

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

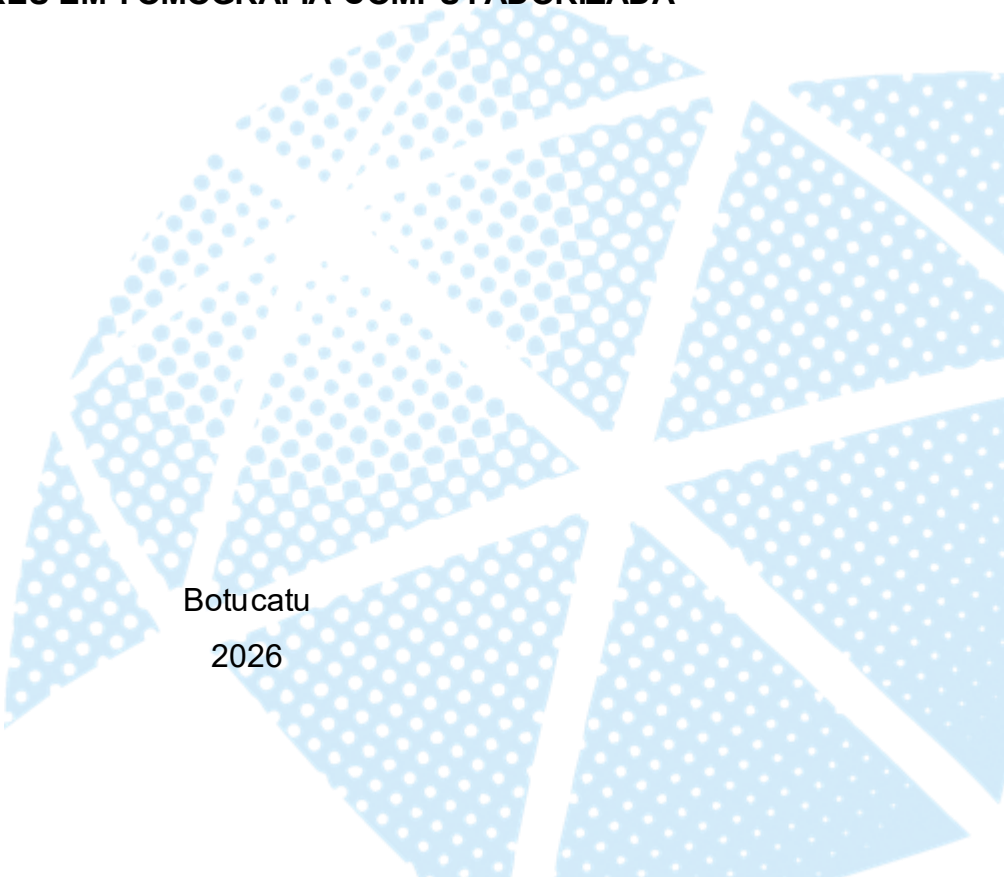
Instituto de Biociências - Campus de Botucatu

NYCOLAS MARIOTTO

***DEEP ENSEMBLE* PARA SEGMENTAÇÃO E CLASSIFICAÇÃO DE NÓDULOS
PULMONARES EM TOMOGRAFIA COMPUTADORIZADA**

Botucatu

2026



NYCOLAS MARIOTTO

***DEEP ENSEMBLE* PARA SEGMENTAÇÃO E CLASSIFICAÇÃO DE NÓDULOS
PULMONARES EM TOMOGRAFIA COMPUTADORIZADA**

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu, para obtenção do título de Mestre em Ciências Biomoleculares e Farmacológicas.

Área de Concentração: Física Biomolecular

Orientador(a): Prof. Dr. Allan Felipe Fattori
Alves

Botucatu

2026

M342d	Mariotto, Nycolas Deep ensemble para segmentação e classificação de nódulos pulmonares em tomografia computadorizada / Nycolas Mariotto. — Botucatu, 2026 63 p. : il., tabs. Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu Orientador: Allan Felipe Fattori Alves 1. Tomografia. 2. Deep learning (Machine learning). 3. Redes neurais (Computação). 4. Inteligência artificial. 5. Processamento de imagens. I. Título.
-------	--

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

Esta pesquisa apresenta impacto científico, técnico e social ao contribuir para o desenvolvimento de métodos baseados em *deep learning* e *ensemble* aplicados à segmentação e classificação de nódulos pulmonares em tomografia computadorizada, com potencial uso como ferramenta de apoio à decisão clínica em contextos de triagem.

POTENTIAL IMPACT OF THIS RESEARCH

This research presents scientific, technical, and social impact by contributing to the development of deep learning and ensemble-based methods for pulmonary nodule segmentation and classification in computed tomography, with potential application as a clinical decision support tool in triage settings.

NYCOLAS MARIOTTO

**DEEP ENSEMBLE PARA SEGMENTAÇÃO E CLASSIFICAÇÃO DE NÓDULOS
PULMONARES EM TOMOGRAFIA COMPUTADORIZADA**

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu, para obtenção do título de Mestre em Ciências Biomoleculares e Farmacológicas.

Área de Concentração: Física Biomolecular

Data da defesa: 10/02/2026

Banca Examinadora:

Documento assinado digitalmente
 **ALLAN FELIPE FATTORI ALVES**
Data: 13/03/2026 23:50:30-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Allan Felipe Fattori Alves
UNESP – Instituto de Biociências de Botucatu

Documento assinado digitalmente
 **SUSANA MARRERO IGLESIAS**
Data: 16/03/2026 11:36:41-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dra. Susana Marrero Iglesias
UESC - Universidade Estadual de Santa Cruz

Documento assinado digitalmente
 **FILIPPE WELSON LEAL PEREIRA**
Data: 17/03/2026 16:49:58-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Filipe Welson Leal Pereira
UNESP - Faculdade de Medicina de Botucatu

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter sido a minha fortaleza durante toda esta jornada.

Aos meus pais Marcos Ronaldo Mariotto e Silvana Teixeira Mariotto, à minha esposa Bruna da Cruz Carvalho Mariotto e à minha irmã Bruna Teixeira Oliveira que sempre me incentivaram e apoiaram nos meus sonhos.

Ao meu orientador Prof. Dr. Allan Felipe Fattori Alves que esteve comigo nesta etapa importante, tornando possível a conclusão desta dissertação. Obrigado por ter acreditado em mim e em todos os seus ensinamentos.

A Universidade Estadual Paulista “Júlio de Mesquita Filho” e ao Instituto de Biociências de Botucatu (IBB), que me permitiram realizar esse sonho.

Ao financiamento concedido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

A todos que contribuíram de alguma forma para o que sou.

“As pessoas não são perturbadas pelas coisas, mas pelos juízos que fazem sobre elas.”

Epicteto, *Enchiridion* (Manual), §5.

RESUMO

O câncer de pulmão é a principal causa de morte por câncer no mundo, em grande parte devido ao diagnóstico tardio da doença. Nesse contexto, a tomografia computadorizada desempenha papel central na detecção de nódulos pulmonares, que representam a forma mais frequente de apresentação da doença em estágios iniciais. No entanto, a análise desses exames é desafiadora em razão da variabilidade morfológica das lesões, da heterogeneidade dos protocolos de aquisição e da elevada carga de trabalho imposta aos especialistas. Este trabalho propõe e avalia um *pipeline* computacional baseado em redes neurais convolucionais e estratégias de *ensemble* para segmentação e classificação de nódulos pulmonares em tomografia computadorizada. Foram conduzidos três experimentos progressivos utilizando bases públicas amplamente empregadas na literatura. Inicialmente, realizou-se a classificação multiclasse de nódulos na base de dados LNDb, seguida pela classificação binária de nódulos malignos e benignos na base de dados LIDC-IDRI. Por fim, foi implementado um *pipeline end-to-end* que incorpora segmentação pulmonar e nodular antes da etapa de classificação. Os resultados indicaram que arquiteturas modernas de redes neurais convolucionais apresentam desempenho competitivo na classificação de nódulos pulmonares, enquanto estratégias de *ensemble* contribuem para maior robustez e estabilidade das predições. Em cenários mais complexos e heterogêneos, observou-se que o *ensemble* tende a favorecer métricas de sensibilidade, como o *recall*, em detrimento de métricas globais de desempenho. Esses achados reforçam o potencial de modelos computacionais como ferramentas de apoio à decisão clínica, especialmente em contextos de triagem e priorização visual, sem substituir a atuação do radiologista.

Palavras-chave: tomografia computadorizada; nódulos pulmonares; segmentação; classificação; deep ensemble; redes neurais convolucionais; LIDC-IDRI; LNDb.

