270 graus. As imagens também foram invertidas horizontalmente e verticalmente. O modelo final obteve acurácia entre 76,6% e 77,0% nas quatro classificações BI-RADS®. Na implementação clínica, os autores obtiveram um Linear κ de 0,67 no conjunto de testes e no estudo com observadores, a evidenciar, assim, uma concordância entre o modelo e os radiologistas.

DENG et al. (2020) elaboraram um modelo de rede neural baseado em *Squeeze and Excitation attention* (SE-Attention) de forma a aumentar a acurácia e reduzir o tempo de inferência. Para isso, utilizaram uma base de imagens FFDM do *First Hospital of Shanxi Medical University*, composta por 18.157 imagens de mamografia de 4.982 pacientes. O pré-processamento dessas imagens envolveu o recorte do fundo e do músculo peitoral, além da melhoria das imagens a partir da transformação em escala de cinza, normalização das imagens e aumento dos dados por

meio de cortes e rotações aleatórias. O aspecto principal desse estudo é a arquitetura criada pelos autores, que combina uma CNN com o mecanismo SE-Attention, aplicado após as camadas de convolução. Esta arquitetura está representada na imagem Figura 3.2. Segundo os autores, as vantagens desta abordagem são melhorar a acurácia a partir da introdução de características relevantes em poucos parâmetros e reforçar a propagação das características obtidas. Implementaram o mecanismo de atenção em três CNNs diferentes, pré-treinadas no *ImageNet*: *Inception-V4*, *ResNeXt*, *DenseNet* para experimentação. Utilizaram validação cruzada em 10 dobras, reunindo os resultados após a inferência. O modelo que obteve mais acurácia foi o *Inception-V4*, com ACC de 92,17% e F1 médio de 90,33%. De acordo com os autores, isso demonstra que o mecanismo SE-Attention aumenta a acurácia geral do modelo. Ao analisar a performance por classe do BI-RADS®, observou-se que as classes A e D foram as menos performáticas. Os autores supõem que isso se deve à menor quantidade de imagens disponíveis para essas classes.

Figura 3.2: Framework desenvolvido por DENG et al. (2020). Nota-se que, após cada camada de convolução, os valores são copiados para um bloco de atenção e armazenados em um vetor global. Isso permite ter informações do modelo durante todas as etapas de processamento. Extraído e adaptado de DENG et al. (2020).



BUSALEH et al. (2022) propõem uma arquitetura para a classificação da densidade mamária a partir das duas visualizações Crânio-Caudal (CC) e Médio-Oblíquo Lateral (MLO) das mamografias de rastreamento, denominada *TwoViewDensityNet*. O modelo foi treinado a partir das bases de imagens *Digital Database for Screening Mammography* (DDSM) e *Integrated Breast Dataset* (InBreast). Os autores destacam que esse método apresenta uma performance superior em comparação com modelos de visualização única existentes no estado da arte. Para o treinamento do modelo proposto, foram utilizadas imagens com resolução de 336x224 pixels.

No pré-processamento, aplicou-se o filtro de magma, removeram-se os artefatos e isolou-se a mama, a fim de melhorar o aprendizado do modelo. A partir disso, as imagens são enviadas para arquitetura, que emprega duas CNNs como base para extrair as características das duas visualizações. Após a extração, uma camada *Global Average Pooling* (GAP) é inserida para reduzir as dimensões de saída e concatenar as informações obtidas. Foram avaliados três CNNs para a base da arquitetura: *ResNet50*, *EfficientNetB0* e *DenseNet201*. Esses modelos foram treinados no DDSM, e como resultado, o *ResNet50* apresentou o melhor desempenho. Também foram testadas três funções de perda: *Focal Loss* (FL LOSS), CE LOSS e *Sum of Squared Errors* (SSE LOSS). A FL LOSS gerou os melhores resultados. As métricas obtidas nas bases InBreast e DDSM com essa arquitetura demonstram uma acurácia de 96% e 95,83%, AUC de 97,44% e 99,51% e sensibilidade de 97,14% e 98,63%, respectivamente.

BIROŠ et al. (2024) avaliaram a performance de um sistema CAD baseado em DL. A arquitetura, denominada *DLAD*, realiza agregações dos pesos de múltiplos modelos *EfficientNet* independentemente treinados, a gerar um modelo único que incorpora todas as características e configurações relevantes. Os autores utilizaram um conjunto de dados composto por 5.130 estudos de mamografia, a totalizar 20.520 imagens para o treinamento do modelo. Além disso, separaram um conjunto de 122 estudos de mamografias, com 488 imagens obtidas de três instalações de oncologia diferentes, que foram enviados a radiologistas para criar um *ground-truth* dos dados, a fim de avaliar o modelo. Os resultados obtidos pela CNN nesse conjunto mostram um Cohen's κ de 0,708, F1 de 0,798 e ACC de 0,819. A ACC dos radiologistas variou, sendo a maior 0,875 com um Cohen's κ de 0,8, a demonstrar a semelhança de classificação entre o modelo e o radiologista. Um importante resultado foi a dificuldade do modelo em classificar imagens com lesões malignas significativas, artefatos de metal, clipes ou outros objetos presentes na imagem. No entanto, o modelo demonstrou uma performance semelhante a de um profissional. Além disso, os autores mencionam que a origem diversa dos dados permitiu uma maior generalização do modelo, devido à diversidade das imagens.

NGUYEN et al. (2021) desenvolveram um modelo de multi-classificação e multi-visualização de DL para classificar o câncer de mama e a densidade de acordo com o BI-RADS®, demonstrando também que esse método supera o mesmo modelo com apenas uma visualização. Para isso, utilizaram as bases de dados *VinDr-Mammo* (VinDr), DDSM e uma base coletada do *Hanoi Medical University Hospital*, que contém 8.509 estudos de imagens FFDM. No pré-processamento das imagens, foi realizado o recorte da mama para remover a área do fundo, utilizando o *You Only Look Once Version 5* (YOLOv5). Para a extração das características de cada visualização, as CNNs foram implementadas sem a camada totalmente conectada ao final da arquitetura, utilizando o vetor de características dessa camada em um classificador *LightGBM*, que recebe uma média dos vetores. O modelo escolhido para os quatro *backbones* foi o *EfficientNet-B2*, que superou o outro modelo testado pelos autores, o *ResNet50*. As imagens foram redimensionadas para 512x512 pixels e divididas em três canais de 32x24x1408 pixels. Foi utilizada a CE LOSS para calcular o erro, e a métrica F1 foi empregada para as classifica-

ções BI-RADS®. O modelo multi-visualização superou o modelo de única visualização no F1, obtendo 61,65%, enquanto o outro modelo obteve 6% a menos, demonstrando a vantagem de utilizar as quatro imagens da mamografia.

PAWAR et al. (2022) investigam o uso de modelos com arquitetura multi-camadas com *DenseNet-121*. O estudo utilizou a base de imagens DDSM, selecionando 200 imagens de cada visualização e lateralidade da mama para o treinamento, a totalizar 800 imagens de mamografia. A etapa seguinte envolveu a remoção do músculo peitoral e dos artefatos por meio de um algoritmo de *Depth First Search* (DFS), que remove áreas de alto contraste na imagem. Em seguida, aplicaram o *Contrast Limited Adaptive Histogram Equalization* (CLAHE) para melhorar o contraste das imagens. Na arquitetura do modelo, usaram quatro *DenseNet-121*, cada uma responsável por aprender as características de uma visualização e lateralidade. Após as camadas densas e de convolução, fizeram a concatenação das camadas para uma camada totalmente conectada e uma camada de classificação. Foi utilizado o algoritmo *Stochastic Gradient Descend* (SGD) para otimização, e a função de perda foi a CE LOSS. Com essa abordagem, os autores alcançaram um *Overall AUC* (OAUC) de 0,9625 e um F1 de 0,886. Comparando com outros estudos, os autores comentam que a performance da arquitetura multi-camadas e multi-visualização mostrou-se mais confiável do que a classificação utilizando apenas uma visualização.

MATTHEWS et al. (2021) propõem treinar um modelo de DL com imagens FFDM para depois utilizar os pesos para o *fine-tuning* do modelo em imagens de mamografias 2D sintéticas. As imagens FFDM foram obtidas de um centro médico acadêmico localizado em *Midwestern US*, consistindo de 187.627 exames. Para o primeiro treinamento do modelo, foi utilizada a arquitetura *ResNet34* pré-ativada, na qual as camadas de normalização de lotes foram substituídas por normalizações em grupo. As imagens foram redimensionadas para 416x320 e normalizadas no intervalo [0,1]. Cada lote de treinamento foi balanceado de forma que as classes B e C do BI-RADS® fossem quatro vezes mais frequentes do que as classes A e D. Houve aumento dos dados por meio de inversões horizontais e verticais. O modelo obteve 82,2% de acurácia, 0,75 de Cohen's κ e 0,952 de OAUC, mostrando-se competitivo em comparação com outros modelos analisados pelo autor na classificação das imagens FFDM.

MAGHSOUDI et al. (2021) apresenta um algoritmo para classificação de densidade baseado na geração de *superpixels* e aprendizado de máquina radiométrica, a utilizar um conjunto de dados de imagens FFDMs considerado "padrão de ouro" pelos radiologistas. Foram utilizados seis conjuntos de dados diferentes, cada um com uma função específica, obtidos pelo *Hospital of the University of Pennsylvania* e *Mayo Clinic*. Neste trabalho, apenas três conjuntos são relevantes. O primeiro conjunto consistiu em 11.200 imagens bilaterais, utilizadas para o treinamento e validação de um modelo baseado na *UNet*, responsável por remover o fundo das imagens. O segundo conjunto consistiu em 1.100 imagens MLO, e foi utilizado para treinar um modelo baseado na *UNet* na tarefa de segmentação binária do músculo peitoral. Após essas dois conjuntos, o terceiro consistiu em 3.314 imagens bilateral CC, da base de imagens de *Mayo*.

Clinic, que continha o "padrão de ouro", foi usado para treinar uma SVM após realizar uma série de cálculos e pré-processamentos. Para esse último algoritmo, a mama foi particionada em *superpixels* utilizando valores de intensidade, e então aplicaram-se características radiométricas nas imagens, que foram enviadas para o modelo para classificar a mama como densa ou não densa. A avaliação final dos modelos indicou que os resultados possuem forte correlação com a análise de um leitor experiente, de acordo com os autores.

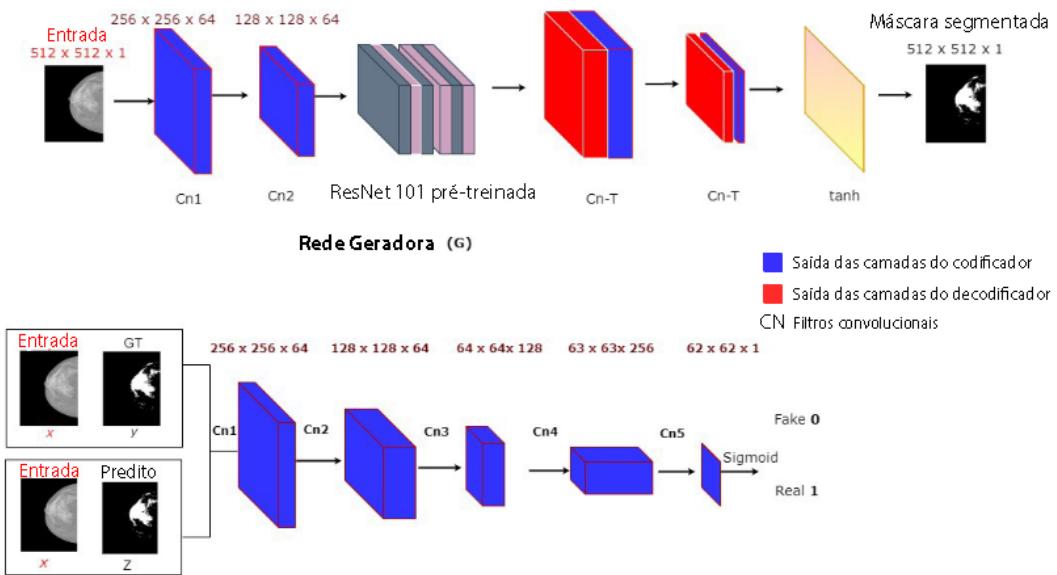
CHAN; HELVIE (2019) abordam um modelo de DL utilizando a arquitetura *ResNet*, treinado em um conjunto de 41.000 imagens de mamografias provenientes de 27.684 pacientes. Os resultados demonstraram que o modelo concordou com a avaliação do radiologista em 77% dos casos, com um Cohen's κ de 0,67.

COGAN; TAMIL (2020) arquitetam um modelo chamado *DualViewNet*, que infere a densidade mamária com base na visualização das mamografias, ao invés de considerar a densidade por imagem ou por paciente. O autor conjectura que imagens com visualização MLO são mais adequadas para o modelo aprender, permitindo uma melhor captura das características específicas de cada visão. A base de dados utilizada foi a *Curated Breast Imaging Subset of DDSM* (CBIS-DDSM), e as imagens passaram por um pré-processamento que incluiu filtragem com o filtro magma e o redimensionamento para 336x224 pixels. Houve o aumento das imagens a partir de recortes, rotações e inversões aleatórias para mitigar o *overfitting*. Por fim, os pixels foram normalizados utilizando os valores do ImageNet. A arquitetura proposta envolve o uso de duas redes neurais separadas, uma para as imagens MLO e outra para as imagens CC. O autor optou pela rede *MobileNetV2* para fazer a classificação, embora ressalte que outras redes do estado da arte poderiam ser utilizadas. O resultado do *DualViewNet* foi ligeiramente superior ao modelo de visualização única, atingindo *Macro AUC* (M-AUC) de 0,897.

SAFFARI et al. (2020) desenvolvem uma arquitetura que combina uma *conditional GAN* com uma CNN para classificação da densidade mamária. A *conditional GAN* é empregada para gerar uma máscara binária do tecido denso nas mamografias, utilizada posteriormente no treinamento da CNN. O conjunto de dados selecionado para treinamento e validação do modelo foi o InBreast. A primeira etapa do processo envolve a remoção do músculo peitoral e do fundo das imagens de mamografia, utilizando um algoritmo de crescimento de região, seguido pelo redimensionamento das imagens para 512x512 pixels. Essas imagens são então processadas pelo modelo gerativo para gerar a máscara binária, conforme ilustrado na Figura 3.3. No modelo de classificação, foi utilizado o WCE LOSS devido ao desbalanceamento do conjunto de dados InBreast. O otimizador escolhido foi o *Root Mean Square Propagation* (RMSProp). O desempenho do modelo foi avaliado em conjuntos de dados平衡ados e desbalanceados, alcançando a melhor acurácia de 98,75% no conjunto de dados balanceado.

YI et al. (2019) implementa um sistema de DL que classifica a lateralidade, visualização e densidade de mamografias digitais (FFDM). Para isso, utilizou-se o conjunto de dados DDSM. As imagens foram convertidas em *Portable Network Graphics* (PNG) e redimensionadas para 256x256 pixels. Além disso, artefatos nas imagens foram removidos durante o

Figura 3.3: Arquitetura da cGAN. A arquitetura de cima é a etapa do *encoder*, que irá produzir uma máscara binária da região da densidade da imagem de mamografia. A segunda arquitetura compreende a arquitetura do *decoder*, que irá utilizar camadas de convolução para extrair as características da máscara e retornar a densidade da imagem. Extraído e adaptado de SAFFARI et al. (2020).



pré-processamento. O modelo escolhido foi uma *Deep Convolutional Neural Network* (DCNN) baseada na arquitetura *ResNet-50*, otimizado com o algoritmo SGD. Ademais, produziram mapas de calor para indicar a lateralidade, a partir do algoritmo *Class Activation Maps* (CAM), a fim de estudar as partes mais impactantes escontradas pelo modelo DCNNs. No entanto, o modelo de classificação de densidade mamária não alcançou métricas satisfatórias, com ACC de 68%, sensitividade de 90% e especificidade de 53%.