ABSTRACT

Lung cancer is the leading cause of cancer death worldwide, largely due to late-stage diagnosis. In this context, computed tomography plays a central role in the detection of pulmonary nodules, which represent the most common radiological manifestation of the disease at early stages. However, the analysis of these examinations is challenging because of lesion morphological variability, heterogeneous acquisition protocols, and the high workload imposed on specialists. This study proposes and evaluates a computational pipeline based on convolutional neural networks and deep ensemble strategies for pulmonary nodule segmentation and classification in computed tomography. Three progressive experimental scenarios were conducted using publicly available datasets widely adopted in the literature. First, a multiclass classification of pulmonary nodules was performed on the LNDb dataset. Subsequently, binary classification of malignant and benign nodules was carried out using the LIDC-IDRI dataset. Finally, an end-to-end pipeline incorporating lung and nodule segmentation before classification was implemented. The results demonstrate that modern convolutional neural network architectures achieve competitive performance in pulmonary nodule classification, while ensemble strategies contribute to increased robustness and prediction stability. In more complex and heterogeneous scenarios, the ensemble approach tended to favor sensitivity-related metrics, such as recall, at the expense of global performance measures. These findings highlight the potential of computational models as clinical decision support tools, particularly in triage and visual prioritization tasks, without replacing the role of the radiologist.

Keywords: computed tomography; pulmonary nodules; segmentation; classification; deep ensemble; convolutional neural networks; LIDC-IDRI; LNDb

LISTA DE FIGURAS

Figura 1. Números absolutos de incidência e mortalidade por câncer, ambos os sexos, no mundo, em 2022	15
Figura 2. Números absolutos de incidência e mortalidade por câncer, ambos os sexos, no Brasil, em 2022	16
Figura 3. Exemplo de funcionamento de uma tomografia computadorizada.....	19
Figura 4. Escala ilustrativa de unidades HU.....	20
Figura 5. Representação esquemática do conceito de janelamento em tomografia computadorizada.....	21
Figura 6. Efeito da largura da janela na visualização de uma tomografia computadorizada.....	21
Figura 7. Diferença visual entre essas duas configurações de janelamento em uma tomografia computadorizada de tórax	22
Figura 8. Exemplos de tomografias computadorizadas de tórax	23
Figura 9. Estrutura geral de uma rede neural convolucional aplicada à classificação de imagens.....	28
Figura 10. Esquema conceitual de um <i>ensemble</i> por <i>Soft Voting</i> de redes neurais convolucionais.....	31
Figura 11. Fluxograma metodológico geral adotado.....	34
Figura 12. Síntese esquemática dos três experimentos realizados	35
Figura 13. Distribuições de características anotadas pelos radiologistas na base de dados LIDC-IDRI.....	38
Figura 14. Exemplos de técnicas de <i>data augmentation</i> aplicadas a cortes de tomografia computadorizada de tórax	38
Figura 15. Sobreposição das quatro anotações de um mesmo nódulo e máscara consensual de 50%	39
Figura 16. Exemplos da variabilidade morfológica dos nódulos pulmonares presentes na base de dados LIDC-IDRI.....	41
Figura 17. Exemplo do <i>pipeline</i> em resolução reduzida.....	42
Figura 18. Exemplo do <i>pipeline</i> de segmentação aplicado a uma tomografia computadorizada de tórax.	42
Figura 19. Desempenho em acurácia dos modelos individuais de CNN avaliados na classificação multiclasse de nódulos pulmonares na base LNDb	46

Figura 20. Acurácia dos principais modelos individuais e do <i>ensemble</i> na classificação binária de nódulos pulmonares na base de dados LIDC-IDRI	48
Figura 21. Comparação das métricas de desempenho entre os Experimentos 1 e 2	51
Figura 22. Diferença de desempenho entre o <i>ensemble</i> e o melhor modelo individual nas métricas avaliadas nos Experimentos 1 e 2.....	51

LISTA DE TABELAS

Tabela 1. Comparação conceitual entre radiômica e <i>deep learning</i> aplicados à análise de imagens médicas.....	27
Tabela 2. Desempenho dos modelos individuais e do <i>ensemble</i> na classificação multiclasse de nódulos pulmonares na base de dados LNDb.	47
Tabela 3. Desempenho dos modelos no <i>pipeline end-to-end</i> com segmentação prévia de nódulos pulmonares na base de dados LIDC-IDRI.	49
Tabela 4. Comparação do melhor desempenho individual e do <i>ensemble</i> nos três experimentos	50

LISTA DE ABREVIATURAS E SIGLAS

ACR	American College of Radiology
ADAM	Adaptive Moment Estimation
ADAMW	Adam with Decoupled Weight Decay
BCE	Binary Cross-Entropy
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
Dice	Dice Similarity Coefficient
FBP	Filtered Back Projection
GCO	Global Cancer Observatory
GGO	Ground Glass Opacity (Padrão de Vidro Fosco)
GLOBOCAN	Global Cancer Observatory Database
HU	Hounsfield Unit
IARC	International Agency for Research on Cancer
IoU	Intersection over Union
LIDC-IDRI	Lung Image Database Consortium and Image Database Resource Initiative
LNDb	Lung Nodule Database
Lung-RADS	Lung Imaging Reporting and Data System
NLST	National Lung Screening Trial
TCIA	The Cancer Imaging Archive
VGG	Visual Geometry Group

SUMÁRIO

1	INTRODUÇÃO	15
2	DESENVOLVIMENTO.....	18
2.1	REVISÃO DE LITERATURA.....	18
2.1.1	Tomografia computadorizada no câncer de pulmão.....	18
2.1.2	Nódulos pulmonares: caracterização e manejo clínico	23
2.1.3	Métodos computacionais em imagens médicas	24
2.1.4	Arquiteturas de redes neurais convolucionais.....	28
2.1.5	<i>Ensembles</i> e métricas de avaliação de desempenho.....	29
2.1.5.1	Estratégias de combinação	31
2.1.5.2	Métricas de avaliação de desempenho.....	33
2.2	MATERIAIS E MÉTODOS	33
2.2.1	Bases de dados.....	36
2.2.2	Preparação e curadoria dos dados	37
2.2.3	Pré-processamento e segmentação	41
2.2.4	Treinamento e <i>ensemble</i> de modelos	44
2.2.5	Avaliação experimental.....	45
2.3	RESULTADOS.....	46
2.3.1	Classificação multiclasse de nódulos pulmonares na base de dados LNDb (Experimento 1).....	46
2.3.2	Classificação binária de nódulos pulmonares na base de dados LIDC-IDRI (Experimento 2).....	47
2.3.3	<i>Pipeline end-to-end</i> com segmentação prévia na base de dados LIDC-IDRI (Experimento 3).....	49
2.3.4	Síntese comparativa e análise integrada dos experimentos	50
2.4	DISCUSSÃO.....	52
3.	CONCLUSÃO.....	55
	REFERÊNCIAS.....	57
	DADOS CURRICULARES.....	61

1 INTRODUÇÃO

O câncer de pulmão é a principal causa de morte por câncer no mundo, com aproximadamente 1,8 milhão de óbitos anuais. Conforme apresentado na Figura 1, dados da *International Agency for Research on Cancer* (IARC) indicam que, em 2022, o câncer de pulmão foi a neoplasia mais incidente e a mais letal globalmente¹.

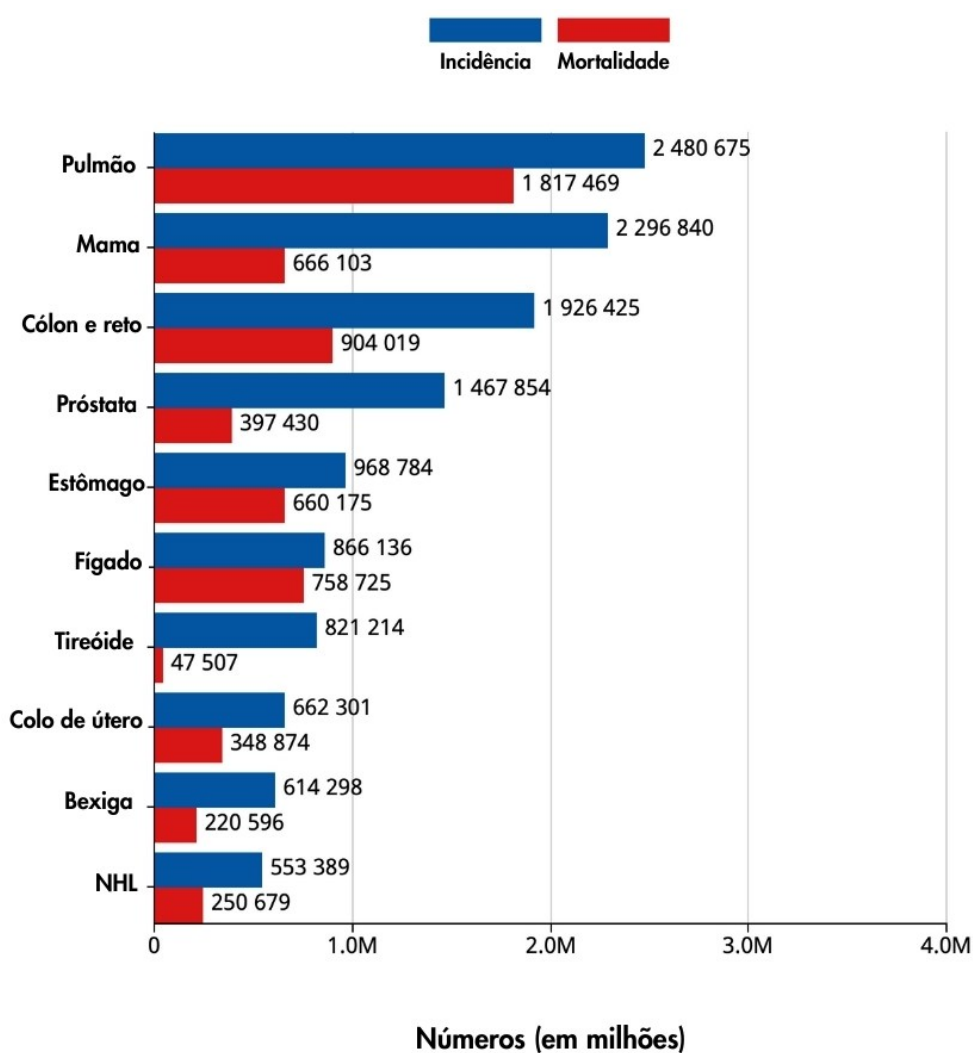


Figura 1. Números absolutos de incidência e mortalidade por câncer, ambos os sexos, no mundo, em 2022. Fonte: Adaptado de Global Cancer Observatory (GCO) – Cancer Today; GLOBOCAN 2022 (v1.1; 2024-02-08). International Agency for Research on Cancer (IARC).

No contexto brasileiro, observa-se um padrão semelhante, com elevada letalidade associada à doença, como ilustrado na Figura 2, reforçando a necessidade de estratégias eficazes de detecção precoce tanto em escala global quanto nacional^{1,2}.

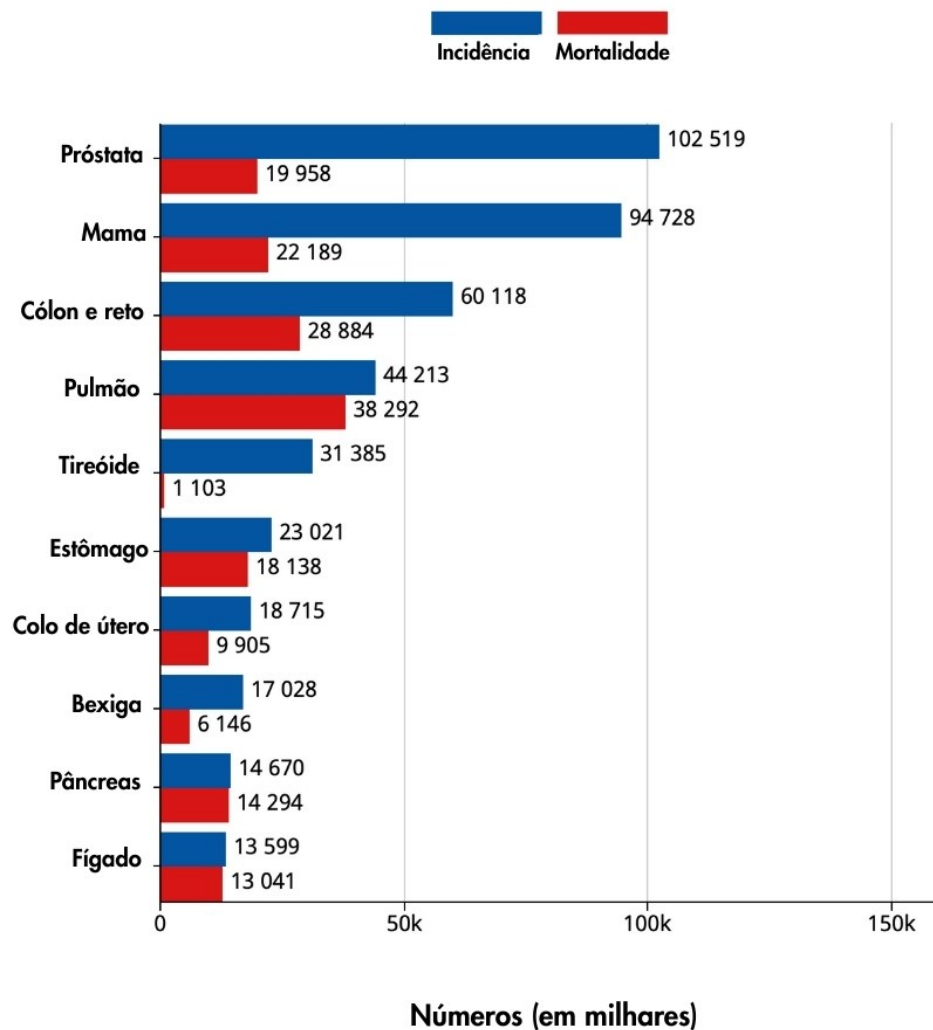


Figura 2. Números absolutos de incidência e mortalidade por câncer, ambos os sexos, no Brasil, em 2022. Fonte: Adaptado e traduzido de Global Cancer Observatory (GCO) – Cancer Today; GLOBOCAN 2022 (v1.1; 2024-02-08). International Agency for Research on Cancer (IARC).

Ensaio clínico randomizado, como o *National Lung Screening Trial* (NLST), demonstraram redução significativa da mortalidade por câncer de pulmão com o rastreamento por tomografia computadorizada de baixa dose, consolidando essa modalidade como ferramenta fundamental para a detecção em estágios iniciais³.

Embora o câncer de pulmão possa se manifestar sob diferentes padrões radiológicos — incluindo massas extensas, consolidações e opacidades difusas em vidro fosco — os nódulos pulmonares representam a forma mais frequente de apresentação em estágios iniciais da doença, especialmente em exames de rastreamento por tomografia computadorizada. A identificação e caracterização de nódulos pulmonares têm relevância clínica central, uma vez que lesões pequenas e assintomáticas podem corresponder a neoplasias em fase inicial, nas quais o diagnóstico precoce impacta diretamente o prognóstico e a sobrevida dos pacientes³.

Além disso, protocolos clínicos consolidados, como os sistemas de estratificação do *American College of Radiology* (ACR), são especificamente direcionados à avaliação de nódulos pulmonares, reforçando a relevância desse padrão radiológico como alvo prioritário para o desenvolvimento de métodos computacionais de apoio à decisão⁴.

Na tomografia computadorizada, os nódulos pulmonares são visualizados como opacidades arredondadas de até 3 cm e incluem morfologias sólidas, parcialmente sólidas e de padrão vidro fosco. A detecção diagnóstica é desafiadora devido à semelhança com outras estruturas, ao baixo contraste e à variabilidade morfológica, fatores que podem elevar o risco de falsos positivos⁵. Além disso, a variabilidade interobservador afeta a categorização por sistemas padronizados, como o *Lung Imaging Reporting and Data System* (Lung-RADS), desenvolvido pelo ACR, influenciando decisões diagnósticas e condutas subsequentes adotadas pelos médicos especialistas⁶.

Em resposta a esses desafios, surgiram os primeiros sistemas de diagnóstico auxiliado por computador (*Computer-Aided Design*, CAD), que buscavam automatizar a detecção de nódulos. No entanto, estudos iniciais demonstraram que esses sistemas apresentavam sensibilidade limitada, especialmente para nódulos menores, e não superaram significativamente a análise por radiologistas experientes, funcionando mais como um segundo leitor⁷. A evolução para técnicas de *machine learning* e, posteriormente, para o *deep learning*, representou um avanço substancial. Diferentes dos sistemas CAD tradicionais baseados em regras, os modelos de *deep learning* podem aprender a identificar características complexas e sutis diretamente dos dados de imagem, abordando os problemas de baixo contraste e variabilidade morfológica⁸. Essa capacidade de extração automática de características abriu caminho para o desenvolvimento de ferramentas mais robustas e precisas.