LIZZI et al. (2022) utilizam uma CNN residual para classificação de densidade a partir de imagens de mamografia digital (FFDM). Os autores utilizaram uma base de dados privada com 6648 imagens de 1662 pacientes para o treinamento. O pré-processamento das imagens incluiu a conversão de JPEG sem perdas para PNG de 8 bits, recorte da mama para remoção do fundo preto e exclusão do músculo peitoral nas imagens com visualização MLO. As imagens foram invertidas verticalmente para manter uma orientação única, e um filtro Gaussiano foi aplicado para reduzir ruídos. Além disso, utilizou-se técnicas de *data augmentation*, com *zooms* aleatórios de 20% e rotações aleatórias de 10° graus. A arquitetura da rede utilizada foi a implementada em HE et al. (2015), e o treinamento foi realizado em quatro modelos distintos, um para cada visualização. A classificação final para cada mama, direita ou esquerda, foi determinada pela maior classificação atribuída pelos modelos. Os autores treinaram o modelo em três distribuições diferentes de dados. A melhor foi a definida pelo Atlas BI-RADS®, com (A: 10%, B: 40%, C: 40%, D: 10%), alcançando 87,9% de precisão e 80,1% de recall. A pior distribuição foi a uniforme, com 78% de precisão e 77,8% de recall. A terceira distribuição utilizou a base

de dados completa, porém obteve um resultado quase semelhante ao da primeira distribuição. Adicionalmente, o modelo foi treinado com e sem a presença do músculo peitoral nas imagens. O melhor desempenho foi obtido com o modelo treinado sem o músculo peitoral, indicando que sua remoção exerce uma influência positiva na classificação da densidade mamária.

LI et al. (2020) aprimoraram um modelo de classificação de densidade mamária com duas visualizações com aprendizagem residual dilatada e guiado por atenção. Para isso, foi utilizada uma base de dados privada contendo 1985 imagens de 500 pacientes, e o InBreast foi escolhido para o conjunto de testes da validação do modelo. O pré-processamento das imagens incluiu recorte da mama, redimensionamento para 224x224 pixels e técnicas de *data augmentation*, com inversões horizontais e verticais, além de rotações entre -90 e 90 graus. O treinamento foi realizado em validação cruzada com 5 dobras. Ao aplicar a aprendizagem dilatada, as camadas de GAP tiveram sua resolução ajustada para 1/8, ao invés de 1/32, e um bloco de atenção foi introduzido para destacar as características importantes das diferentes camadas. Isso aumentou o campo receptivo da rede sem perder a resolução da imagem. O otimizador Adam foi utilizado para o treinamento em 30 épocas, e a ResNet50 foi empregada como a CNN base para adicionar as técnicas de aprendizagem dilatada e guiada por atenção. Após o treinamento, o modelo apresentou melhor desempenho ao incorporar essas duas estruturas. Nos experimentos, foi notado que as imagens de visualização CC foram mais importantes do que as imagens de visualização MLO, dado que o modelo treinado com duas imagens CC teve desempenho superior. Além disso, uma observação dos autores é que um modelo com duas visualizações contém informações complementares que podem ser aproveitadas para melhorar ainda mais a performance do modelo.

ZHAO et al. (2021) implementam uma arquitetura com CNN com o objetivo de aprimorar a detecção de características dimensionais e espaciais das mamografias por meio de um novo modelo bilateral denominado BASCNet. Este modelo foi desenvolvido para discriminar características espaciais e foi comparado com soluções unilaterais. Para isso, os autores utilizaram o DDSM para treinamento do modelo, e o validaram no InBreast. O pré-processamento das imagens incluiu a remoção da parte escura das mamografias através do isolamento da mama e o redimensionamento das imagens para 224x224 pixels. Além disso, técnicas de *data augmentation* foram utilizadas para evitar *overfitting*, com rotações de 90° em 90° graus e inversões horizontais nas imagens. A arquitetura criada pelos autores, *BASCNet*, foi criada devido a necessidade de um campo receptivo maior nos modelos, uma vez que foi identificado que a informação espacial global das glândulas mamárias é fundamental para distinguir a densidade. Para isso, usaram a ResNet50 como base para essa arquitetura. Assim, os autores criaram dois módulos, denominados *Adaptive Channel Attention Module* (ACAM), que explora a independência dos canais da imagem para identificar os canais mais relevantes para a classificação, e o *Adaptive Spatial Attention Module* (ASAM), que captura informações globais na dimensão espacial e se concentra nas informações discriminantes por meio de uma estrutura de atenção. Após o treinamento do modelo, as métricas obtidas superaram os modelos anteriores treinados

no InBreast e no DDSM. Os autores também observaram, após os experimentos, que a profundidade de uma rede convolucional não resulta necessariamente em uma melhoria significativa de performance. Também foi observado que as imagens de visualização CC proporcionavam um melhor treinamento dos modelos em comparação com as imagens de visualização MLO.

3.6 Conclusão dos resultados

A partir das análises realizadas nesta revisão da literatura, observou-se um número significativo de estudos que empregam técnicas de DL e CNNs para a classificação da densidade mamária. No entanto, há uma grande variabilidade nas técnicas de pré-processamento de imagens e nas arquiteturas dos modelos utilizados, com poucos pontos em comum entre os trabalhos analisados. Além disso, a maioria dos estudos não explora de forma abrangente diferentes métodos e combinações de técnicas para aprimorar o desempenho dos modelos de classificação. Essa diversidade no tratamento das imagens e nas abordagens arquiteturais evidencia uma lacuna na literatura, abrindo espaço para pesquisas que investiguem novas combinações de técnicas e métodos visando superar os resultados atuais.

Diante dessa lacuna, este trabalho propõe uma abordagem multiclasse para a classificação da densidade mamária conforme o sistema BI-RADS®, explorando uma variedade de técnicas de pré-processamento, arquiteturas de CNNs e ajustes de hiperparâmetros. O objetivo é identificar as melhores combinações de métodos e, assim, buscar superar o estado da arte na classificação de densidade mamária. Além disso, serão investigados modelos e técnicas ainda não explorados na literatura, contribuindo com inovações para a área.

Um dos diferenciais deste trabalho é a adoção de uma abordagem multiclasse inovadora. Serão desenvolvidos quatro modelos binários independentes, cada um treinado para identificar uma categoria do BI-RADS® como positiva, enquanto as demais são consideradas negativas. Durante a inferência, uma imagem FFDM será analisada pelos quatro modelos, e a classificação final será determinada por uma camada adicional de decisão. Essa estratégia tem o potencial de superar métodos convencionais, proporcionando uma nova perspectiva para a classificação da densidade mamária.

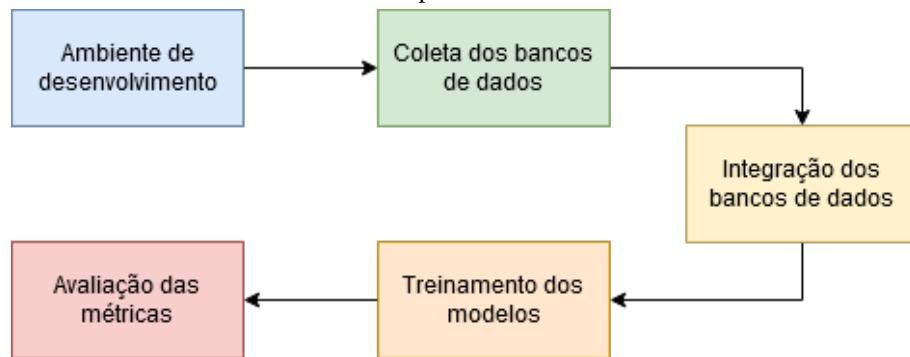
4

Metodologia

4.1 Organização do trabalho

A metodologia deste trabalho foi realizada em cinco etapas: configuração do ambiente de desenvolvimento, coleta dos bancos de dados, integração dos dados, treinamento das CNNs e avaliação do desempenho do modelo, como visto na Figura 4.1.

Figura 4.1: Fluxograma da metodologia. As três primeiras etapas focam na infraestrutura e preparação dos dados, enquanto as duas últimas concentram-se na realização e avaliação dos experimentos.



4.2 Ambiente de desenvolvimento

Para a execução deste trabalho, foi configurado um ambiente computacional baseado em Linux, dedicado ao armazenamento, processamento dos dados e treinamento dos modelos. A infraestrutura utilizada consiste em uma máquina equipada com um processador AMD EPYC 7713 64-Core, um SSD NVMe de 1TB para armazenamento de alta velocidade, 256GB de memória RAM e três placas de vídeo NVIDIA RTX A4500 com 20GB de VRAM cada, garantindo capacidade suficiente para processamento paralelo e manipulação de uma grande quantidade de dados. O sistema operacional utilizado foi o Ubuntu 22.04 LTS, escolhido por sua estabilidade e suporte a ferramentas de desenvolvimento em ML.

A implementação foi realizada em Python 3.12, utilizando o framework PyTorch para o treinamento e avaliação dos modelos de DL. Para auxiliar no pré-processamento e manipulação dos dados, foram empregadas as bibliotecas opencv-python, pandas, pydicom e NumPy, enquanto

a divisão dos dados em conjuntos de treinamento, validação e teste foi realizada com o auxílio da biblioteca Scikit-learn. Adicionalmente, foram utilizadas outras ferramentas, como Matplotlib e Seaborn, para visualização de dados e análise dos resultados.

4.3 Coleta dos bancos de dados

Na etapa de coleta dos conjuntos de dados para treinamento e validação dos modelos, priorizou-se a seleção de bases públicas e anônimas mencionadas na revisão da literatura. Inicialmente, foram escolhidas as bases VinDr e InBreast. A base DDSM, também citada em diversos artigos, foi considerada, mas, ao verificar sua disponibilidade, identificou-se que o site oficial poderia retornar imagens com erros devido ao uso de software desatualizado. Como alternativa, optou-se pela base *Mini Digital Database for Screening Mammography* (Mini-DDSM), uma versão atualizada e mais acessível do DDSM.

Para ampliar a generalização e o aprendizado dos modelos, buscou-se na literatura bases de dados públicas e anônimas adicionais, que contenham informações sobre a densidade mamária. Dessa forma, foram selecionadas as bases *Radiological Society of North America* (RSNA) (CARR et al., 2022), originária de uma competição de classificação de câncer de mama, e *Breast Micro-Calcifications Dataset* (BMCD), que, embora focada na classificação de microcalcificações, também inclui dados sobre a densidade. Outras bases públicas, como o *OPTIMAM* (HALLING-BROWN et al., 2021), foram consideradas. No entanto, seu uso foi desconsiderado pelas limitações de *hardware*, como espaço de armazenamento insuficiente e o tempo necessário para realizar o *fine-tuning* dos modelos, que poderiam ultrapassar o prazo do projeto devido ao enorme volume de dados.

Portanto, as bases selecionadas para este trabalho foram BMCD, InBreast, Mini-DDSM, RSNA e VinDr. Todas reúnem imagens anônimas e públicas de mamografias, ou seja, que passaram por um rigoroso processo de anonimidade, garantindo a preservação da privacidade e a confidencialidade dos dados dos pacientes, em conformidade com as normas éticas e legais vigentes.

4.3.1 BMCD

O conjunto de dados BMCD (LOIZIDOU et al., 2021) é composto por 100 pares de mamografias, coletadas em dois momentos temporais sequenciais. O conjunto inclui as mamografias anteriores e recentes de cada paciente, tanto na vista CC quanto na vista MLO. Trata-se de um banco de dados completo para a detecção e classificação de microcalcificações mamárias de acordo com o sistema BI-RADS®, utilizando mamografias digitais, que também inclui a classificação da densidade mamária. A coleta desse banco foi realizada por meio do endereço eletrônico *Zenodo*.

4.3.2 InBreast

O conjunto de dados InBreast (MOREIRA et al., 2012) é uma base pública de imagens FFDM. Esse banco de dados é composto por 115 exames, totalizando 410 imagens. Desses exames, 90 correspondem a mulheres com ambas as mamas afetadas por câncer (quatro imagens por exame), enquanto 25 exames são de pacientes mastectomizadas (duas imagens por exame). A base de dados inclui diversos tipos de anotações, que incluem a densidade, câncer, lesões, calcificações, assimetrias e distorções. Além disso, fornece contornos precisos das lesões, anotados por especialistas. Neste trabalho, foi utilizado apenas as informações da densidade. A coleta desse banco foi realizada por meio do endereço eletrônico *Kaggle largest AI and ML community* (Kaggle).

4.3.3 MiniDDSM

O Mini-DDSM (LEKAMLAJE et al., 2020) é uma versão leve e atualizada do popular conjunto de dados DDSM, que atualmente está obsoleto. Essa base surgiu a partir do interesse de manter a base original acessível na web, visto que a base de dados original, mantida pela Universidade do Sul da Flórida, possui imagens compactadas em formato *JPEG* sem perdas, geradas por meio de um software desatualizado ou com problemas, conforme descrito no site do DDSM. A criação do Mini-DDSM demandou um esforço significativo em termos de tempo, codificação e poder de processamento para organizar e preparar os dados. A coleta desse banco foi realizada por meio do endereço eletrônico Kaggle.

4.3.4 RSNA

O conjunto de dados RSNA (CARR et al., 2022) é uma base pública de imagens FFDM, compilada para uma competição de predição de câncer de mama na plataforma *Kaggle*. O objetivo da competição era identificar casos de câncer de mama em mamografias provenientes de exames de rastreamento. Muitas das imagens contêm anotações sobre a densidade mamária, o que torna o conjunto de dados particularmente útil para estudos relacionados à classificação de densidade e diagnóstico assistido por computador. A coleta desse banco foi realizada por meio do endereço eletrônico Kaggle.

4.3.5 VinDr

O conjunto de dados VinDr (PHAM; NGUYEN TRUNG; NGUYEN, 2022) é uma base de imagens FFDM em larga escala, projetada para o desenvolvimento e avaliação de algoritmos voltados à classificação de câncer de mama e à avaliação da densidade mamária de acordo com o sistema BI-RADS®. O conjunto disponibiliza 1.000 exames teste e 4.000 para treinamento modelos, em que cada exame possui 4 imagens. Além disso, o conjunto de dados também pode ser utilizado para outras tarefas em imagens médicas e visão computacional em

geral. Em relação à divisão entre treinamento e teste, diferentes partições podem ser criadas conforme a necessidade, uma vez que todo o conjunto de dados foi gerado por meio de um único procedimento padronizado. Essa flexibilidade permite que pesquisadores adaptem o uso do conjunto de dados a diversos cenários e objetivos de pesquisa. A coleta desse banco foi realizada por meio da plataforma *Physionet* (GOLDBERGER et al., 2000).

4.4 Integração dos dados

Após a coleta dos dados, foi necessária uma etapa de pré-processamento para normalizar os metadados associados a cada imagem, converter as imagens no formato DICOM para PNG e aplicar filtros e recortes visando uma padronização comum entre os conjuntos de dados.

A normalização dos metadados consistiu na remoção de informações irrelevantes para o trabalho, como detalhes sobre o equipamento utilizado para captura das mamografias ou dados demográficos dos pacientes. Além disso, foram excluídos casos em que as mamas apresentavam implantes, pois esses poderiam introduzir ruídos e prejudicar o aprendizado do modelo. Ao final desse processo, os arquivos de metadados continham apenas as informações essenciais para associar a densidade mamária às respectivas imagens.

As imagens no formato DICOM foram convertidas para PNG em 8-bits e em escala de cinza, garantindo compatibilidade com as bibliotecas do Python utilizadas para processamento e maior eficiência no carregamento dos dados durante o treinamento dos modelos.

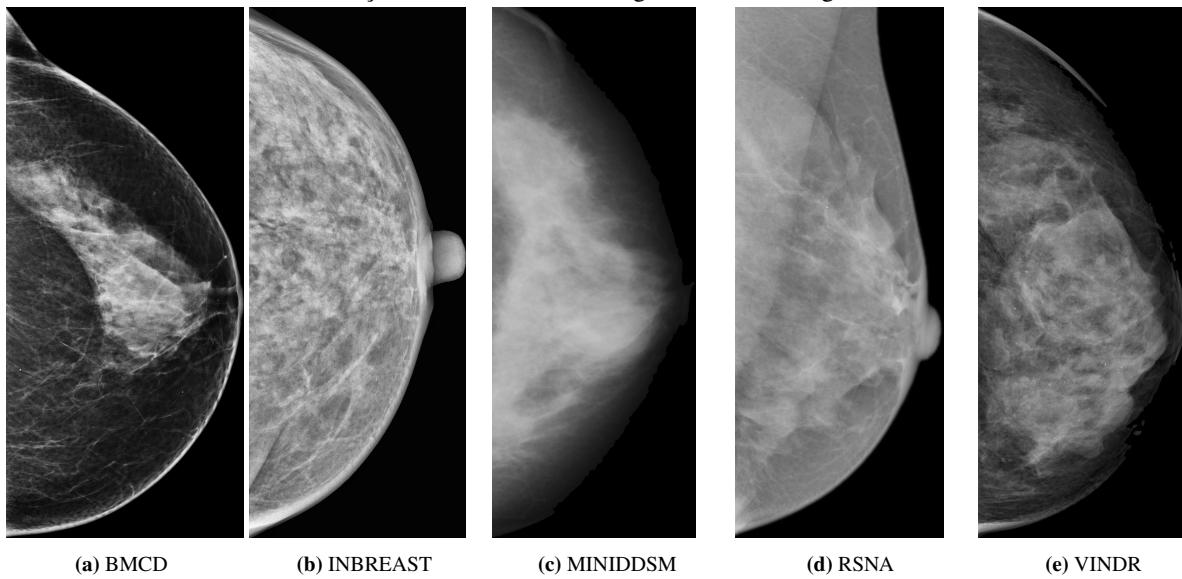
Para uniformizar as imagens, foi realizado um recorte automático com o objetivo de remover as regiões escuras ao redor da mama, focando apenas na área de interesse. Para isso, utilizou-se um método de computação gráfica, que consiste no uso de operações morfológicas para isolar a mama (MINA; ISA, 2015). Adicionalmente, todas as imagens foram padronizadas para a lateralidade esquerda, sendo espelhadas horizontalmente quando necessário. Por fim, foram aplicados os seguintes procedimentos específicos para cada base de dados:

- **BMCD:** As imagens apresentam bom contraste, portanto, nenhum pré-processamento adicional foi necessário.
- **InBreast:** Foi aplicado o filtro CLAHE para realçar o contraste das imagens, conforme recomendado por HUANG; LIN (2020).
- **MiniDDSM:** Como as imagens já estavam no formato PNG, a conversão de DICOM não foi necessária. No entanto, devido à presença de artefatos e bordas brancas que poderiam interferir na classificação, foi realizado um recorte das bordas brancas para eliminar essas interferências, e os artefatos foram removidos pelo recorte automática da mama.
- **RSNA:** A técnica de *windowing*, disponível nos metadados do DICOM, foi aplicada para melhorar o contraste das imagens.