Redes neurais convolucionais (CNNs) são um exemplo de ferramentas tecnológicas que podem auxiliar em tarefas computacionais relacionadas ao diagnóstico de doenças pulmonares, incluindo nódulos pulmonares. As CNNs permitem aprender representações hierárquicas diretamente dos *voxels*, dispensando a construção manual de atributos. Em tomografia computadorizada torácica, há evidência de bom desempenho para detecção, segmentação e estratificação de risco de nódulos pulmonares, com potencial para reduzir carga de trabalho e padronizar decisões diagnósticas⁸. Para robustez e quantificação de incertezas, aspectos críticos em aplicações clínicas, técnicas denominadas *ensembles* permitem a combinação de

predições de múltiplas redes e tendem a melhorar a generalização ao mesmo tempo em que fornecem medidas úteis de incerteza⁹.

Diante desse panorama, emergem questões centrais que orientam o desenvolvimento deste trabalho. Em particular, busca-se compreender em que medida arquiteturas modernas de *deep learning* apresentam desempenho consistente na classificação de nódulos pulmonares quando avaliadas em bases públicas com diferentes níveis de complexidade e heterogeneidade. Além disso, investiga-se se estratégias baseadas em *ensemble* contribuem para maior robustez e sensibilidade diagnóstica em comparação a modelos individuais, bem como qual é o impacto da segmentação prévia no desempenho final da classificação. Ao abordar essas questões de forma progressiva, por meio de experimentos conduzidos em diferentes cenários e bases de dados, este estudo visa contribuir tanto para o avanço metodológico em visão computacional aplicada à saúde quanto para o desenvolvimento de ferramentas de apoio à decisão clínica em tomografia computadorizada.

Este trabalho organiza, de forma reprodutível, um *pipeline* aplicado à segmentação e à classificação de nódulos pulmonares em exames de TOMOGRAFIA COMPUTADORIZADA. O escopo abrange pré-processamento e *data augmentation*, teste de diferentes arquiteturas, estratégias de treinamento, validação e análise crítica de métricas. Por fim, avaliamos modelos baseados em CNNs, incluindo estratégias de *ensemble*, como apoio ao diagnóstico de câncer de pulmão.

2 DESENVOLVIMENTO

2.1 REVISÃO DE LITERATURA

2.1.1 Tomografia computadorizada no câncer de pulmão

A tomografia computadorizada é uma das principais modalidades do diagnóstico por imagem, cujo princípio físico baseia-se na medição da atenuação de fótons de raios X ao atravessarem o volume de interesse em múltiplos ângulos. Os padrões de atenuação registrados pelos detectores de radiação originam as projeções, que estão relacionadas ao coeficiente de atenuação linear (μ) de cada *voxel*. Essas projeções compõem o sinograma e podem ser reconstruídas em

imagens axiais por métodos analíticos, como a retroprojeção filtrada (*filtered backprojection*, FBP), ou por métodos iterativos¹⁰. Em sistemas helicoidais, os dados são adquiridos de forma contínua enquanto o tubo e os detectores giram ao redor do paciente, com interpolação ao longo do eixo z, conforme ilustrado na Figura 3.

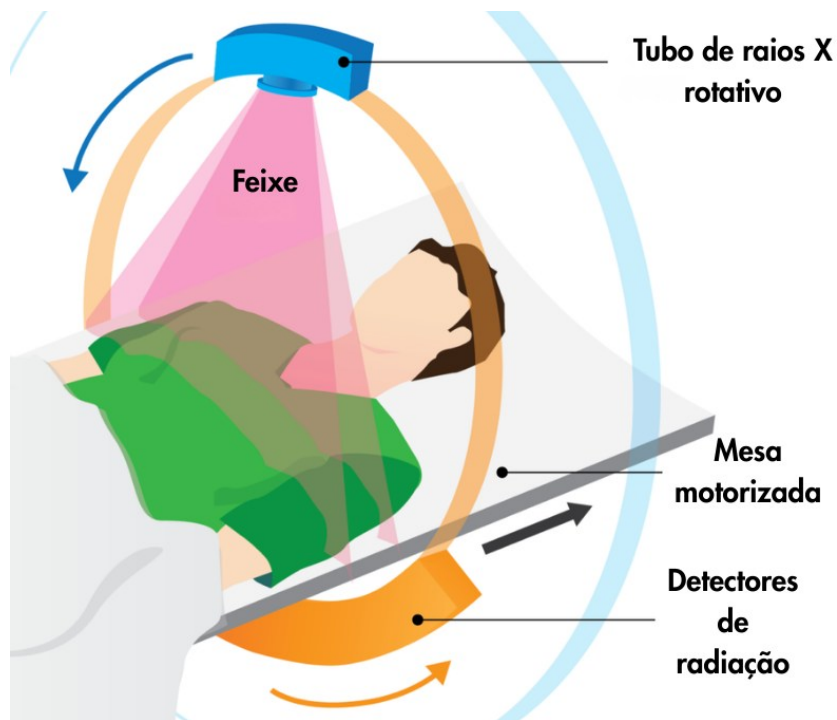


Figura 3. Exemplo de funcionamento de uma tomografia computadorizada helicoidal. Fonte: Adaptado de Lecturio. Tomografia Computorizada (TC).

A imagem reconstruída por TOMOGRAFIA COMPUTADORIZADA é composta por *voxels* cujos valores de intensidade são expressos em unidades *Hounsfield* (HU), que refletem a atenuação relativa dos tecidos em relação à água, sendo representada por 0 HU e o ar por -1000 HU. Tecidos moles geralmente situam-se entre 30 e 60 HU, a gordura apresenta valores inferiores a -50 HU e o osso cortical frequentemente excede +1000 HU¹¹. A Figura 4 apresenta uma ilustração da escala de unidades HU aplicada a diferentes materiais e tecidos biológicos.

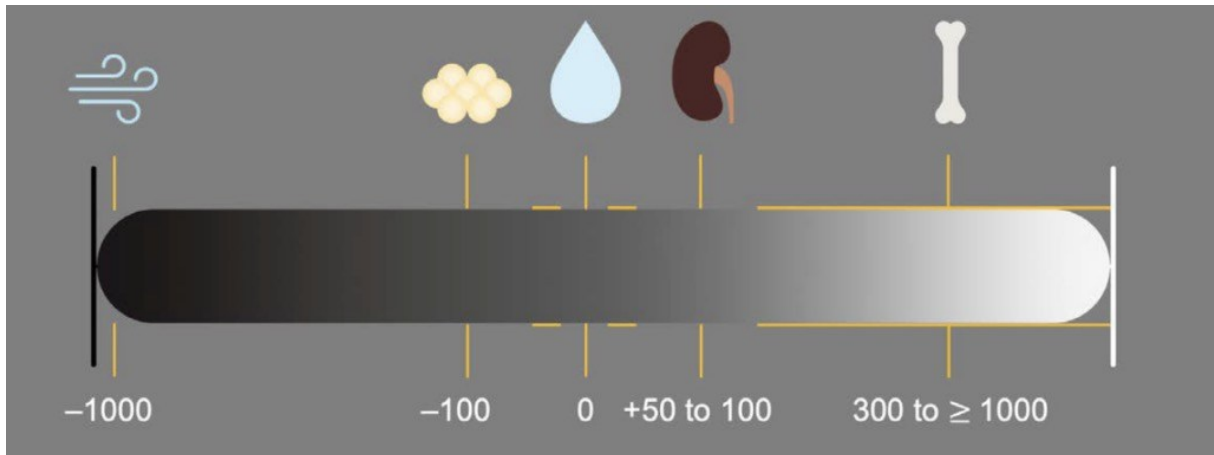


Figura 4. Escala ilustrativa de unidades HU. Fonte: Extraído de Hartung & Brown, LITFL.

A qualidade e a interpretabilidade das imagens de TOMOGRAFIA COMPUTADORIZADA dependem não apenas da reconstrução, mas também de parâmetros de aquisição e processamento, incluindo resolução espacial, contraste, janelamento, espessura de corte, nível de ruído e presença de artefatos¹². A resolução espacial está associada à capacidade de distinguir estruturas adjacentes, enquanto o contraste reflete a diferença de atenuação entre tecidos. O janelamento ajusta a faixa de valores de HU exibida, permitindo realçar estruturas específicas de acordo com o objetivo diagnóstico.

Do ponto de vista operacional, o janelamento é definido por dois parâmetros fundamentais: a largura da janela (*window width*) e o nível da janela (*window level*)¹². A largura da janela estabelece o intervalo de valores de unidades HU que será mapeado para a escala de cinza exibida, determinando o contraste global da imagem. Janelas mais largas acomodam uma faixa maior de atenuações, reduzindo o contraste entre tecidos com valores próximos, enquanto janelas mais estreitas aumentam o contraste, à custa da saturação de valores fora desse intervalo.

O nível da janela corresponde ao valor central de HU em torno do qual a largura é distribuída e define quais tecidos são priorizados na visualização. Ajustes no nível deslocam a janela ao longo da escala de HU, permitindo enfatizar diferentes estruturas anatômicas conforme sua densidade relativa. A relação entre a largura e o nível da janela e seu impacto na visualização das intensidades de TOMOGRAFIA COMPUTADORIZADA são ilustrados na Figura 5.

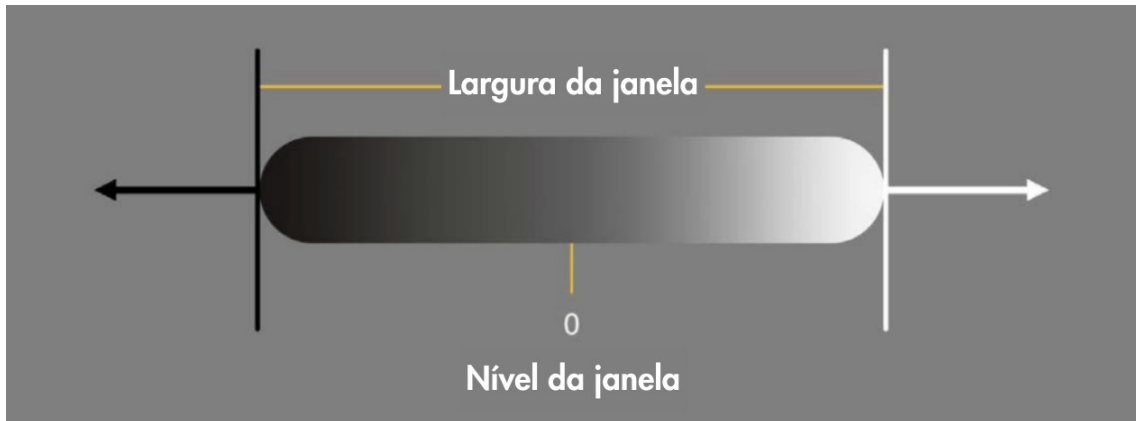


Figura 5. Representação esquemática do conceito de janelamento em tomografia computadorizada. Fonte: Adaptado de Hartung & Brown, LITFL.

A combinação entre largura e nível da janela influencia diretamente a distribuição das intensidades na escala de cinza e, conseqüentemente, a percepção de contraste na imagem reconstruída¹². Para um mesmo nível de janela, larguras mais amplas comprimem diferenças de atenuação, resultando em imagens com menor contraste aparente, enquanto larguras mais estreitas expandem essas diferenças, realçando variações sutis entre tecidos com valores de HU próximos. Esse efeito é ilustrado na Figura 6, que compara a visualização de uma mesma região anatômica sob diferentes larguras de janela.



Figura 6. Efeito da largura da janela na visualização de uma tomografia computadorizada. Fonte: Adaptado de Hartung & Brown, LITFL.

Na avaliação torácica, utilizam-se janelas específicas para otimizar a visualização das diferentes estruturas. A janela pulmonar, tipicamente definida por largura em torno de 1.600 HU e nível aproximado de -600 HU, é adequada para o

parênquima pulmonar e para a detecção de nódulos, enquanto a janela mediastinal privilegia tecidos moles e estruturas vasculares, como exemplificado na Figura 7¹².

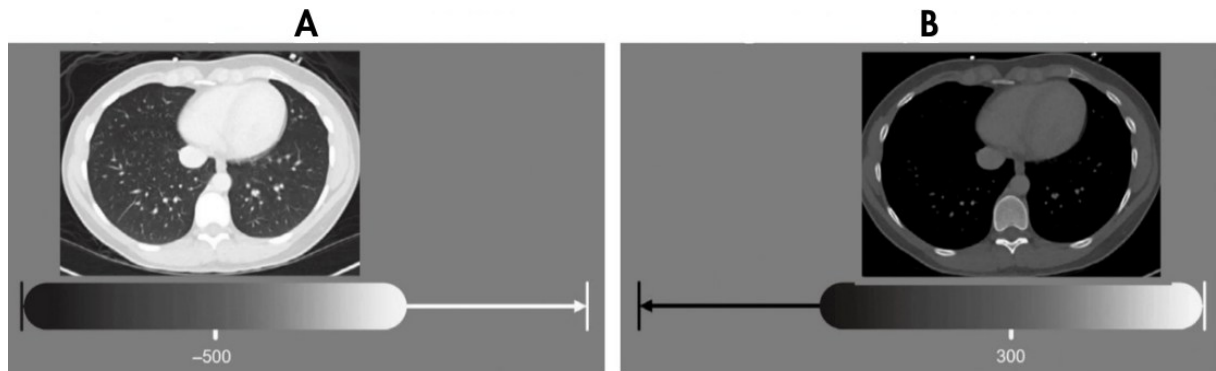


Figura 7. Diferença visual entre essas duas configurações de janelamento em uma tomografia computadorizada de tórax. (A) Comparação entre janelas pulmonares e (B) mediastinais em uma tomografia computadorizada de tórax. Fonte: Adaptado de Hartung & Brown, LITFL.

Além do janelamento, a qualidade da imagem em tomografia computadorizada é influenciada por ruídos e artefatos. O ruído decorre principalmente da estatística dos fótons detectados e tende a aumentar com cortes finos, baixa carga de corrente-tempo (mAs) e o uso de *kernels* de reconstrução mais agudos. Estratégias de mitigação incluem a otimização dos parâmetros de aquisição e o emprego de métodos de reconstrução iterativa¹³.

Artefatos frequentes incluem enrijecimento do feixe, com escurecimento central e estrias, efeito de volume parcial, movimento respiratório e presença de metais e artefatos em anel. A redução desses efeitos envolve a escolha adequada de parâmetros de aquisição, correções específicas durante a reconstrução, sincronização respiratória e técnicas dedicadas à redução de artefatos¹³.

Compreendidos os fundamentos técnicos da formação e da qualidade da imagem de tomografia computadorizada, torna-se possível abordar as características morfológicas e radiológicas dos nódulos pulmonares, que constituem o principal alvo de análise neste estudo. A tomografia computadorizada de baixa dose caracteriza-se pela aquisição de imagens com redução significativa da dose de radiação ionizante, mantendo qualidade diagnóstica suficiente para a detecção de nódulos pulmonares. Evidências provenientes de estudos clínicos randomizados de grande escala demonstram redução da mortalidade por câncer de pulmão quando o rastreamento por tomografia computadorizada de baixa dose é aplicado sob critérios de elegibilidade e protocolos bem definidos³.

Nesse contexto, diretrizes como as da *Fleischner Society* (2017) fornecem recomendações para o manejo de nódulos incidentes em adultos, considerando dimensões, morfologia e fatores de risco, e orientam o seguimento clínico e a conduta adequada¹⁴. Esse conjunto de evidências sustenta a importância de fluxos analíticos reprodutíveis e de métricas robustas para avaliação de desempenho em cenários de triagem e diagnóstico.

2.1.2 Nódulos pulmonares: caracterização e manejo clínico

Os nódulos pulmonares são achados radiológicos que indicam anormalidades focais do parênquima pulmonar. Esses podem ser classificados, do ponto de vista radiológico, em sólidos, parcialmente sólidos e em padrão de vidro fosco (*Ground Glass Opacity*, GGO), categorias associadas a diferentes probabilidades de malignidade, crescimento e comportamento biológico, o que reforça a complexidade da tarefa de classificação automática¹⁵. A Figura 8 ilustra exemplos de nódulo sólido, parcialmente sólido e vidro fosco em janelas de HU usuais para avaliação pulmonar.