- **VinDr:** O mesmo procedimento de *windowing* utilizado na base RSNA foi aplicado a essa base.

O resultado do pré-processamento realizado pode ser encontrado na Figura 4.2.

Figura 4.2: Imagens resultantes após o processamento. Mesmo após a integração, ainda é notável uma diferença nos contrastes das imagens, devido a origem dos dados.



4.4.1 Análise estatística dos dados

No total, foram obtidas 56.510 imagens de mamografia, cuja distribuição entre as classes de densidade mamária está detalhada na Figura 4.3 e na Figura 4.4. A Classe C é a mais representada, com 29.081 imagens, correspondendo a aproximadamente 51,5% do conjunto de dados. Em seguida, a Classe B conta com 17.468 imagens, representando cerca de 30,9% do total. As Classes A e D, por sua vez, possuem 4.362 e 5.599 imagens, respectivamente, correspondendo a 7,7% e 9,9% do conjunto. Essa distribuição reflete a realidade clínica, uma vez que a maioria dos pacientes tende a apresentar mamas com densidade nas Classes B e C, as mais frequentes na população.

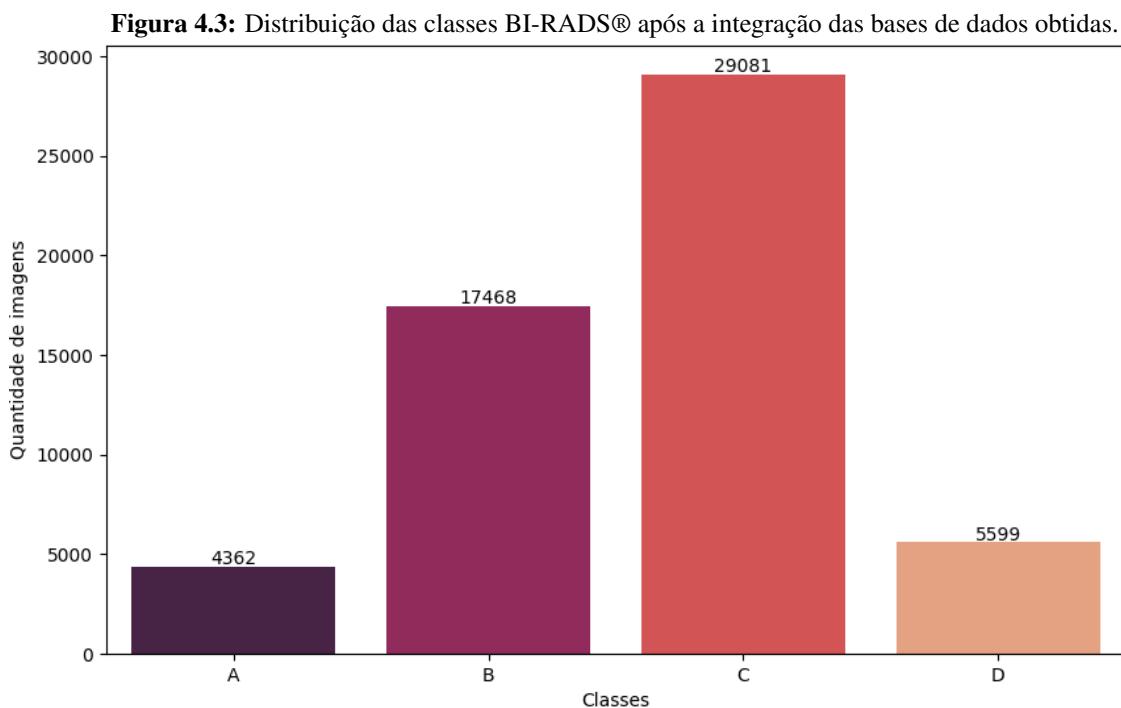
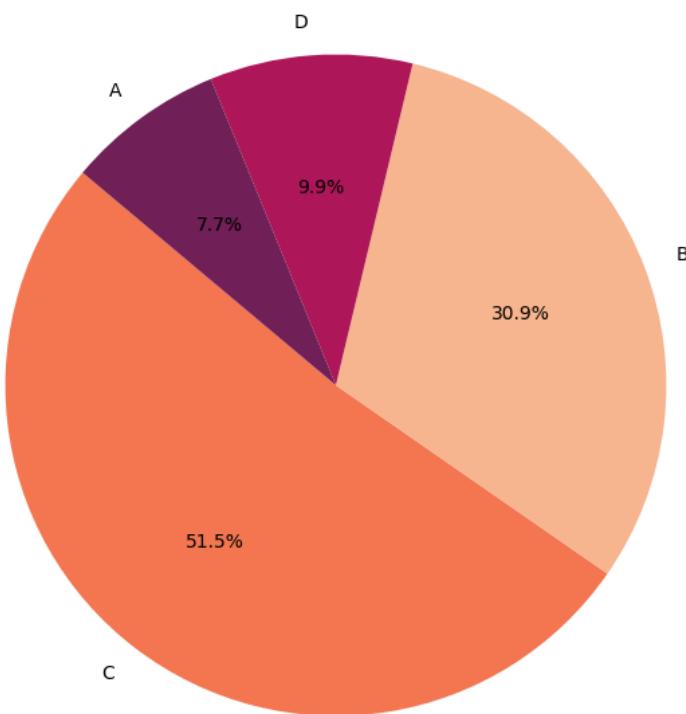


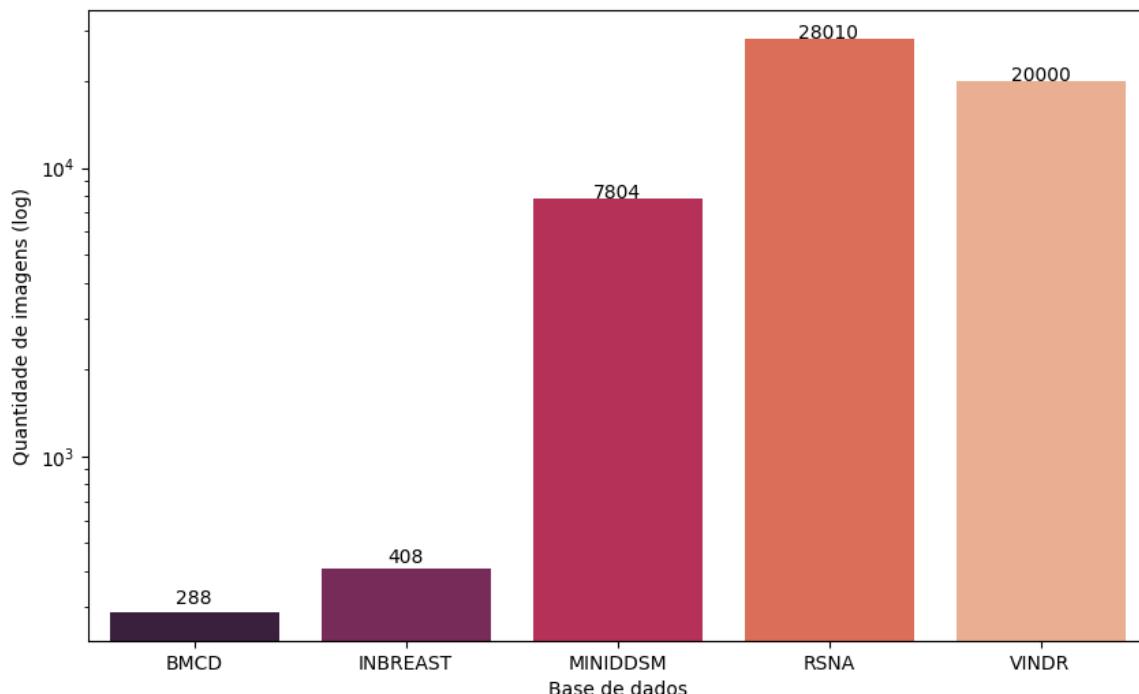
Figura 4.4: Porcentagem da distribuição de cada classe na base de dados



Também é importante destacar que algumas bases de dados sofreram uma redução significativa no número de imagens utilizadas. Um exemplo é a base de dados RSNA, que possuía mais de 50 mil imagens, porém teve uma redução para 28.010 imagens, dado que apenas

essas possuíam a classificação da densidade mamária e não tinham implantes. A quantidade final de imagens utilizadas em cada base de dados pode ser visualizada na Figura 4.5.

Figura 4.5: Quantidade de imagens obtidas em cada base de dados, após aplicar os filtros e processamentos, em escala logarítmica.



Vale destacar que em todas bases deste trabalho houve uma distribuição desbalanceada das classificações BI-RADS®, como pode ser observado na Figura 4.6 e na Figura 4.7. Essa distribuição desigual pode introduzir viés no treinamento do modelo, favorecendo classes com maior número de dados, como a classe B e C, o que pode dificultar o aprendizado das outras classes, como A e D. Assim, foram adotadas técnicas de balanceamento para mitigar esse problema.

Figura 4.6: Gráfico da contribuição de cada base de dados na quantidade e classe das imagens em escala logarítmica.

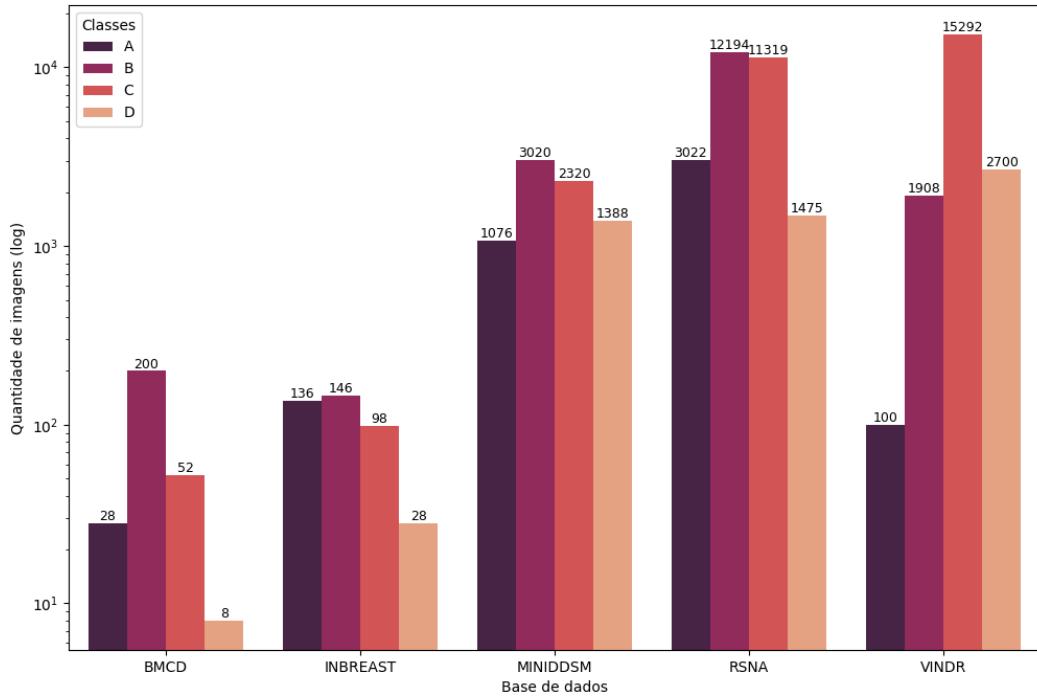
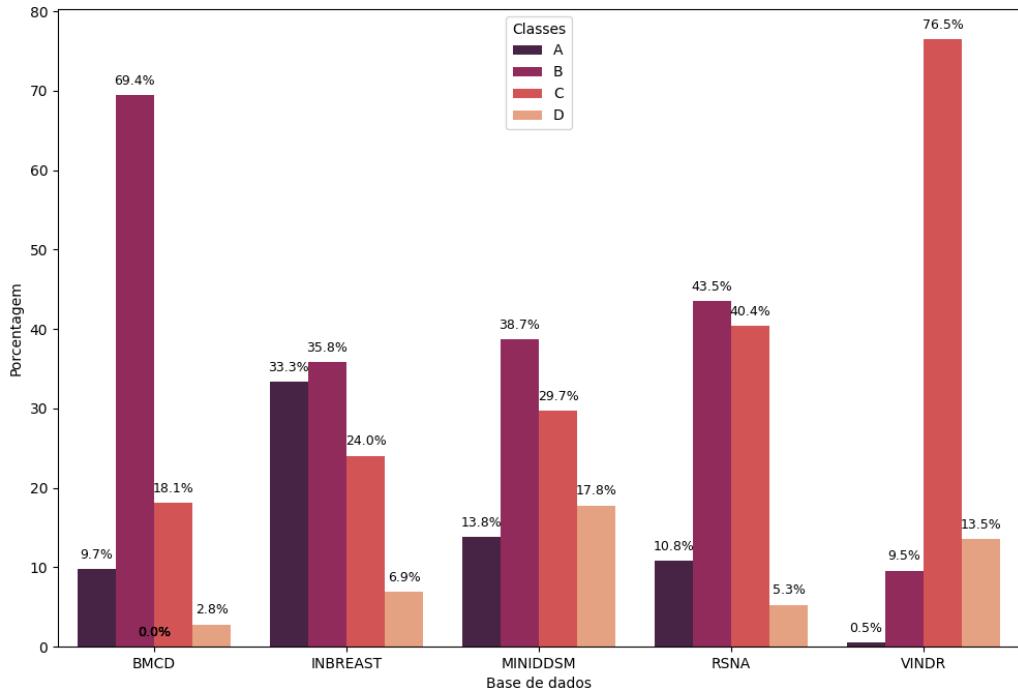


Figura 4.7: Proporção das classes em cada base de dados.

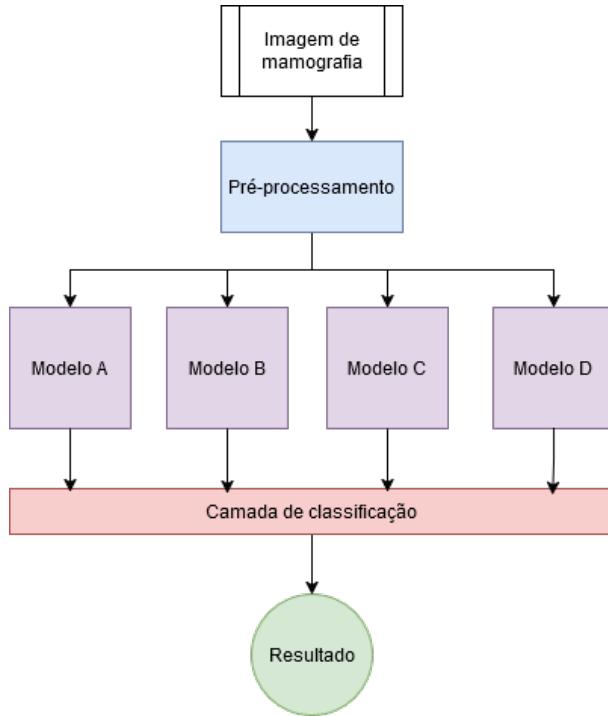


4.5 Treinamento das redes

Para o treinamento das redes neurais, foram exploradas duas abordagens distintas para a classificação da densidade mamária. A primeira abordagem, ilustrada na Figura 4.8, propõe uma arquitetura que utiliza quatro modelos binários independentes. A segunda abordagem consiste

em treinar um único modelo multiclasse capaz de classificar diretamente a imagem em uma das quatro categorias do sistema BI-RADS®.