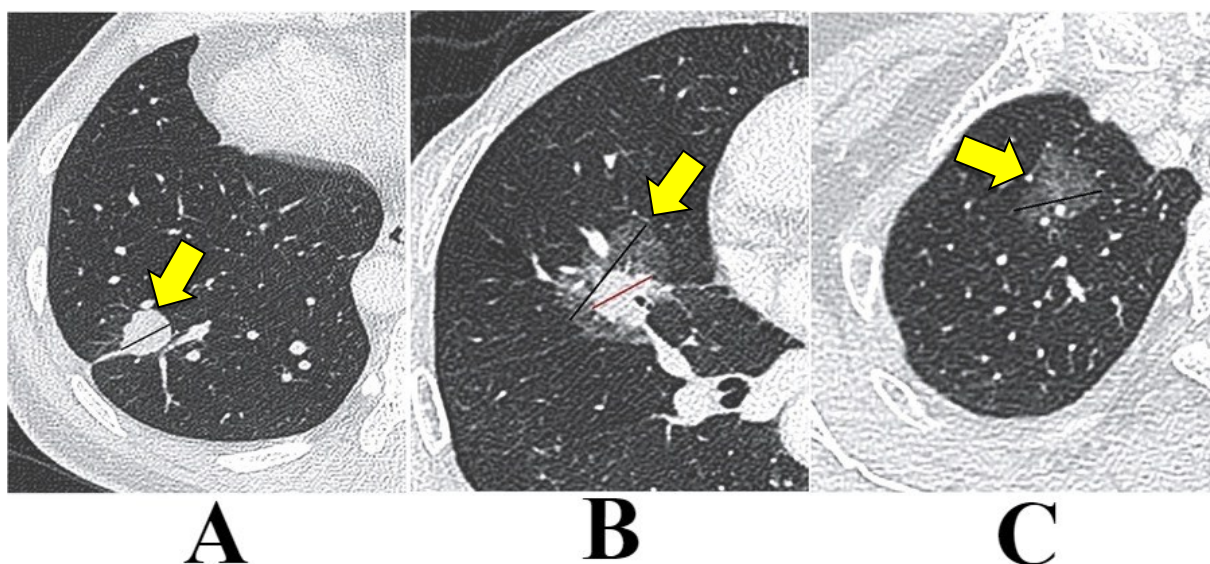


Figura 8. Exemplos de tomografias computadorizadas de tórax. (A) Exemplos de nódulo sólido, (B) parcialmente sólido e (C) padrão de vidro fosco em uma tomografia computadorizada de tórax. Fonte: Adaptado de Ye et al., 2019, Fig. 1 (Journal of Thoracic Oncology).

A estratificação adequada é crítica porque a doença metastática está associada a desfechos piores em comparação com casos localizados ou regionalizados, o que justifica protocolos rigorosos de acompanhamento e indicação

de biópsia¹⁶. Além disso, a interpretação diagnóstica pode apresentar variabilidade interobservador, com impacto na detecção e na caracterização de nódulos, especialmente em lesões pequenas ou parcialmente sólidas¹⁷. Diante dessas limitações humanas e da crescente demanda por padronização diagnóstica, o desenvolvimento de ferramentas computacionais tem emergido como uma alternativa promissora para auxiliar na análise, na estratificação do risco de nódulos pulmonares e no acompanhamento progressivo de alguns casos, oferecendo uma quantificação, em área ou volume, bem precisa das estruturas de interesse.

2.1.3 Métodos computacionais em imagens médicas

No suporte computacional à decisão em imagens médicas, destacam-se duas vertentes complementares. A radiômica, que se baseia na extração sistemática de atributos quantitativos definidos manualmente, capazes de capturar padrões de textura, forma e intensidade¹⁸. Em contraste, os métodos de *deep learning* aprendem representações hierárquicas diretamente dos dados de entrada, dispensando a engenharia manual de características, e têm demonstrado desempenho competitivo em tarefas de detecção, segmentação e classificação de nódulos pulmonares em tomografia computadorizada, conforme evidenciado por desafios padronizados e estudos clínicos¹⁹. Esses avanços motivam a construção de *pipelines* reprodutíveis que integrem pré-processamento, segmentação de suporte e modelos de classificação robustos.

Entre as abordagens de *deep learning*, as CNNs constituem o arcabouço mais empregado para análise de imagens¹⁹. Essas exploram operações de convolução e compartilhamento de pesos para aprender padrões espaciais locais, como bordas, texturas e estruturas anatômicas, organizando-os em representações progressivamente mais abstratas ao longo da rede. Essa propriedade as torna particularmente adequadas para o processamento de imagens médicas, nas quais a informação espacial é determinante para a interpretação diagnóstica.

A aplicação de CNNs envolve a definição de uma função de perda apropriada à tarefa e de um método de otimização para o ajuste dos parâmetros do modelo. Em tarefas de classificação binária, a função de perda mais utilizada é a entropia cruzada binária (*binary cross-entropy* ou BCE). Em cenários com desbalanceamento de classes, versões ponderadas da BCE ou a perda focal podem

ser empregadas para aumentar o peso de exemplos raros ou difíceis. Em segmentação semântica, perdas por *pixel* são frequentemente combinadas com métricas de sobreposição, como o coeficiente de *Dice* e a *Intersection over Union* (IoU), de modo a equilibrar a otimização local e regional das regiões de interesse²⁰.

O processo de otimização é conduzido por algoritmos que ajustam iterativamente os pesos da rede com o objetivo de minimizar a função de perda. Além do gradiente descendente estocástico, otimizadores adaptativos como o *Adaptive Moment Estimation* (Adam) e *Adam with Decoupled Weight Decay* (AdamW) ajustam o passo de atualização de cada parâmetro com base no histórico dos gradientes, o que geralmente resulta em convergência mais rápida e menor necessidade de ajuste manual de hiperparâmetros. O AdamW, em particular, desacopla o decaimento de peso do termo de atualização dos gradientes, promovendo uma regularização mais adequada e contribuindo para a redução do sobreajuste^{21,22}.

A taxa de aprendizado controla a magnitude das atualizações dos pesos e exerce influência direta na estabilidade e na velocidade do treinamento. Valores excessivamente altos podem levar à instabilidade, enquanto valores muito baixos tornam o processo lento ou propenso a mínimos locais. Estratégias comuns incluem aquecimento inicial da taxa de aprendizado nas primeiras épocas, seguido de esquemas de redução progressiva, como decaimento baseado em platô ou decaimento cosseno. Técnicas adicionais, como normalização em lote (*batch normalization*) e *dropout*, contribuem para a estabilização do treinamento e a redução do sobreajuste^{23,24}.

A avaliação confiável de modelos em imagens médicas requer um particionamento rigoroso dos dados. O conjunto de treinamento é utilizado para o ajuste dos pesos, o de validação orienta a escolha de hiperparâmetros e o ponto de parada, e o conjunto de teste permanece reservado para estimar o desempenho final em dados não vistos. A estratificação preserva a proporção de classes em cada partição, aspecto relevante em cenários desbalanceados. Quando há amostra suficiente, esquemas *holdout* simples são comuns, enquanto em bases menores emprega-se validação cruzada em k-dobras. Em todos os casos, a divisão em nível de paciente é fundamental para evitar vazamento de informação. Esses princípios gerais de treinamento, otimização e validação fornecem a base conceitual para a aplicação prática das CNNs.

Na literatura, observa-se uma transição progressiva de abordagens baseadas em radiômica, fundamentadas na extração manual de atributos quantitativos e modelagem estatística, para métodos de *deep learning*, em especial CNNs, que aprendem representações diretamente das imagens e apresentam desempenho competitivo em tarefas de detecção, segmentação e classificação de nódulos pulmonares em tomografia computadorizada¹⁹. Nesse contexto, abordagens baseadas em *ensemble* têm sido propostas como estratégia para aumentar a robustez e reduzir a variância entre modelos, combinando arquiteturas distintas e apresentando ganhos dependentes do protocolo experimental, do particionamento por paciente e da heterogeneidade entre bases de dados²⁵. Esses aspectos motivam o desenho experimental progressivo adotado neste trabalho, comparando diferentes cenários em bases públicas e avaliando o impacto da segmentação prévia no desempenho dos classificadores. A Tabela 1 apresenta uma comparação conceitual entre abordagens baseadas em radiômica e *deep learning* aplicadas à análise de imagens médicas.

Tabela 1. Comparação conceitual entre radiômica e deep learning aplicados à análise de imagens médicas.

Aspecto	Radiômica	Deep learning
Definição	Extração manual de medidas numéricas da imagem (ex.: forma, textura, densidade)	Aprendizado automático de padrões diretamente das imagens
Funcionamento	Especialistas definem quais características medirem → modelo estatístico ou <i>machine learning</i> clássico	A rede aprende sozinha quais padrões são relevantes durante o treinamento
Quantidade de dados	Geralmente menor	Geralmente maior
Interpretação dos resultados	Mais interpretável (atributos explícitos)	Menor interpretabilidade direta
Sensibilidade ao protocolo	Alta (segmentação e aquisição influenciam fortemente).	Moderada a alta, dependendo do treinamento.
Desempenho em tarefas complexas	Limitado quando padrões são sutis.	Elevado em detecção, segmentação e classificação.
Uso típico em TOMOGRAFIA COMPUTADORIZADA pulmonar	Análise exploratória, biomarcadores.	<i>Pipelines end-to-end</i> e sistemas de apoio à decisão (CAD)
Custo computacional	Baixo/moderado (extração + <i>machine learning</i> clássico)	Alto (treino/inferência), especialmente 3D
Pontos fortes	Controle de variáveis, comparabilidade e rastreabilidade das decisões	Capta padrões sutis e interações complexas; escalável para pipelines multietapas
Limitações	Pode perder padrões não modelados; depende de <i>feature design</i>	Precisa de muitos dados e poder computacional; risco de sobreajuste e dependência do conjunto

Fonte: Elaborado pelo autor.