Figura 4.8: Arquitetura baseada em modelos binários. A imagem de entrada é processada por quatro modelos binários independentes, cada um treinado para identificar uma classe específica (positiva) enquanto agrupa as demais como negativas. As saídas dos modelos são então enviadas para uma camada de classificação, que aplica uma lógica para determinar a classe final da imagem.



Na abordagem com modelos binários, cada modelo é especializado em identificar uma categoria específica do BI-RADS®. Durante o treinamento de cada modelo, a classe positiva corresponde à categoria específica que ele representa, enquanto as outras três categorias são agrupadas como classe negativa. No pipeline de inferência, uma imagem FFDM é processada simultaneamente pelos quatro modelos, e suas saídas são enviadas para uma camada de classificação responsável por determinar a densidade mamária final.

Nos experimentos iniciais com a abordagem de modelos binários, a camada de classificação utilizou uma lógica simples: a classe final é determinada pelo modelo que produziu a maior pontuação. Esta escolha considera que a mama pode apresentar regiões heterogêneas, com áreas de densidade variável. Por exemplo, mesmo que a maior parte da mama seja classificada como lipossubstituída (categoria A), a presença de uma única região mais densa já é suficiente para elevar a classificação geral da mama, refletindo a complexidade e a variabilidade da densidade mamária observada na prática clínica.

Nos experimentos seguintes, implementou-se uma lógica de classificação diferente baseada em limiares. Nessa abordagem, cada modelo binário possui um valor limiar específico que determina a fronteira entre as classes positiva e negativa. Para determinar a classificação final, calcula-se a diferença entre o valor predito por cada modelo e seu respectivo limiar. A

classe correspondente ao modelo que apresentar a maior diferença em relação ao seu limiar é selecionada como a classificação final da densidade mamária.

A segunda abordagem investigada utiliza um único modelo de classificação multiclasse, onde a rede neural é treinada para classificar diretamente a imagem FFDM em uma das quatro categorias de densidade mamária do sistema BI-RADS®. Nesta arquitetura, a rede processa a imagem de entrada e produz um vetor de quatro probabilidades na sua camada de saída, uma para cada categoria. A categoria com a maior probabilidade é selecionada como a classificação final da densidade mamária. Esta abordagem é mais direta e requer o treinamento de apenas um modelo, porém exige que a rede aprenda simultaneamente as características distintivas de todas as classes e suas relações.

4.5.1 Heurística de treinamento

A abordagem adotada para o treinamento dos modelos baseou-se em uma heurística empírica e sequencial, focada na experimentação iterativa dos hiperparâmetros. Essa estratégia consistiu em testar diferentes configurações em rodadas sucessivas de experimentos, ajustando um parâmetro por vez com base nos resultados obtidos em execuções anteriores. Dessa forma, priorizou-se a otimização dos hiperparâmetros com maior impacto teórico no desempenho, permitindo uma análise detalhada da influência de cada escolha no aprendizado dos modelos.

Para garantir a robustez e a generalização dos resultados, foi empregada a técnica de validação cruzada com *k-fold*, na qual o conjunto de dados foi dividido em múltiplas partições de treinamento e validação. Essa metodologia reduz o risco de *overfitting* ao assegurar que o modelo seja avaliado em diferentes subdivisões dos dados, fornecendo uma estimativa mais confiável de seu desempenho. assim, foram exploradas duas estratégias distintas para a obtenção das previsões finais:

- **Média das previsões dos *k-folds*:** As previsões geradas por todos os modelos treinados nas diferentes iterações do *k-fold* foram combinadas por meio de uma média aritmética. Essa abordagem visa aproveitar a diversidade dos modelos treinados, resultando em previsões mais estáveis e reduzindo a variância do modelo final.
- **Seleção do melhor modelo do *k-fold*:** Alternativamente, utilizou-se apenas o modelo que obteve o melhor desempenho entre todas as iterações do *k-fold*, cujas previsões foram adotadas diretamente para a classificação. Essa estratégia simplifica o processo de inferência, pois elimina a necessidade de armazenar múltiplos modelos e calcular médias, embora possa não capturar toda a diversidade do aprendizado obtido ao longo do *k-fold*.

A combinação dessas estratégias possibilitou uma avaliação mais precisa do impacto dos hiperparâmetros e da estabilidade dos modelos, garantindo um equilíbrio entre desempenho, robustez e eficiência computacional.

4.5.2 Arquiteturas das redes neurais

As CNNs utilizadas para o treinamento foram selecionadas com base em seu uso na literatura relacionada à classificação de densidade mamária e em seu excelente desempenho em tarefas de classificação de imagens. A escolha das arquiteturas levou em consideração suas características técnicas, eficiência computacional e capacidade de generalização. As arquiteturas selecionadas foram:

- **Residual Networks (ResNets):** Essas redes são amplamente reconhecidas por introduzir conexões residuais, que facilitam o treinamento de redes profundas ao mitigar problemas como a degradação de gradiente. Essa característica torna as ResNets ideais para tarefas complexas, como a classificação de densidade mamária. Os modelos escolhidos foram ResNet-18, ResNet-34, ResNet-50 e ResNet-101, que variam em profundidade e capacidade de aprendizado, permitindo uma análise do impacto da complexidade da rede no desempenho do modelo.
- **Convolutional Neural Networks inspired by ViTs (ConvNeXts):** Projetadas como uma evolução moderna das CNNs, as ConvNeXts incorporam avanços arquiteturais inspirados em *Vision Transformer* (ViT), como o uso de camadas de atenção e técnicas de normalização mais eficientes. Essa abordagem permite um equilíbrio entre desempenho e custo computacional. Os modelos escolhidos foram ConvNeXt-Tiny, ConvNeXt-Small, ConvNeXt-Base e ConvNeXt-Large, que oferecem diferentes níveis de complexidade para avaliação.
- **Efficient Convolutional Neural Networks (EfficientNets):** Essas redes utilizam um método de dimensionamento balanceado que otimiza a relação entre profundidade, largura e resolução da rede, garantindo alta eficiência computacional sem comprometer a precisão. Essa característica é relevante para aplicações em larga escala, como a classificação de mamografias. Os modelos escolhidos foram EfficientNet-b0, EfficientNet-b1 e EfficientNet-b2, que variam em tamanho e capacidade de processamento.
- **Densely Connected Convolutional Networks (DenseNets):** Essas redes empregam conexões densas entre todas as camadas, promovendo um fluxo de gradiente mais eficiente durante o treinamento e permitindo a reutilização de características extraídas em diferentes níveis da rede. Essa arquitetura é útil para tarefas que exigem a identificação de padrões complexos, como a classificação de densidade mamária. Os modelos escolhidos foram DenseNet-121 e DenseNet-169, que oferecem um equilíbrio entre profundidade e eficiência.

Para adaptar esses modelos à tarefa de classificação de densidade mamária, a última camada totalmente conectada de cada arquitetura foi modificada para retornar um vetor com

o número de classes definido no experimento. Além disso, os pesos iniciais de cada modelo foram obtidos de versões pré-treinadas na base de dados ImageNet, o que permitiu aproveitar o conhecimento adquirido em um grande conjunto de dados e acelerar o processo de convergência durante o treinamento.

O pré-treinamento dessas redes foi realizado previamente por seus desenvolvedores, utilizando milhões de imagens da base ImageNet para aprender representações visuais genéricas. Esse processo permite que os modelos adquiram características úteis, como a detecção de bordas, texturas e padrões, que podem ser transferidas para outras tarefas de visão computacional. Dessa forma, ao reutilizar essas redes já treinadas, evita-se a necessidade de treinar um modelo do zero, reduzindo o tempo de computação e melhorando a convergência do aprendizado. Cabe destacar que essa etapa de pré-treinamento não foi realizada neste trabalho, sendo utilizada apenas a versão já treinada dos modelos.

Para o treinamento das redes, será aplicada a técnica de ajuste-fino *fine-tuning*. Essa abordagem consiste em utilizar redes neurais previamente treinadas em grandes bases de dados, como a ImageNet, e adaptar seus pesos para a classificação de densidade de mama. Isso é realizado de forma que as camadas convolucionais sejam refinadas a partir dos dados de treinamento. Dessa forma, as representações previamente aprendidas, como detecção de bordas, texturas e padrões, são preservadas e ajustadas para capturar características mais relevantes ao problema. Assim, será empregada uma taxa de aprendizado reduzida para evitar a destruição do conhecimento adquirido durante o pré-treinamento, garantindo uma adaptação eficiente ao novo domínio. Além disso, estratégias como parada antecipada e regularização serão utilizadas para mitigar o *overfitting* e melhorar a generalização dos modelos.

4.5.3 Processamento de Imagens

A primeira etapa de processamento envolveu a aplicação de uma rotação aleatória nas imagens, variando entre -20° e 20°. Como todas as imagens estavam na lateralidade esquerda, essa técnica foi fundamental para introduzir variabilidade nos dados e prevenir o *overfitting*. A rotação aleatória simula diferentes orientações das imagens, o que ajuda o modelo a generalizar melhor e a não se ajustar excessivamente aos padrões específicos do conjunto de treinamento.

Com base na revisão da literatura, as imagens foram redimensionadas para diferentes tamanhos com o objetivo de avaliar o impacto da resolução na performance do modelo. Os tamanhos escolhidos para redimensionamento foram:

- **224x224** pixéis;
- **336x224** pixéis;
- **516x516** pixéis;

Após a conversão para o formato *float32*, as imagens foram normalizadas utilizando as médias 0,485, 0,456 e 0,406 e os desvios padrão 0,229, 0,224 e 0,225 para os canais de cor

RGB, respectivamente. Esses valores de normalização foram extraídos do processamento das imagens da base ImageNet, utilizada no pré-treinamento das redes selecionadas.

Por fim, também foi utilizado o filtro magma em certos treinamentos, o qual é uma técnica de pré-processamento de imagens mencionada na literatura para melhorar o contraste e a visualização de estruturas em imagens médicas. O filtro magma aplica uma transformação de cores que realça detalhes sutis nas imagens, facilitando a identificação de padrões e características relevantes para a classificação da densidade mamária.

4.5.4 Funções de Perda

As funções de perda desempenham um papel crucial no treinamento, pois quantificam a diferença entre as previsões do modelo e os valores reais, guiando o processo de otimização. Para este trabalho, foram selecionadas três funções de perda:

- **Cross-Entropy Loss:** A função de perda de entropia cruzada é uma escolha clássica para problemas de classificação multiclasse. Ela mede a diferença entre a distribuição de probabilidade predita pelo modelo e a distribuição real dos rótulos. Essa função é amplamente utilizada devido à sua simplicidade e eficácia em uma variedade de tarefas de classificação.
- **Binary Cross Entropy Loss:** A função de perda binária de entropia cruzada é uma variação da entropia cruzada, específica para problemas de classificação binária.
- **Focal Loss:** Esta função é uma extensão da entropia cruzada binária, projetada para lidar com conjuntos de dados desbalanceados. Ela atribui um peso maior aos exemplos difíceis de classificar, reduzindo a influência de exemplos fáceis e, assim, permitindo que o modelo foque em padrões mais complexos.

A escolha dessas funções foi baseada na natureza do problema de classificação de densidade mamária e na necessidade de lidar com diferentes cenários, como desbalanceamento de classes e classificação multiclasse e binária.

4.5.5 Preparação dos Dados

Para avaliar a influência da quantidade e distribuição dos dados no desempenho dos modelos propostos, além de otimizar o tempo de treinamento, foram estabelecidos diferentes cenários experimentais. Assim, os dados foram estratificados em diferentes proporções, cada uma projetada para avaliar o comportamento dos modelos e agilizar o treinamento:

- **Treinamento com 20% da base de dados:** Esse cenário foi projetado para avaliar a capacidade dos modelos de aprender padrões relevantes mesmo com uma quantidade reduzida de amostras. Além disso, a utilização de um subconjunto menor permite

realizar múltiplos experimentos de forma mais eficiente, reduzindo o tempo computacional necessário para cada teste e possibilitando ajustes mais rápidos nos modelos e hiperparâmetros.

- **Treinamento com 90% e 95% da base de dados:** Estas configurações visam estabelecer o limite superior de desempenho dos modelos quando alimentados com uma quantidade substancial de dados de treinamento. A comparação entre os resultados obtidos com 90% e 95% também permite avaliar se há ganhos significativos com o incremento marginal de dados de treinamento.

Além disso, existe o cenário no qual usamos o balanceamento dos dados. Nele, o objetivo foi mitigar o impacto de possíveis vieses na distribuição original dos dados por meio de técnicas de reamostragem, garantindo que cada categoria de densidade mamária fosse representada por uma quantidade similar de imagens no conjunto de treinamento. Esta abordagem é especialmente importante para evitar que o modelo desenvolva tendências em favor das classes majoritárias.

Ademais, para cada cenário, a divisão dos dados seguiu um protocolo para garantir a representatividade dos conjuntos e a validade dos experimentos:

1. A estratificação foi realizada mantendo a proporcionalidade das classes em cada conjunto, exceto no cenário de dados balanceados.
2. A divisão foi realizada de forma aleatória, mas com uma semente fixa para garantir a reproduzibilidade dos experimentos.
3. Para cada configuração, foram mantidos conjuntos de teste independentes, não utilizados durante o treinamento.

Esta metodologia de divisão dos dados permite uma avaliação sobre os requisitos mínimos de dados para alcançar um desempenho relevante e aceito nas clínicas.

4.5.6 Otimizadores

A escolha do otimizador determina como os pesos do modelo são atualizados durante o processo de aprendizado. Neste trabalho, foram utilizados dois otimizadores distintos, cada um com características específicas que os tornam adequados para diferentes cenários:

- **Stochastic Gradient Descend:** Um otimizador clássico e amplamente utilizado devido à sua simplicidade e eficácia em uma variedade de tarefas de aprendizado profundo. O SGD atualiza os pesos do modelo com base no gradiente da função de perda em relação aos parâmetros, utilizando uma taxa de aprendizado fixa ou ajustável. Apesar de sua simplicidade, o SGD pode ser eficaz quando combinado com técnicas como *momentum* e agendadores de taxa de aprendizado.

- **Adaptive Moment Estimation Weighted:** Uma variação do otimizador Adam que incorpora decaimento de peso, uma técnica de regularização que ajuda a evitar *overfitting*. O *Adaptive Moment Estimation Weighted* (AdamW) combina as vantagens do Adam, como a adaptação dinâmica da taxa de aprendizado para cada parâmetro, com a capacidade de controlar a magnitude dos pesos, resultando em um treinamento mais estável e eficiente.

A seleção desses otimizadores foi baseada em sua eficácia comprovada na literatura e na necessidade de equilibrar desempenho e generalização do modelo.

4.5.7 Agendadores de Taxa de Aprendizado

A taxa de aprendizado é um hiperparâmetro crítico no treinamento de modelos de aprendizado profundo, influenciando diretamente a velocidade e a qualidade da convergência. Para ajustar dinamicamente a taxa de aprendizado durante o treinamento, foram utilizados dois agendadores:

- **ReduceLROnPlateau:** Este agendador reduz a taxa de aprendizado quando uma métrica de validação para de melhorar após um número definido de épocas. Essa abordagem é útil para evitar que o modelo fique preso em platôs durante o treinamento, permitindo que ele continue a convergir para uma solução melhor.
- **Cosine Annealing Scheduler:** Este agendador ajusta a taxa de aprendizado seguindo uma função cosseno, que diminui gradualmente a taxa ao longo do tempo e, em alguns casos, reinicia ciclicamente para escapar de mínimos locais.

Além disso, foram realizados treinamentos sem o uso de agendadores, com uma taxa de aprendizado fixa, para comparar o impacto do ajuste dinâmico da taxa de aprendizado. Essa comparação permitiu avaliar a eficácia dos agendadores em melhorar o desempenho do modelo e evitar problemas como estagnação ou divergência durante o treinamento.

4.5.8 Amostragem de dados

O desbalanceamento de classes é um desafio comum em tarefas de classificação, especialmente em conjuntos de dados médicos, onde algumas classes podem ser significativamente menos representadas que outras. Para lidar com esse problema, foi utilizada uma amostragem ponderada durante o carregamento dos dados. Essa técnica garante que o modelo seja exposto a um número equilibrado de exemplos de cada classe, promovendo um treinamento mais justo e melhorando o desempenho em classes minoritárias. Além da amostragem ponderada, também foram realizados treinamentos com amostragem aleatória, onde os dados são carregados sem ajustes de peso.

4.6 Avaliação das Métricas

A avaliação do desempenho dos modelos propostos foi realizada por meio de métricas consolidadas na literatura, escolhidas de acordo com a natureza do problema de classificação de densidade mamária. As principais métricas utilizadas foram:

- **Acurácia:** Mede a proporção de previsões corretas em relação ao total de exemplos. Embora seja uma métrica intuitiva, ela pode ser enganosa em conjuntos de dados desbalanceados, onde a maioria das previsões pode pertencer à classe majoritária.
- **F1 Score:** Combina precisão e *recall* em uma única métrica, sendo especialmente útil para avaliar o desempenho do modelo em classes desbalanceadas.

Além dessas métricas, foram utilizadas a precisão e a sensibilidade para avaliar o desempenho do modelo. Essas métricas permitiram uma análise mais detalhada do comportamento do modelo em diferentes cenários de classificação.

Os resultados obtidos foram comparados com os de trabalhos recentes na literatura, destacando as contribuições deste estudo em relação ao estado da arte.

5

Resultados e Discussão

Os resultados apresentados neste capítulo foram organizados de acordo com os treinamentos binários e multiclasses, e foram divididos em etapas de experimentos. Cada etapa representa etapas nas quais um ou mais hiperparâmetros dos modelos fossem ajustados, com o objetivo de melhorar as métricas de desempenho, como acurácia, F1, precisão e sensibilidade. Seguiu-se uma abordagem iterativa, na qual os resultados de cada etapa definiram as decisões para as etapas subsequentes.

5.1 Modelos Binários

A primeira abordagem explorada neste trabalho foi o treinamento de modelos binários, seguindo a estrutura apresentada na Figura 4.8. Essa abordagem consiste em treinar quatro modelos binários independentes, cada um responsável por classificar uma das quatro categorias de densidade mamária definidas pelo sistema BI-RADS®. A predição final é obtida de acordo com a camada de classificação ao final da estrutura.

5.1.1 Primeira Etapa de Experimentos

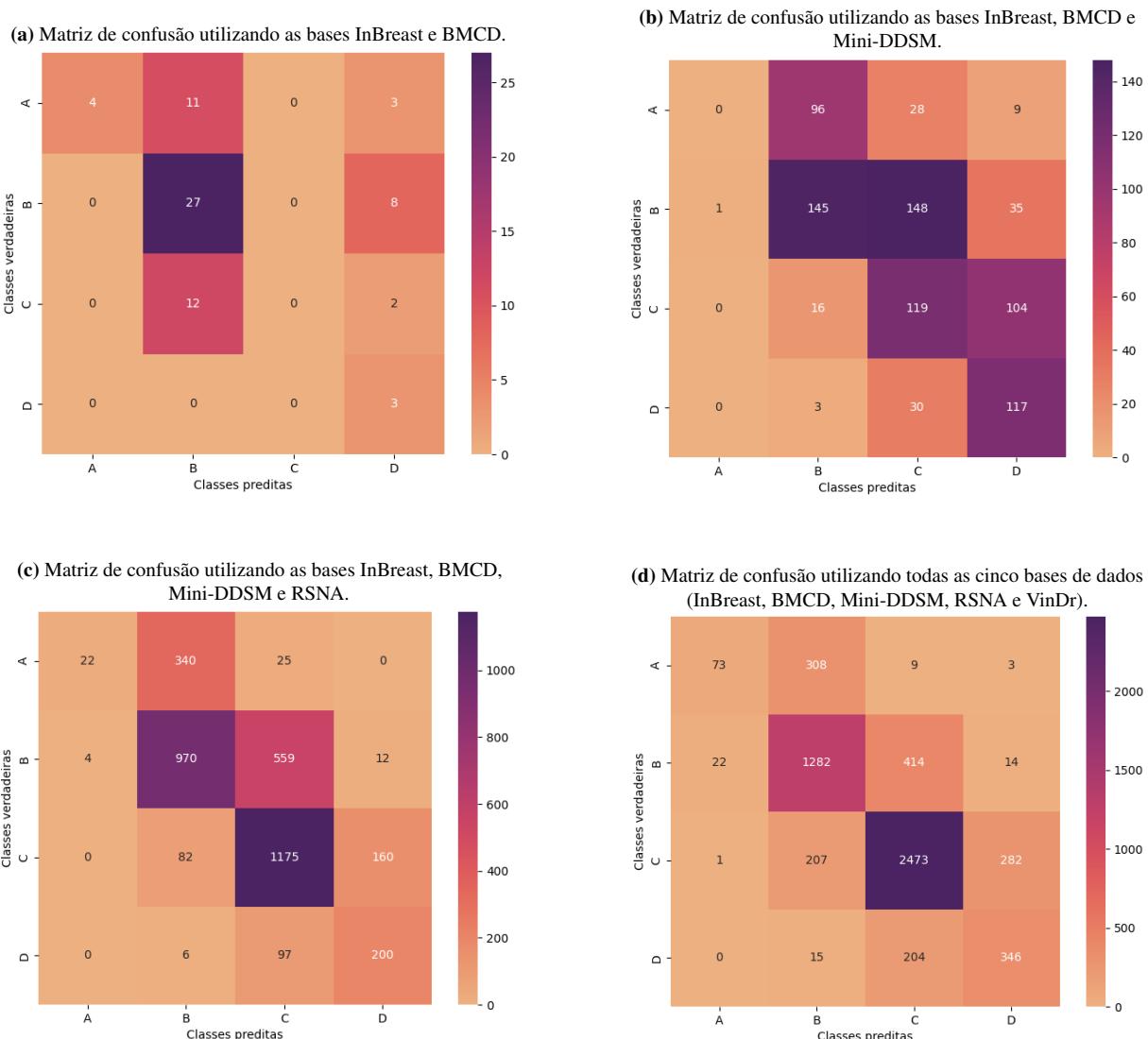
A primeira etapa de experimentos com modelos binários teve como objetivo principal implementar e validar a abordagem proposta, avaliando seu desempenho nos conjuntos de dados selecionados. Para isso, a lógica escolhida para a camada de classificação foi selecionar a classe com a maior densidade retornada pelos modelos.

Dessa forma, foi escolhido inicialmente o modelo ResNet50 para a classificação da densidade. A configuração inicial dos hiperparâmetros foi definida de acordo com a Tabela 5.1. Os treinamentos dos modelos foram realizados de forma incremental, utilizando diferentes combinações de bases de dados para avaliar a influência da quantidade e da diversidade dos dados no desempenho do modelo. Inicialmente, foram utilizadas as bases InBreast e BMCD para o primeiro treinamento. Em seguida, a base Mini-DDSM foi adicionada. Logo, na iteração seguinte, a base RSNA. Finalmente, foi integrada a base VinDr. As matrizes de confusão resultantes desses treinamentos são apresentadas na Figura 5.1.

Tabela 5.1: Configuração inicial dos hiperparâmetros utilizados na primeira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	BCE LOSS
Otimizador	AdamW
Agendador	<i>ReduceLROnPlateau</i>
<i>k-folds</i>	Média de 2 <i>folds</i>
Épocas	30
Resolução	336×224 pixels
Divisão de dados	90% para treinamento, 10% para teste

Figura 5.1: Matrizes de confusão dos treinamentos realizados na primeira etapa de experimentos, mostrando a influência das bases de dados no desempenho do modelo.



A análise das matrizes de confusão revelou que muitas previsões estavam sendo classificadas como uma classe acima da original, como pode ser observado na Figura 5.1b, Figura 5.1c e Figura 5.1d. Esse comportamento sugere que a lógica inicial de classificação, baseada na maior classe predita entre os quatro modelos binários, pode não ser adequada para lidar com a heterogeneidade das imagens e a complexidade da tarefa.

O último treinamento realizado, disponível na Figura 5.1d., com a adição da base VinDr, teve o aumento do número de *folds* na validação cruzada de 2 para 3. Essa mudança visava garantir que o modelo aprendesse de forma mais generalizada, dada a maior quantidade e diversidade de dados disponíveis.

Os resultados das métricas de desempenho para todos os treinamentos são apresentados na Tabela 5.2. Observa-se que o desempenho do modelo melhorou significativamente com a adição de mais bases de dados, atingindo uma ACC de 0,738 e um F1 de 0,727 no treinamento com todas as cinco bases. No entanto, a análise das matrizes de confusão indica que a lógica de classificação atual ainda pode ser aprimorada, uma vez que muitas previsões estão sendo classificadas incorretamente como uma classe acima da original.

Tabela 5.2: Resultados das métricas dos treinamentos da primeira etapa de experimentos.

Bases de dados	Acurácia	Precisão	Sensibilidade	F1
Duas bases	0,486	0,535	0,486	0,425
Três bases	0,448	0,396	0,448	0,408
Quatro bases	0,648	0,673	0,648	0,618
Cinco bases	0,738	0,741	0,738	0,727

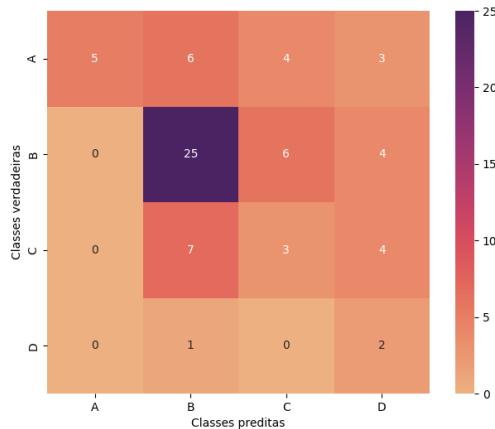
Diante desses resultados, a lógica da camada de classificação será alterada para selecionar a previsão com a maior diferença no limiar de confiança, visando melhorar a precisão das previsões e reduzir os erros de classificação. Essa abordagem será explorada na próxima etapa de experimentos.

5.1.2 Segunda Etapa de Experimentos

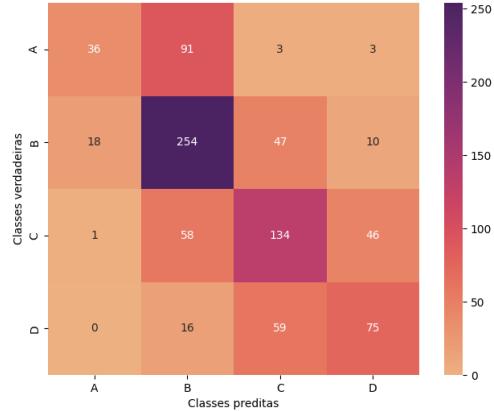
Na segunda etapa de experimentos, a lógica da camada final de classificação foi alterada para melhorar a discriminação entre as classes. Em vez de selecionar a classe com a maior densidade entre os modelos binários, optou-se por escolher a classe com a maior diferença entre a previsão do modelo e o limiar de classificação. Essa abordagem visa reduzir erros em casos onde as previsões estão próximas ao limiar, aumentando a confiança nas classificações. Como o treinamento dos modelos não é influenciado por essa camada, foram reutilizados os dados e os pesos obtidos na primeira etapa de experimentos. A Figura 5.2 apresenta as matrizes de confusão resultantes dessa nova lógica de classificação.

Figura 5.2: Matrizes de confusão dos resultados obtidos na segunda etapa de experimentos, utilizando a nova lógica de classificação.

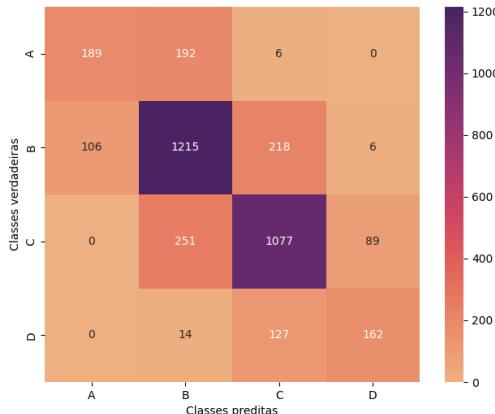
(a) Matriz de confusão utilizando as bases InBreast e BMCD.



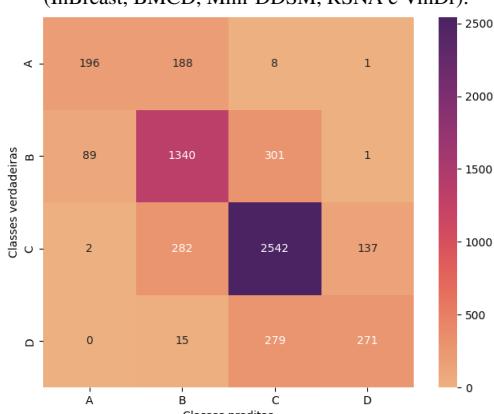
(b) Matriz de confusão utilizando as bases InBreast, BMCD e Mini-DDSM.



(c) Matriz de confusão utilizando as bases InBreast, BMCD, Mini-DDSM e RSNA.



(d) Matriz de confusão utilizando todas as cinco bases de dados (InBreast, BMCD, Mini-DDSM, RSNA e VinDr).



Os resultados das métricas obtidos com a nova lógica de classificação são apresentados na Tabela 5.3. Observa-se que a nova abordagem trouxe melhorias significativas, com destaque para o treinamento utilizando todas as cinco bases de dados, que alcançou uma ACC de 0,769 e um F1 de 0,764. No entanto, ainda persiste uma margem de erro de uma classificação de diferença entre as classes, como pode ser observado nas matrizes de confusão. Esse comportamento pode estar relacionado a fatores como o treinamento de cada *fold*, a distribuição desbalanceada dos dados, a arquitetura e aos hiperparâmetros utilizados, que podem ser refinados em experimentos futuros.

Tabela 5.3: Resultados das métricas dos treinamentos realizados na segunda etapa de experimentos.

Bases de dados	Acurácia	Precisão	Sensibilidade	F1
Duas bases	0,500	0,630	0,500	0,505
Três bases	0,586	0,590	0,586	0,572
Quatro bases	0,724	0,720	0,724	0,720
Cinco bases	0,769	0,764	0,769	0,764

Por fim, foi explorada uma última modificação nessa etapa: utilizar apenas o melhor modelo dos *k-folds*, em vez de realizar a média entre eles. Essa abordagem visa avaliar se a seleção do melhor modelo individual poderia superar o desempenho obtido com a média das previsões. A matriz de confusão resultante dessa exploração é apresentada na Figura 5.3, e as métricas obtidas são detalhadas na Tabela 5.4.

Figura 5.3: Matriz de confusão obtida ao utilizar apenas o melhor modelo dos *k-folds*, sem realizar a média das previsões.

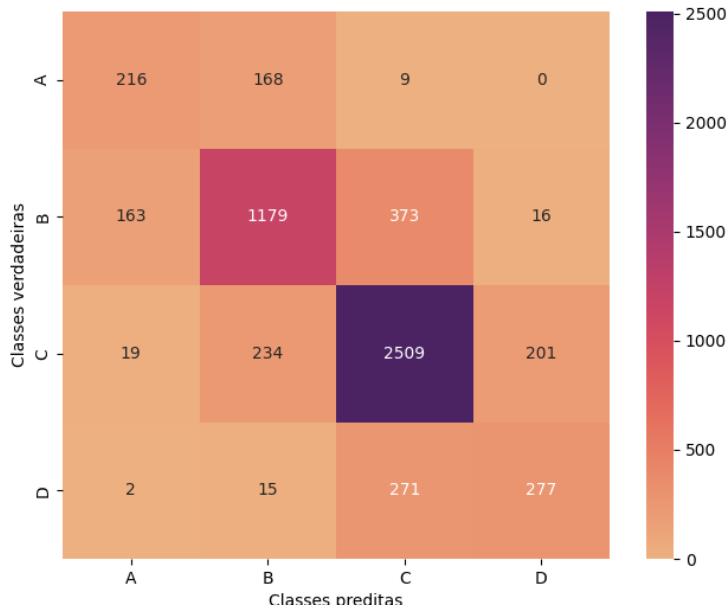


Tabela 5.4: Resultados das métricas ao utilizar apenas o melhor modelo dos *k-folds*.

Acurácia	Precisão	Sensibilidade	F1
0,740	0,736	0,740	0,737

Os resultados obtidos foram inferiores aos alcançados com a média das previsões dos *k-folds*, indicando que a estratégia de combinar múltiplos modelos proporciona maior robustez e generalização. Diante disso, optou-se por seguir com a abordagem de utilizar a média das previsões nos próximos experimentos, além de seguir com o uso de todas as bases de dados nos treinamentos.

5.1.3 Análise da Abordagem Binária

Os treinamentos realizados duraram desde algumas horas até uma semana e meia para serem finalizados, especialmente quando todos os dados disponíveis foram utilizados. Esse tempo é decorrente da arquitetura proposta, que envolve muitas instâncias de modelos binários independentes. Essa característica tornaria inviável testar diferentes combinações de hiperparâmetros e arquiteturas de modelos dentro do prazo disponível para o trabalho, uma vez que cada experimento demandaria um tempo significativo para ser concluído.