2.1.4 Arquiteturas de redes neurais convolucionais

As CNNs podem assumir diferentes arquiteturas, que variam em profundidade, conectividade e estratégia de extração de características, impactando diretamente o desempenho, a estabilidade do treinamento e o custo computacional. Essas arquiteturas organizam-se, em geral, em estágios sucessivos de convolução e redução de dimensionalidade para extração automática de características, seguidos por camadas totalmente conectadas ou cabeças de classificação responsáveis pela predição final. A Figura 9 ilustra a estrutura genérica de uma CNN aplicada à classificação de imagens médicas, destacando as etapas de extração de características e classificação.

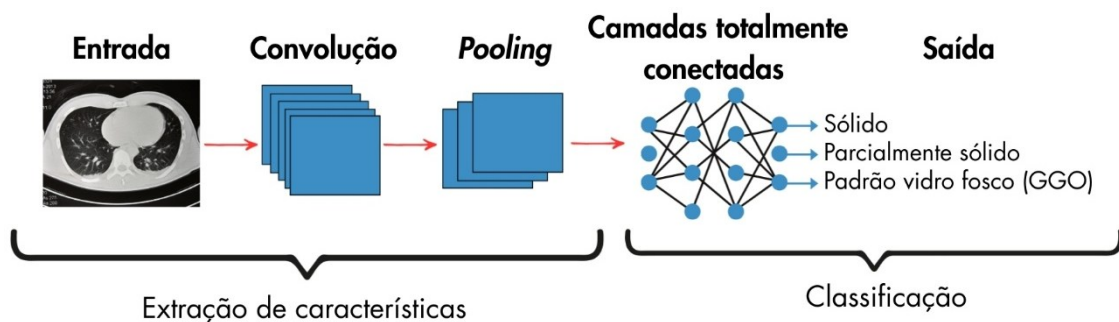


Figura 9. Estrutura geral de uma rede neural convolucional aplicada à classificação de imagens. Fonte: Elaborado pelo autor.

Entre as diversas arquiteturas de CNNs propostas na literatura, as redes da família ResNet (Residual Network) caracterizam-se pelo uso de conexões residuais, nas quais a entrada de um bloco é somada à sua saída²⁶. Essa estratégia facilita o fluxo do gradiente durante o treinamento e permite o aprendizado eficaz de redes profundas, mitigando problemas como o desvanecimento do gradiente. Os blocos mais utilizados seguem a configuração 1×1 , 3×3 e 1×1 , com um caminho de identidade em paralelo, o que viabiliza o treinamento estável de modelos profundos como ResNet-50 ou ResNet-101. Em imagens médicas, as ResNets são amplamente empregadas tanto como classificadores quanto como extratores de características em tarefas de detecção e segmentação. A principal vantagem prática reside na otimização previsível e na boa relação entre profundidade e desempenho, enquanto o custo está associado ao maior número de parâmetros e ao consumo de memória ao processar imagens de alta resolução.

Como alternativa às arquiteturas residuais, a família DenseNet (*Densely Connected Convolutional Network*) distingue-se por conectar cada camada a todas as anteriores por meio da concatenação dos mapas de características, promovendo reutilização de informações e melhor propagação do gradiente ao longo da rede²⁷. A largura do modelo é controlada por um parâmetro de crescimento, que define quantos novos canais cada camada adiciona. Em geral, as DenseNets alcançam elevado desempenho com menor número de parâmetros quando comparadas a arquiteturas tradicionais, característica particularmente vantajosa em bases médicas menores, nas quais a generalização é um desafio. Como limitação, destaca-se o consumo de memória, que cresce com a resolução espacial e com a concatenação progressiva dos mapas de características, exigindo escolhas cuidadosas do tamanho de entrada e das configurações de transição entre blocos.

Para tarefas de segmentação, a U-Net constitui uma das arquiteturas de referência na literatura²⁸. Essa rede apresenta um caminho de contração (*encoder*), responsável pela agregação de contexto, e um caminho de expansão (*decoder*), que permite a recuperação da resolução espacial. Conexões de atalho entre níveis simétricos transferem mapas de características do *encoder* para o *decoder* por concatenação, preservando detalhes de borda e melhorando a localização de estruturas pequenas, aspecto particularmente relevante em aplicações pulmonares. Variantes bidimensionais e tridimensionais são amplamente utilizadas, assim como extensões que incorporam blocos residuais ou mecanismos de atenção. Em cenários com número limitado de anotações, a U-Net tende a apresentar treinamento estável quando combinada a funções de perda baseadas em sobreposição, como o coeficiente de *Dice*, favorecendo o aprendizado de regiões pouco representadas.

2.1.5 Ensembles e métricas de avaliação de desempenho

Em machine learning, ensembles consistem na combinação sistemática das predições de múltiplos modelos com o objetivo de reduzir variância, mitigar erros específicos e aumentar robustez e estabilidade preditiva. A premissa central é que modelos distintos — seja por arquitetura, inicialização, amostragem dos dados ou estratégia de treinamento — capturam padrões complementares do espaço de características²⁹.

Além de melhorar a robustez e a acurácia, os *ensembles* oferecem um mecanismo para estimar a incerteza preditiva das decisões do modelo. A incerteza pode ser decomposta em duas componentes principais: incerteza aleatória, que é inerente à variabilidade e ao ruído nos dados, e incerteza epistêmica, que reflete a incerteza nos parâmetros do modelo e pode ser reduzida com mais dados de treinamento. Em um *deep ensemble*, a variância entre as predições dos diferentes modelos do comitê pode ser utilizada como uma aproximação da incerteza epistêmica. Uma alta discordância entre os modelos para uma determinada amostra sinaliza uma baixa confiança na predição, o que é uma informação valiosa em um contexto clínico, podendo ser usada para sinalizar casos que necessitam de maior atenção do especialista.

Formalmente, seja $f_i(x)$ função preditiva do modelo i , com $i = 1, \dots, \mathcal{M}$. O *ensemble* constrói uma função agregada, conforme a equação 1:

$$\mathcal{F}(x) = \mathcal{A}(f_1(x), f_2(x), \dots, f_{\mathcal{M}}(x)) \quad (1)$$

Onde $\mathcal{A}(\cdot)$ representa o operador de agregação.

Em aplicações médicas, ensembles são particularmente eficazes em cenários com:

- Alta variabilidade interobservador;
- Pequeno tamanho amostral;
- Desbalanceamento de classes;
- Ruído inerente às imagens.

A Figura 10 ilustra conceitualmente esse processo, mostrando que a agregação das predições individuais tende a produzir estimativas mais estáveis do que modelos isolados.

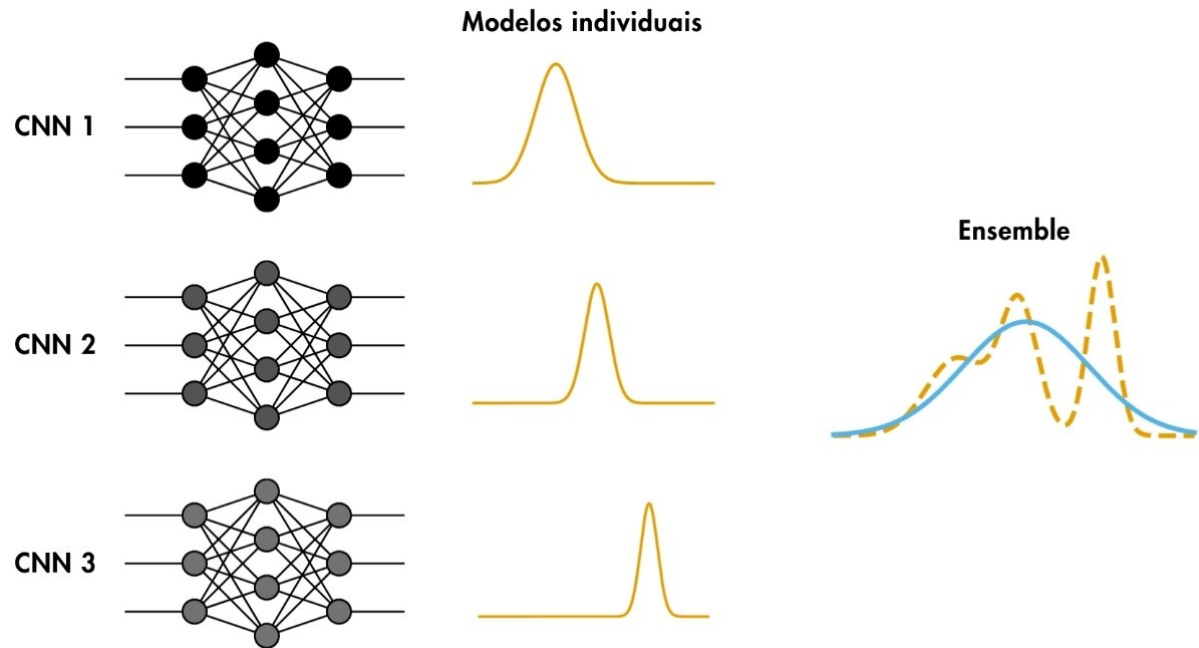


Figura 10. Esquema conceitual de um ensemble por Soft Voting de redes neurais convolucionais. Fonte: Elaborado pelo autor.