Diante disso, optou-se por continuar para a próxima abordagem, a de classificação multiclasse clássica, amplamente utilizada na literatura. Essa abordagem é computacionalmente mais eficiente, permitindo testar diferentes configurações de hiperparâmetros e modelos em um menor tempo.

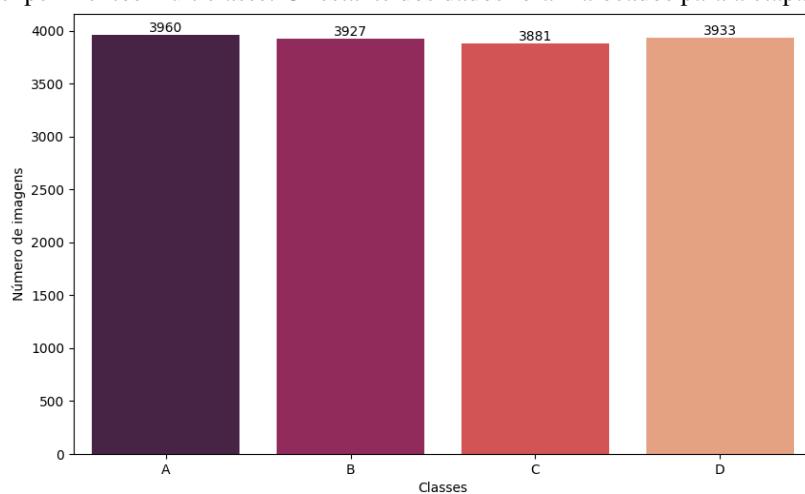
5.2 Modelos Multiclasse

A segunda abordagem explorada neste trabalho utiliza um único modelo para classificar todas as classes simultaneamente. A predição final é determinada pelo maior valor no vetor de quatro classes obtido ao final da rede neural. Essa abordagem tem como foco principal a análise das métricas de desempenho, como acurácia, precisão, sensibilidade e F1-score, em vez de se concentrar exclusivamente nas matrizes de confusão. Isso se deve ao fato de que o problema observado na abordagem anterior (classificação incorreta como uma classe acima) não se manifestou de forma tão evidente nessa arquitetura. Além disso, a maior eficiência computacional da abordagem multiclasse permitiu a realização de um número significativamente maior de treinamentos, possibilitando uma avaliação mais abrangente.

5.2.1 Primeira Etapa de Experimentos

A primeira etapa de experimentos multiclasse teve como objetivo identificar o modelo com as melhores métricas em um conjunto balanceado de dados. A distribuição das classes no conjunto balanceado é apresentada na Figura 5.4.

Figura 5.4: Distribuição das classes no conjunto de dados balanceado utilizado na primeira etapa de experimentos multiclasse. O restante dos dados foram alocados para a etapa de teste.



Os treinamentos foram realizados com os mesmos parâmetros para garantir uma comparação justa entre os modelos. A configuração inicial dos hiperparâmetros é detalhada na Tabela 5.5.

Tabela 5.5: Configuração inicial dos hiperparâmetros utilizados na primeira etapa de experimentos multiclasse.

Hiperparâmetros	Valores
Função de perda	FL LOSS
Otimizador	AdamW
Agendador	<i>ReduceLROnPlateau</i>
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	25
Resolução	336×224 pixels
Conjunto de dados	Balanceado
Divisão de dados	90% para treinamento, 10% para teste
Amostragem	Ponderada

Foram avaliadas as arquiteturas ResNet, ConvNeXt, EfficientNet e DenseNet, com suas respectivas variações de profundidade. As métricas obtidas nessa primeira etapa são apresentadas na Tabela 5.6.

Tabela 5.6: Resultados da primeira etapa de experimentos multiclasse.

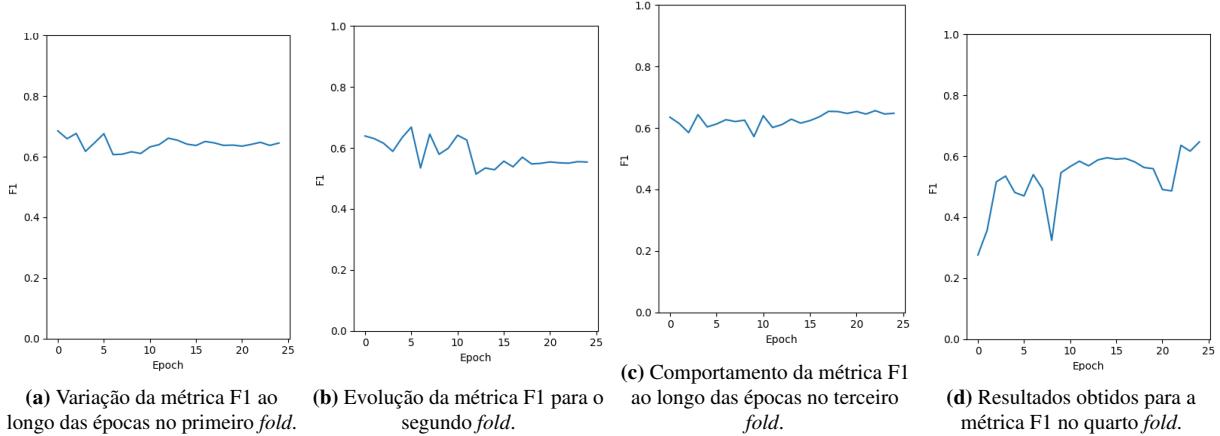
Modelo	Acurácia	Precisão	Sensibilidade	F1
ResNet34	0,669	0,684	0,669	0,672
ResNet50	0,662	0,681	0,662	0,665
ResNet101	0,653	0,660	0,653	0,655
ConvNeXt-tiny	0,275	0,076	0,275	0,119
ConvNeXt-small	0,275	0,076	0,275	0,119

Os experimentos foram interrompidos antes que todos os modelos fossem avaliados, pois o desempenho do ConvNeXt não se mostrou adequado aos parâmetros escolhidos. Observou-se que as métricas obtidas por esse modelo foram comprometidas devido à incompatibilidade do otimizador com a arquitetura.

Algumas observações relevantes sobre o processo de aprendizado indicam que, ao analisar a evolução da métrica F1 ao longo das épocas, o modelo poderia se beneficiar de um maior número de épocas de treinamento. Isso pode ser observado nos gráficos de cada *fold* de validação apresentados na Figura 5.5, onde a métrica ainda demonstra tendência de crescimento, sugerindo que o modelo não atingiu sua convergência total.

Figura 5.5: Evolução da métrica F1 ao longo das épocas para diferentes *folds* de validação.

Observa-se que a pontuação ainda apresenta tendência de crescimento, indicando que o treinamento pode ser estendido para um melhor ajuste do modelo.



Diante disso, optou-se por uma segunda etapa de experimentos, na qual os parâmetros de treinamento foram ajustados para garantir uma melhor generalização em todos os modelos.

5.2.2 Segunda Etapa de Experimentos

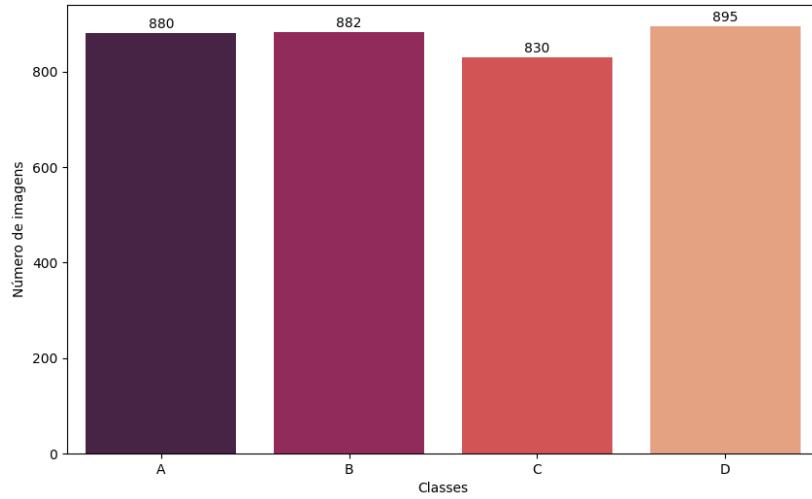
Nessa segunda etapa, os parâmetros de treinamento foram ajustados com base nos resultados da etapa anterior. A configuração dos hiperparâmetros é detalhada na Tabela 5.7.

Tabela 5.7: Configuração dos hiperparâmetros utilizados na segunda etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	CE LOSS
Otimizador	SGD
Agendador	Nenhum
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	50
Resolução	336×224 pixels
Conjunto de dados	Balanceado
Divisão de dados	20% para treinamento, 80% para teste
Amostragem	Ponderada
Parada Antecipada	25 épocas

Com o aumento do número de épocas para 50, conforme identificado na etapa anterior, o tempo total de treinamento foi duplicado. Para otimizar esse tempo, o conjunto de dados balanceado utilizado na primeira etapa foi reduzido para 20% do seu tamanho original. A distribuição das classes nesse subconjunto reduzido é apresentada na Figura 5.6. A amostragem ponderada também foi adicionada nesta etapa para evitar o desbalanceamento das classes. O restante dos dados foi reservado para a etapa de teste dos modelos treinados.

Figura 5.6: Distribuição das classes no conjunto de dados reduzido utilizado na segunda etapa de experimentos multiclasse. O restante dos dados foram alocados para a etapa de teste.



Além disso, para evitar que modelos que convergem mais rapidamente sejam prejudicados por um número maior de épocas, foi incorporado um critério de parada antecipada (*early stopping*) com um limite de 25 épocas sem melhoria na métrica de validação.

Os resultados obtidos nessa etapa são apresentados na Tabela 5.8. Observa-se que o modelo ConvNeXt-base obteve o melhor desempenho geral, com ACC de 0,690 e F1 de 0,706. Em comparação, os modelos da família ResNet e EfficientNet apresentaram desempenho inferior, com acuráncias variando entre 0,602 e 0,674. O DenseNet169 também se destacou, com uma acurácia de 0,671, mas ainda abaixo do ConvNeXt-base.

Tabela 5.8: Resultados da segunda etapa de experimentos com modelos multiclasse.

Modelo	Acurácia	Precisão	Sensibilidade	F1
ResNet18	0,647	0,706	0,647	0,662
ResNet34	0,626	0,564	0,665	0,588
ResNet50	0,674	0,721	0,674	0,686
ResNet101	0,602	0,677	0,602	0,614
EfficientNet-b0	0,642	0,692	0,642	0,654
EfficientNet-b1	0,659	0,705	0,659	0,667
EfficientNet-b2	0,629	0,706	0,629	0,647
DenseNet121	0,641	0,703	0,641	0,654
DenseNet169	0,671	0,709	0,671	0,682
ConvNeXt-tiny	0,681	0,748	0,681	0,696
ConvNeXt-small	0,654	0,731	0,654	0,672
ConvNeXt-base	0,690	0,752	0,690	0,706
ConvNeXt-large	0,685	0,751	0,685	0,700

Com base nos resultados, o modelo ConvNeXt-base foi selecionado como o principal

para os futuros experimentos, devido ao seu desempenho superior em todas as métricas avaliadas. Essa escolha permitiu avançar para a terceira etapa de experimentos, na qual o foco será o refinamento do modelo e a exploração de técnicas adicionais para melhorar as métricas.

5.2.3 Terceira Etapa de Experimentos

Na terceira etapa de experimentos, o objetivo foi explorar os hiperparâmetros do modelo ConvNeXt-base, buscando a configuração ideal para melhorar as métricas. Para isso, foram realizados sete treinamentos com diferentes configurações, conforme detalhado a seguir.

5.2.3.1 Primeiro Treinamento

No primeiro treinamento, utilizou-se os mesmos parâmetros da etapa anterior, porém a base de dados, ainda com o conjunto balanceado, foi aumentada para 90% do total para treino e validação, reservando 10% para teste. A distribuição da base de dados para treinamento pode ser encontrada na Figura 5.4. Além disso, foi adicionado o agendador *ReduceLRonPlateau* para ajustar dinamicamente a taxa de aprendizado. A configuração dos hiperparâmetros é detalhada na Tabela 5.9.

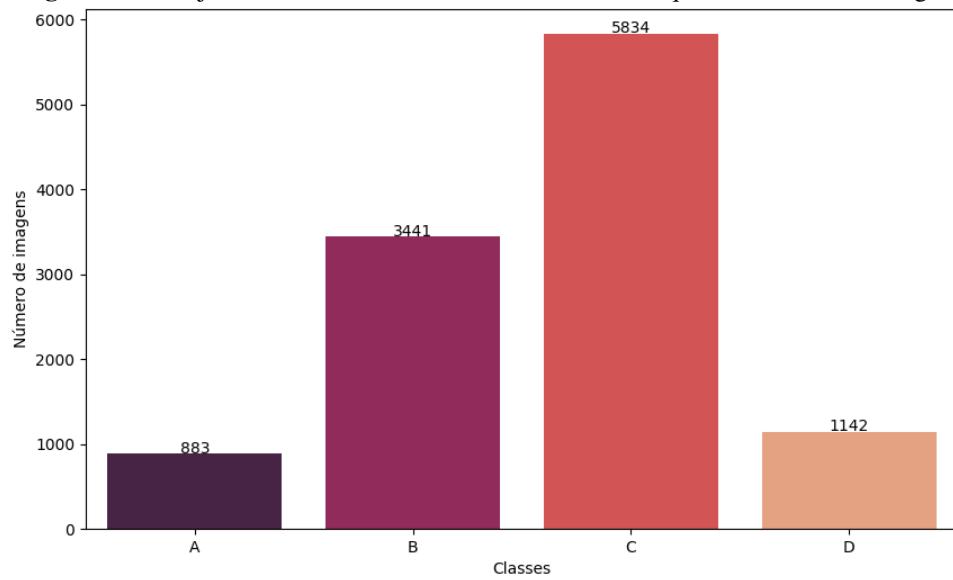
Tabela 5.9: Configuração dos hiperparâmetros utilizados no primeiro treinamento da terceira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	CE LOSS
Otimizador	SGD
Agendador	<i>ReduceLRonPlateau</i>
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	50
Resolução	336×224 pixels
Conjunto de dados	Balanceado
Divisão de dados	90% para treinamento, 10% para teste
Amostragem	Ponderada
Parada Antecipada	25 épocas

O modelo obteve uma ACC de 0,691 e um F1 de 0,733, indicando um desempenho inicial promissor. Esses resultados serviram como linha de base para as próximas iterações.

5.2.3.2 Segundo Treinamento

Para o segundo treinamento, optou-se por utilizar 20% dos dados da base de dados completa, visando avaliar a robustez do modelo com uma amostra maior e mais diversificada. A distribuição das classes no conjunto de dados utilizado é apresentada na Figura 5.7.

Figura 5.7: Conjunto de dados de treinamento com 20% da quantidade total de imagens.

A função de perda foi alterada para FL LOSS, devido ao desbalanceamento dos dados, e o agendador foi substituído por um agendador cosseno, com parada antecipada de 15 épocas. A configuração dos hiperparâmetros é detalhada na Tabela 5.10.

Tabela 5.10: Configuração dos hiperparâmetros utilizados no segundo treinamento da terceira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	FL LOSS
Otimizador	SGD
Agendador	Cosseno
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	50
Resolução	336×224 pixels
Conjunto de dados	Completo
Divisão de dados	20% para treinamento, 80% para teste
Amostragem	Ponderada
Parada Antecipada	15 épocas

O modelo obteve ACC de 0,733 e F1 de 0,725, mostrando uma melhora na acurácia, mas um leve decréscimo no F1 em comparação ao primeiro treinamento.

5.2.3.3 Terceiro Treinamento

No terceiro treinamento, testou-se a alteração da resolução das imagens, reduzindo-as de 336x224 para 224x224 pixels, a fim de verificar o impacto dessa mudança nas métricas.

Essa alteração resultou em uma queda no desempenho, com uma ACC de 0,712 e um F1 de 0,698, sugerindo que a resolução original era mais adequada para o modelo. A configuração dos hiperparâmetros é detalhada na Tabela 5.11.

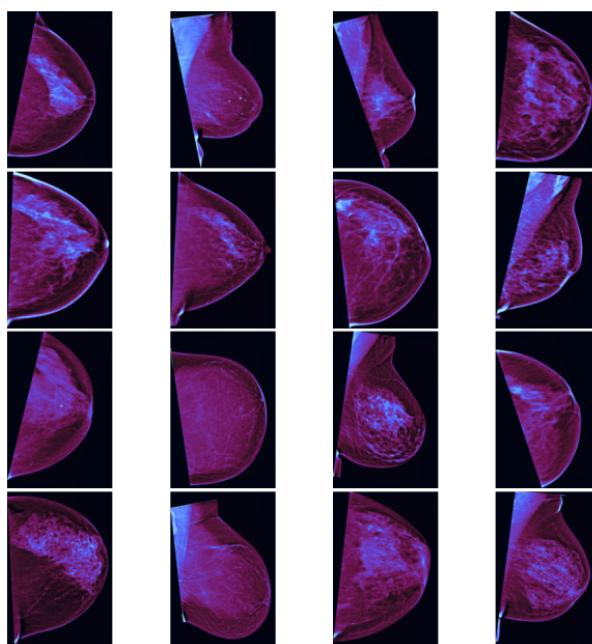
Tabela 5.11: Configuração dos hiperparâmetros utilizados no terceiro treinamento da terceira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	FL LOSS
Otimizador	SGD
Agendador	Cosseno
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	50
Resolução	224 × 224 <i>pixels</i>
Conjunto de dados	Completo
Divisão de dados	20% para treinamento, 80% para teste
Amostragem	Ponderada
Parada Antecipada	15 épocas

5.2.3.4 Quarto Treinamento

No quarto treinamento, a resolução foi mantida em 336x224 *pixels*, e foi aplicado o filtro magma, utilizado em alguns trabalhos por melhorar o contraste das imagens mamográficas. Um exemplo de lote de treinamento com o filtro aplicado é apresentado na Figura 5.8.

Figura 5.8: Exemplo de um lote de treinamento com o filtro magma aplicado nas imagens.



Essa abordagem resultou em ACC de 0,753 e F1 de 0,746, indicando que o filtro teve um impacto positivo no desempenho do modelo. A configuração dos hiperparâmetros é detalhada na Tabela 5.12.

Tabela 5.12: Configuração dos hiperparâmetros utilizados no quarto treinamento da terceira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	FL LOSS
Otimizador	SGD
Agendador	Cosseno
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	50
Resolução	336×224 pixels
Conjunto de dados	Completo
Divisão de dados	20% para treinamento, 80% para teste
Amostragem	Ponderada
Parada Antecipada	15 épocas
Filtro	Magma

5.2.3.5 Quinto Treinamento

No quinto treinamento, a resolução das imagens foi aumentada para 516x516 pixels, mantendo o filtro magma. No entanto, as métricas diminuíram levemente, com ACC de 0,746 e F1 de 0,730, sugerindo que o aumento da resolução não trouxe benefícios adicionais ao modelo. A configuração dos hiperparâmetros é detalhada na Tabela 5.13.

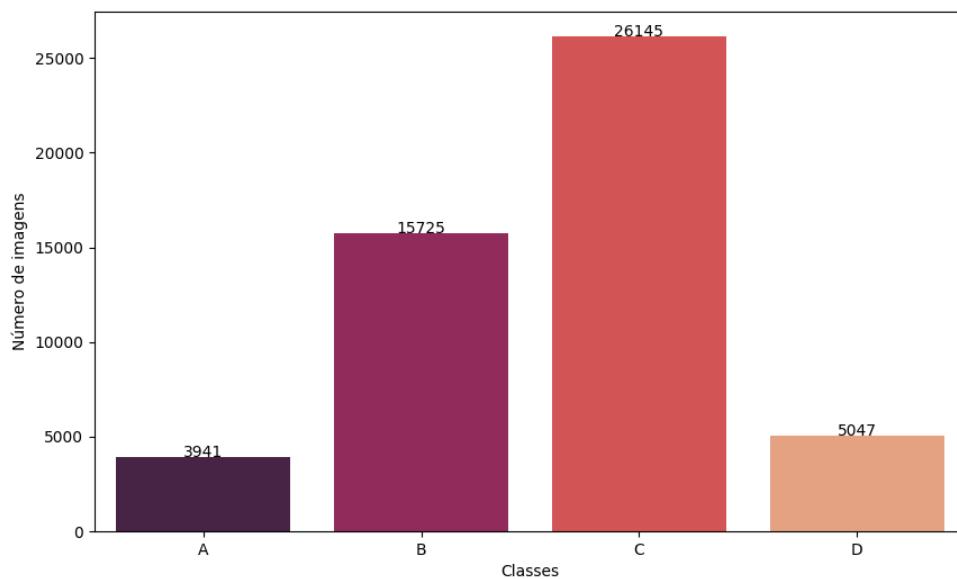
Tabela 5.13: Configuração dos hiperparâmetros utilizados no quinto treinamento da terceira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	FL LOSS
Otimizador	SGD
Agendador	Cosseno
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	50
Resolução	516×516 pixels
Conjunto de dados	Completo
Divisão de dados	20% para treinamento, 80% para teste
Amostragem	Ponderada
Parada Antecipada	15 épocas
Filtro	Magma

5.2.3.6 Sexto Treinamento

No sexto treinamento, a resolução foi mantida em 336×224 pixels, e utilizou-se 90% dos dados para treinamento, a fim de verificar o desempenho do modelo com uma quantidade maior de dados. A distribuição das classes no conjunto de dados utilizado é apresentada na Figura 5.9.

Figura 5.9: Conjunto de dados de treinamento com 90% da quantidade total de imagens.



Os resultados levaram a uma ACC de 0,755 e um F1 de 0,742, representando uma melhoria em relação ao treinamento anterior, mas ainda abaixo do esperado considerando o aumento no volume de dados. A configuração dos hiperparâmetros é detalhada na Tabela 5.14.

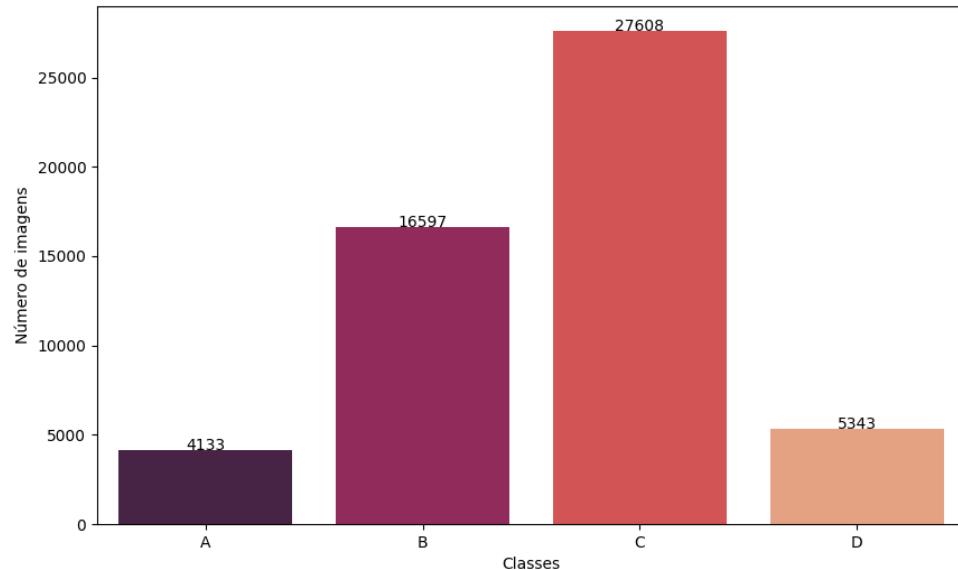
Tabela 5.14: Configuração dos hiperparâmetros utilizados no sexto treinamento da terceira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	FL LOSS
Otimizador	SGD
Agendador	Cosseno
<i>k-folds</i>	Média de 4 <i>folds</i>
Épocas	50
Resolução	336×224 pixels
Conjunto de dados	Completo
Divisão de dados	90% para treinamento, 10% para teste
Amostragem	Ponderada
Parada Antecipada	15 épocas
Filtro	Nenhum

5.2.3.7 Sétimo Treinamento

No sétimo e último treinamento da etapa, utilizou-se 95% do total de dados para treinamento, com 5% reservado para teste, conforme ilustrado na Figura 5.10. Essa configuração permitiu validar o modelo com um conjunto de dados mais representativo.

Figura 5.10: Conjunto de dados de treinamento com 95% da quantidade total de imagens.



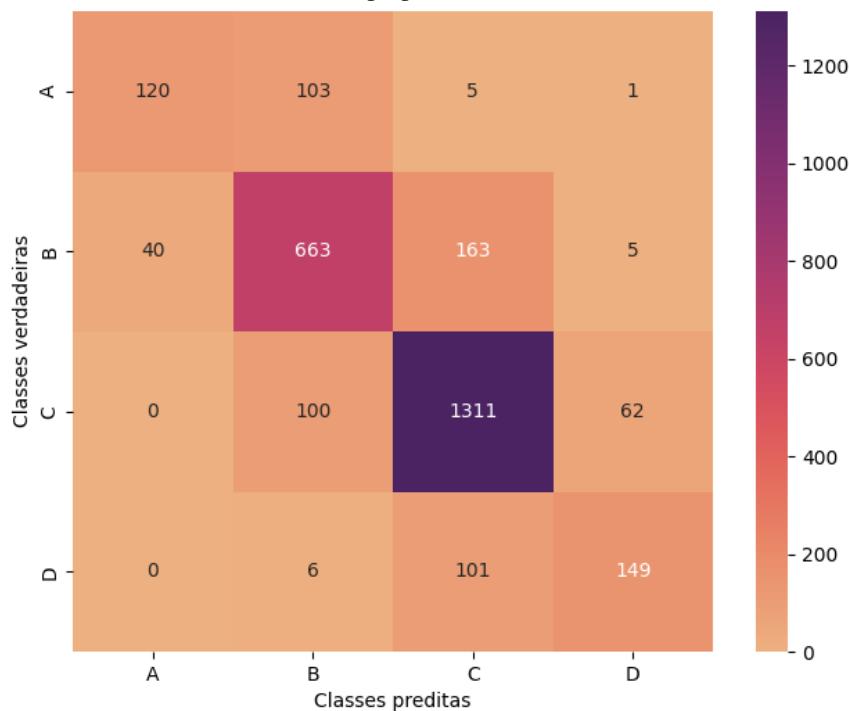
O modelo alcançou ACC de 0,793 e F1 de 0,788, demonstrando uma melhora significativa no desempenho e confirmando a capacidade de generalização da arquitetura. A configuração dos hiperparâmetros é detalhada na Tabela 5.15.

Tabela 5.15: Configuração dos hiperparâmetros utilizados no sétimo treinamento da terceira etapa de experimentos.

Hiperparâmetros	Valores
Função de perda	FL LOSS
Otimizador	SGD
Agendador	Cosseno
k -folds	Média de 4 folds
Épocas	50
Resolução	336×224 pixels
Conjunto de dados	Completo
Divisão de dados	95% para treinamento, 5% para teste
Amostragem	Ponderada
Parada Antecipada	15 épocas
Filtro	Nenhum

A matriz de confusão desse treinamento se encontra na Figura 5.11.

Figura 5.11: Matriz de confusão do melhor modelo multiclasse obtido após várias iterações de hiperparâmetros.



5.2.3.8 Resultados da Etapa

Os resultados de todos os treinamentos são apresentados na Tabela 5.16. Observa-se que o último treinamento obteve o melhor desempenho geral, enquanto o uso do filtro magma no quarto treinamento também se destacou como uma abordagem promissora.

Tabela 5.16: Resultados consolidados da terceira etapa de experimentos.

Treinamento	Acurácia	Precisão	Sensibilidade	F1
1º	0,691	0,818	0,691	0,733
2º	0,733	0,737	0,733	0,725
3º	0,712	0,719	0,712	0,698
4º	0,753	0,754	0,753	0,746
5º	0,746	0,747	0,746	0,730
6º	0,755	0,755	0,755	0,742
7º	0,793	0,789	0,793	0,788

Esses resultados demonstram a importância da escolha adequada de hiperparâmetros, tamanho do conjunto de dados e técnicas de pré-processamento para o desempenho do modelo.

5.3 Resultados Finais

A Tabela 5.17 apresenta uma comparação dos resultados obtidos com trabalhos recentes na literatura. As métricas macroAUC e Linear κ foram calculadas nos melhores modelos, utilizando as probabilidades e previsões obtidas em seus respectivos treinamentos e testes. Entretanto, como a abordagem binária tem uma lógica diferente, a métrica macroAUC não foi calculada por não representar seu valor real.

Tabela 5.17: Comparação de diferentes modelos na classificação da densidade mamária nas quatro classes do BI-RADS®.

Artigo	Acurácia	F1	macroAUC	Linear Kappa
Abordagem multiclasse deste trabalho	79,3	78,8	0,940	0,71
Abordagem binária deste trabalho	76,9	76,4	-	0,68
BIROŠ et al. (2024)	81,9	-	-	-
BUSALEH et al. (2022)	96,0	-	-	-
LIZZI et al. (2022)	68,0	-	-	-
PAWAR et al. (2022)	-	88,6	-	-
NGUYEN et al. (2021)	-	61,6	-	-
MATTHEWS et al. (2021)	82,2	-	0,952	0,75
CHANG et al. (2020)	-	-	-	0,66
COGAN; TAMIL (2020)	-	-	0,897	-
DENG et al. (2020)	92,1	90,3	-	-
LEHMAN et al. (2019)	77,0	-	-	0,67
CHAN; HELVIE (2019)	-	-	-	0,67

Nossa abordagem multiclasse alcançou uma ACC de 79,3%, enquanto a abordagem

binária obteve 76,9%. Esses resultados demonstram que ambas as abordagens estão alinhadas com o estado da arte, embora ainda haja espaço para melhorias, como a verificação de novos hiperparâmetros e configurações de treinamento.

Todas as etapas descritas no Capítulo 4, bem como os resultados obtidos neste capítulo, estão disponíveis publicamente no repositório eletrônico <<https://github.com/loioladev/breast-density-classification>>. O repositório contém os códigos-fonte, *scripts* de treinamento, processamento dos dados e instruções detalhadas para reproduzir os experimentos e validar os resultados.

6

Conclusões e Trabalhos Futuros

6.1 Conclusões

Este trabalho explorou duas abordagens para a classificação da densidade mamária em mamografias digitais: uma baseada em modelos binários independentes e outra utilizando modelos multiclasse. Ambas as abordagens demonstraram resultados competitivos em relação ao estado da arte, com acuráncias de 76,9% e 79,3%, respectivamente. A abordagem multiclasse, além de obter a melhor performance geral, mostrou-se mais eficiente computacionalmente e mais fácil de ajustar, enquanto a abordagem binária ofereceu informações valiosas sobre a complexidade da tarefa e a necessidade de técnicas específicas para lidar com o desbalanceamento de classes.

A exploração de diferentes arquiteturas de redes neurais, como ConvNeXt, ResNet e EfficientNet, permitiu identificar configurações que melhor se adaptam à tarefa de classificação da densidade mamária. Além disso, a experimentação com diferentes estratégias de treinamento, incluindo ajustes finos e variações nos hiperparâmetros, evidenciou o impacto de técnicas como validação cruzada, balanceamento de dados e critérios de parada antecipada na melhoria da robustez e generalização dos modelos.

Apesar dos resultados promissores, ainda há espaço para avanços. A investigação de novas estratégias de pré-processamento de imagens, como realce de contraste e técnicas baseadas em aprendizado auto-supervisionado, pode contribuir para um melhor aproveitamento das informações visuais. Além disso, a incorporação de modelos multimodais, combinando imagens com metadados clínicos, pode fornecer uma abordagem mais abrangente para a classificação da densidade mamária.

Por fim, a disponibilização pública de todo o código e dos dados utilizados neste trabalho visa promover a reproduzibilidade e o avanço da pesquisa na área. Essa iniciativa contribui para o desenvolvimento de ferramentas computacionais que auxiliem radiologistas no diagnóstico precoce do câncer de mama, reforçando o potencial do aprendizado profundo na melhoria da detecção e avaliação da densidade mamária em mamografias digitais.

6.2 Trabalhos Futuros

Com base nos resultados obtidos nos experimentos realizados, identificam-se diversas oportunidades para aprimorar o modelo proposto e expandir sua aplicabilidade. As principais direções futuras incluem:

- **Integração das abordagens binária e multiclasse:** Substituir o modelo binário atual pelos hiperparâmetros e técnicas que apresentaram os melhores resultados na abordagem multiclasse, visando combinar as vantagens de ambas as abordagens.
- **Exploração de técnicas avançadas de classificação:** Substituir a camada de classificação da arquitetura binária por um classificador treinável, como uma rede neural adicional, para capturar relações mais complexas entre as saídas dos modelos binários.
- **Expansão do conjunto de dados:** Incluir novas bases de dados públicas ou privadas para aumentar a diversidade e a representatividade dos dados, melhorando a generalização do modelo.
- **Validação em bases de dados externas:** Testar o modelo em bases de dados externas e não vistas durante o treinamento, a fim de validar sua robustez e aplicabilidade em cenários clínicos reais.
- **Análise de interpretabilidade do modelo:** Explorar técnicas de interpretabilidade, como mapas de ativação (*Grad-CAM*), para entender como o modelo toma decisões e identificar possíveis vieses ou falhas.

Essas direções futuras têm o potencial de aprimorar significativamente o desempenho e a aplicabilidade do modelo, contribuindo para o avanço da área de diagnóstico assistido por computador e, consequentemente, para a melhoria do cuidado com a saúde das pacientes.

REFERÊNCIAS

- ALOM, M. Z. et al. **The History Began from AlexNet**: a comprehensive survey on deep learning approaches. 2018.
- ALPAYDIN, E. **Introduction to Machine Learning, fourth edition**. [S.l.]: MIT Press, 2020. (Adaptive Computation and Machine Learning series).
- ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics Surveys**, [S.l.], v.4, n.none, p.40 – 79, 2010.
- BERG, W. A. et al. Breast Imaging Reporting and Data System. **American Journal of Roentgenology**, [S.l.], v.174, n.6, p.1769–1777, 2000. PMID: 10845521.
- BIDGOOD, W. D. et al. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. **Journal of the American Medical Informatics Association**, [S.l.], v.4, n.3, p.199–212, 05 1997.
- BIROŠ, M. et al. Enhancing Accuracy in Breast Density Assessment Using Deep Learning: a multicentric, multi-reader study. **Diagnostics**, [S.l.], v.14, n.11, 2024.
- BODEWES, F. et al. Mammographic breast density and the risk of breast cancer: a systematic review and meta-analysis. **The Breast**, [S.l.], v.66, p.62–68, 2022.
- BOTCHKAREV, A. A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. **Interdisciplinary Journal of Information, Knowledge, and Management**, [S.l.], v.14, p.045–076, 2019.
- BUSALEH, M. et al. TwoViewDensityNet: two-view mammographic breast density classification based on deep convolutional neural network. **Mathematics**, [S.l.], v.10, n.23, p.4610, 2022.
- CARR, C. et al. **RSNA Screening Mammography Breast Cancer Detection**. Kaggle, <https://kaggle.com/competitions/rsna-breast-cancer-detection>.
- CHAN, H.-P.; HELVIE, M. A. Deep Learning for Mammographic Breast Density Assessment and Beyond. **Radiology**, [S.l.], v.290, n.1, p.59–60, 2019. PMID: 30325286.
- CHANG, K. et al. Multi-Institutional Assessment and Crowdsourcing Evaluation of Deep Learning for Automated Classification of Breast Density. **Journal of the American College of Radiology**, [S.l.], v.17, n.12, p.1653–1662, 2020.
- COGAN, T.; TAMIL, L. Deep Understanding of Breast Density Classification. In: ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY (EMBC), 2020. **Anais...** [S.l.: s.n.], 2020. p.1140–1143.
- COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, [S.l.], v.20, n.1, p.37–46, 1960.
- CRESWELL, A. et al. Generative Adversarial Networks: an overview. **IEEE Signal Processing Magazine**, [S.l.], v.35, n.1, p.53–65, 2018.
- LIU, L.; ÖZSU, M. T. (Ed.). **Cross-Validation**. Boston, MA: Springer US, 2009. p.532–538.

- DENG, J. et al. Classification of breast density categories based on SE-Attention neural networks. **Computer Methods and Programs in Biomedicine**, [S.I.], v.193, p.105489, 2020.
- DERMEVAL, D.; COELHO, J. A. P. d.; BITTENCOURT, I. I. Mapeamento Sistemático e Revisão Sistemática da Literatura em Informática na Educação. **Revista Brasileira de Informática na Educação (RBIE)**, [S.I.], 2020. Acesso em: 23-07-2024.
- DOAN, A.; HALEVY, A.; IVES, Z. 1 - Introduction. In: DOAN, A.; HALEVY, A.; IVES, Z. (Ed.). **Principles of Data Integration**. Boston: Morgan Kaufmann, 2012. p.1–18.
- DONTCHOS, B. N. et al. External Validation of a Deep Learning Model for Predicting Mammographic Breast Density in Routine Clinical Practice. **Academic Radiology**, [S.I.], v.28, n.4, p.475–480, 2021.
- ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science**: algoritmos de machine learning e métodos de análise. [S.I.: s.n.], 2020. 272p.
- FDA. **Mammography Quality Standards Act and Program**. Acesso em: 17 jun. 2024., Disponível em: <<https://www.fda.gov/radiation-emitting-products/mammography-quality-standards-act-and-program>>.
- GEMICI, A. A. et al. Comparison of breast density assessments according to BI-RADS 4th and 5th editions and experience level. **Acta Radiologica Open**, [S.I.], v.9, n.7, p.2058460120937381, 2020. PMID: 32733694.
- GOLDBERGER, A. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. **Circulation**, [S.I.], v.101, n.23, p.e215–e220, 2000. [Online].
- GOODFELLOW, I. et al. Generative adversarial networks. **Commun. ACM**, New York, NY, USA, v.63, n.11, p.139–144, oct 2020.
- GRANDINI, M.; BAGLI, E.; VISANI, G. **Metrics for Multi-Class Classification**: an overview. 2020.
- GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**: concepts, tools, and techniques to build intelligent systems. 3rd.ed. Sebastopol, CA: O'Reilly Media, 2022.
- HALEEM, A.; JAVAID, M.; KHAN, I. H. Current status and applications of Artificial Intelligence (AI) in medical field: an overview. **Current Medicine Research and Practice**, [S.I.], v.9, n.6, p.231–237, 2019.
- HALLING-BROWN, M. D. et al. OPTIMAM Mammography Image Database: a large-scale resource of mammography images and clinical data. **Radiology: Artificial Intelligence**, [S.I.], v.3, n.1, p.e200103, 2021. PMID: 33937853.
- HAYKIN, S. **Redes Neurais**: princípios e práticas. 2^a ed..ed. São Paulo: BOOKMAN, 2001. 900p.
- HE, K. et al. Deep Residual Learning for Image Recognition. **CoRR**, [S.I.], v.abs/1512.03385, 2015.

- HOLZINGER, A. Explainable AI and Multi-Modal Causability in Medicine. **i-com**, [S.l.], v.19, n.3, p.171–179, 2020.
- HUANG, G. et al. **Densely Connected Convolutional Networks**. 2018.
- HUANG, M.-L.; LIN, T.-Y. Dataset of breast mammography images with masses. **Data in Brief**, [S.l.], v.31, p.105928, 2020.
- IARC. **Global Cancer Observatory Cancer Today**. Acesso em: 17 jun. 2024., Disponível em: <<https://gco.iarc.fr/today>>.
- INCA. **Detecção precoce**. Publicado em: 16 set. 2022. Atualizado em: 28 maio 2024. Acesso em: 24 jun. 2024., Disponível em: <<https://www.gov.br/inca/pt-br/assuntos/gestor-e-profissional-de-saude/controle-do-cancer-de-mama/acoes/deteccao-precoce>>.
- KHAN, S.; RAHMANI, H.; SHAH, S. A. A. **A Guide to Convolutional Neural Networks for Computer Vision**. [S.l.]: Morgan & Claypool Publishers, 2018.
- KITCHENHAM, B. et al. Systematic literature reviews in software engineering – A tertiary study. **Information and Software Technology**, [S.l.], v.52, n.8, p.792–805, 2010.
- KOLO, B. **Binary and Multiclass Classification**. [S.l.]: Weatherford Press, 2011.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. **Anais...** Curran Associates: Inc., 2012. v.25.
- LAROBINA, M. Thirty Years of the DICOM Standard. **Tomography**, [S.l.], v.9, n.5, p.1829–1838, 2023.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, [S.l.], v.521, n.7553, p.436–444, 2015.
- LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, [S.l.], v.86, n.11, p.2278–2324, 1998.
- LEHMAN, C. D. et al. Mammographic Breast Density Assessment Using Deep Learning: clinical implementation. **Radiology**, [S.l.], v.290, n.1, p.52–58, 2019. PMID: 30325282.
- LEKAMLAGE, C. D. et al. Mini-DDSM: mammography-based automatic age estimation. In: INTERNATIONAL CONFERENCE ON DIGITAL MEDICINE AND IMAGE PROCESSING (DMIP 2020), 3. **Proceedings...** ACM, 2020.
- LEWIN, J. M. 17 - Digital Mammography. In: HAYAT, M. (Ed.). **Cancer Imaging**. San Diego: Academic Press, 2008. p.455–458.
- LI, C. et al. Multi-View Mammographic Density Classification by Dilated and Attention-Guided Residual Learning. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, [S.l.], v.18, p.1003–1013, 2020.
- LI, Q. et al. Medical image classification with convolutional neural network. In: INTERNATIONAL CONFERENCE ON CONTROL AUTOMATION ROBOTICS AND VISION (ICARCV), 2014. **Anais...** [S.l.: s.n.], 2014. p.844–848.

- LI, Z. et al. **A Survey of Convolutional Neural Networks:** analysis, applications, and prospects. 2020.
- LIU, P. ran et al. Application of Artificial Intelligence in Medicine: an overview. **Current Medical Science**, [S.I.], v.41, n.6, p.1105–1115, dec 2021.
- LIZZI, F. et al. Convolutional Neural Networks for Breast Density Classification: performance and explanation insights. **Applied Sciences**, [S.I.], v.12, n.1, 2022.
- LOBO, L. C. Inteligência artificial, o Futuro da Medicina e a Educação Médica. **Revista Brasileira de Educação Médica**, [S.I.], v.42, n.3, p.3–8, Jul 2018.
- LOIZIDOU, K. et al. **Breast Micro-Calcifications Dataset with Precisely Annotated Sequential Mammograms.** [S.I.]: Zenodo, 2021.
- LOPEZ-ALMAZAN, H. et al. A deep learning framework to classify breast density with noisy labels regularization. **Computer Methods and Programs in Biomedicine**, [S.I.], v.221, p.106885, 2022.
- LUDERMIR, T. B. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, [S.I.], v.35, n.101, p.85–94, Jan 2021.
- MAGHSOUDI, O. H. et al. Deep-LIBRA: an artificial-intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment. **Medical Image Analysis**, [S.I.], v.73, p.102138, 2021.
- MAHARANA, K.; MONDAL, S.; NEMADE, B. A review: data pre-processing and data augmentation techniques. **Global Transitions Proceedings**, [S.I.], v.3, n.1, p.91–99, 2022. International Conference on Intelligent Engineering Approach(ICIEA-2022).
- MATTHEWS, T. P. et al. A Multisite Study of a Breast Density Deep Learning Model for Full-Field Digital Mammography and Synthetic Mammography. **Radiology: Artificial Intelligence**, [S.I.], v.3, n.1, p.e200015, 2021. PMID: 33937850.
- MINA, L.; ISA, N. A. M. **A fully automated breast separation For mammographic images.** [S.I.: s.n.], 2015. 37-41p.
- MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos sobre aprendizado de máquina.** 1. ed..ed. [S.I.]: Rezende, 2003. v.1. Disponível em: <<https://dcm.ffclrp.usp.br/augusto/publications/2003-sistemas-inteligentes-cap4.pdf>>. Acesso em: 24 jun. 2024.
- MOREIRA, I. C. et al. INbreast: toward a full-field digital mammographic database. **Academic Radiology**, [S.I.], v.19, n.2, p.236–248, Feb 2012. Epub 2011 Nov 10. PMID: 22078258.
- MUAMEDYEV, R. Machine learning methods: an overview. **Computer modelling & new technologies**, [S.I.], v.19, n.6, p.14–29, 2015.
- NAHM, F. Receiver operating characteristic curve: overview and practical use for clinicians. **Korean Journal of Anesthesiology**, [S.I.], v.75, 01 2022.
- NGUYEN, H. T. X. et al. A novel multi-view deep learning approach for BI-RADS and density assessment of mammograms. **2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)**, [S.I.], p.2144–2148, 2021.

- OLANIPEKUN, A. et al. Applying a Neural Network-Based Machine Learning to Laser-Welded Spark Plasma Sintered Steel: predicting vickers micro-hardness. **Journal of Manufacturing and Materials Processing**, [S.I.], v.6, p.91, 08 2022.
- OOMS, E. et al. Mammography: interobserver variability in breast density assessment. **The Breast**, [S.I.], v.16, n.6, p.568–576, 12 2007.
- Papers With Code. **Max Pooling**. Accessed: 2024-08-06, <https://paperswithcode.com/method/max-pooling>.
- PAWAR, S. D. et al. Multichannel DenseNet Architecture for Classification of Mammographic Breast Density for Breast Cancer Detection. **Frontiers in Public Health**, [S.I.], v.10, 2022.
- PEGORINI, J. et al. Desafios e Soluções em Sistemas de Votação Eletrônica: um mapeamento sistemático. In: IV WORKSHOP DE TECNOLOGIA ELEITORAL, Porto Alegre, RS, Brasil. **Anais...** SBC, 2019. p.13–24.
- PHAM, H. H.; NGUYEN TRUNG, H.; NGUYEN, H. Q. **VinDr-Mammo**: a large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography (version 1.0.0). [S.I.]: PhysioNet, 2022.
- PISANO, E. D. et al. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. **New England Journal of Medicine**, [S.I.], v.353, n.17, p.1773–1783, 2005.
- PODAREANU, D. et al. Best Practice Guide - Deep Learning. **Partnership for Advanced Computing in Europe**, [S.I.], 02 2019.
- RONNERBERGER, O.; FISCHER, P.; BROX, T. U-Net: convolutional networks for biomedical image segmentation. In: MEDICAL IMAGE COMPUTING AND COMPUTER-ASSISTED INTERVENTION – MICCAI 2015, Cham. **Anais...** Springer International Publishing, 2015. p.234–241.
- SAFFARI, N. et al. Fully Automated Breast Density Segmentation and Classification Using Deep Learning. **Diagnostics**, [S.I.], v.10, n.11, 2020.
- SAIA, R. et al. Credit Scoring by Leveraging an Ensemble Stochastic Criterion in a Transformed Feature Space. **Progress in Artificial Intelligence**, [S.I.], v.10, 05 2021.
- SAKIB, S. et al. An Overview of Convolutional Neural Network: its architecture and applications. **Preprints**, [S.I.], February 2019.
- SHARIFANI, K.; AMINI, M. Machine Learning and Deep Learning: a review of methods and applications. **World Information Technology and Engineering Journal**, [S.I.], v.10, n.07, p.3897–3904, 2023. Disponível em: <https://ssrn.com/abstract=4458723>.
- SHINDE, P. P.; SHAH, S. A Review of Machine Learning and Deep Learning Applications. In: FOURTH INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION CONTROL AND AUTOMATION (ICCUBEA), 2018. **Anais...** [S.I.: s.n.], 2018. p.1–6.
- SICKLES, E. et al. ACR BI-RADS® Mammography. In: **ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System**. Reston, VA: American College of Radiology, 2013.

- SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond Accuracy, F-Score and ROC: a family of discriminant measures for performance evaluation. In: AI 2006: ADVANCES IN ARTIFICIAL INTELLIGENCE, Berlin, Heidelberg. **Anais...** Springer Berlin Heidelberg, 2006. p.1015–1021.
- TAN, M.; LE, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. **CoRR**, [S.I.], v.abs/1905.11946, 2019.
- TING, K. M. Sensitivity and Specificity. In: ENCYCLOPEDIA OF MACHINE LEARNING, Boston, MA. **Anais...** Springer US, 2010. p.901–902.
- VRBANCIC, G.; PODGORELEC, V. Transfer Learning With Adaptive Fine-Tuning. **IEEE Access**, [S.I.], v.8, p.196197–196211, 2020.
- WINKLER, N. S. et al. Breast Density: clinical implications and assessment methods. **RadioGraphics**, [S.I.], v.35, n.2, p.316–324, 2015. PMID: 25763719.
- WU, M. et al. Detecting neonatal acute bilirubin encephalopathy based on T1-weighted MRI images and learning-based approaches. **BMC Medical Imaging**, [S.I.], v.21, 06 2021.
- YANG, X. et al. A Survey on Deep Semi-Supervised Learning. **IEEE Transactions on Knowledge and Data Engineering**, [S.I.], v.35, n.9, p.8934–8954, 2023.
- YI, P. et al. Deep-Learning-Based Semantic Labeling for 2D Mammography and Comparison of Complexity for Machine Learning Tasks. **Journal of Digital Imaging**, [S.I.], v.32, 06 2019.
- ZHANG, L.; WANG, S.; LIU, B. **Deep Learning for Sentiment Analysis : a survey**. 2018.
- ZHAO, W. et al. BASCNet: bilateral adaptive spatial and channel attention network for breast density classification in the mammogram. **Biomedical Signal Processing and Control**, [S.I.], v.70, p.103073, 2021.