2.1.5.1 Estratégias de combinação

Diferentes estratégias podem ser empregadas, dependendo do grau de complexidade desejado.

i. Média aritmética (*Soft Voting*)

Para problemas probabilísticos, a predição final é obtida através da equação 2:

$$\hat{p}(x) = \frac{1}{M} \sum_{i=1}^M f_i(x) \quad (2)$$

Onde $f_i(x)$ representa a probabilidade estimada pelo modelo i . Essa abordagem reduz a variância e mantém estabilidade estatística.

ii. Votação majoritária (*Hard Voting*)

A classe final é determinada pela equação 3:

$$\hat{\mathcal{Y}}(x) = \text{moda}\{y_1(x), y_2(x), \dots, y_M(x)\} \quad (3)$$

Onde $\mathcal{Y}_i(x)$ é a classe predita pelo modelo i .

iii. Stacking

No stacking, um meta-modelo $\mathcal{g}(\cdot)$ Aprende a melhor combinação das saídas dos modelos base conforme a equação 4:

$$\hat{\mathcal{Y}}(x) = \mathcal{g}(f_1(x), f_2(x), \dots, f_M(x)) \quad (4)$$

Essa estratégia aumenta a flexibilidade, porém eleva custo computacional e risco de sobreajuste.

Neste estudo, adotou-se a estratégia de combinação direta das probabilidades preditas por média aritmética (*soft voting*), em virtude de:

- Simplicidade computacional;
- Estabilidade estatística;
- Redução consistente de variância;
- Boa relação desempenho–custo.

Essa escolha é coerente com o objetivo de construir um *deep ensemble* robusto para classificação de nódulos pulmonares.

2.1.5.2 Métricas de avaliação de desempenho

O desempenho dos classificadores foi avaliado por métricas amplamente empregadas em aprendizado supervisionado, particularmente relevantes em cenários clínicos com desbalanceamento de classes.

Sejam:

- TP: verdadeiros positivos;
- TN: verdadeiros negativos;
- FP: falsos positivos;
- FN: falsos negativos.

As métricas utilizadas são definidas conforme as equações abaixo:

$$\text{Acurácia}^{30} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{Precisão}^{31} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall}^{31} = \frac{TP}{TP+FN} \quad (7)$$

$$\text{F1-score}^{31} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (8)$$

O F1-score é particularmente relevante em problemas com desbalanceamento entre classes, pois sintetiza em uma única métrica o compromisso entre precisão e sensibilidade, penalizando modelos que apresentam desempenho desigual entre essas dimensões.

2.2 MATERIAIS E MÉTODOS

A Figura 11 apresenta o fluxograma geral do *pipeline* metodológico adotado neste trabalho, abrangendo as etapas de aquisição das bases de dados, análise exploratória, pré-processamento, treinamento e avaliação dos modelos, e serve como

referência para a descrição detalhada das etapas metodológicas apresentadas a seguir.

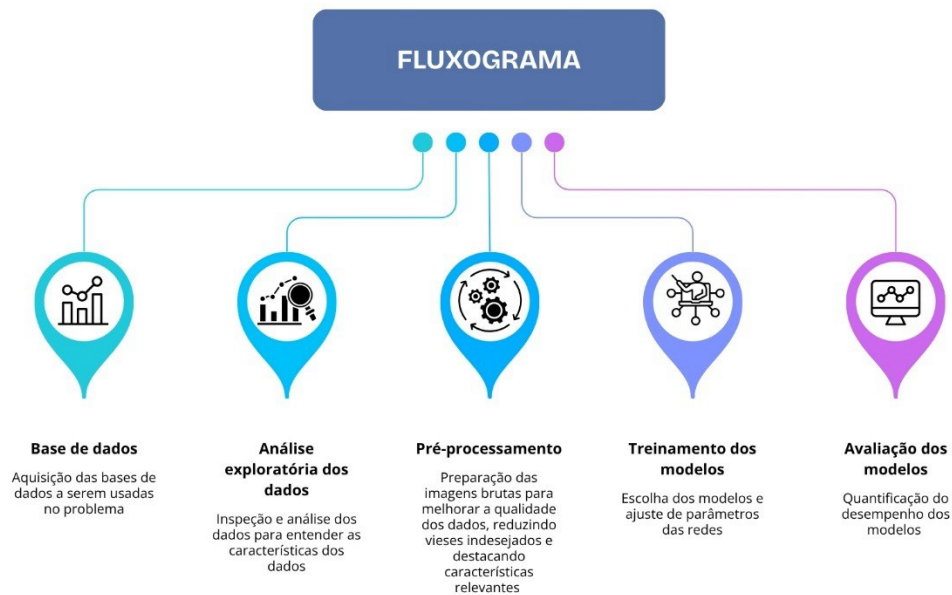
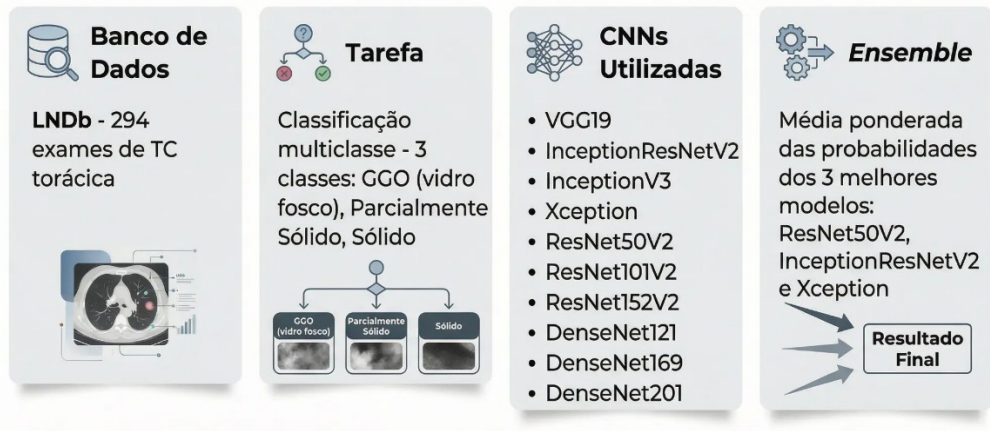


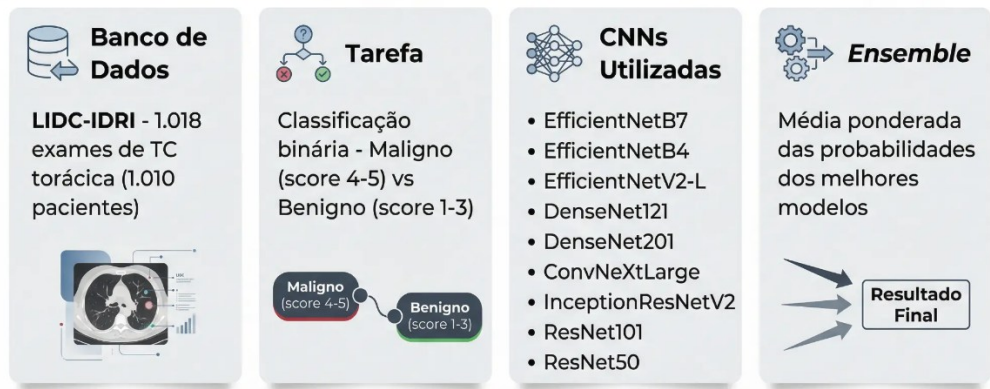
Figura 11. Fluxograma metodológico geral adotado. Fonte: Elaborado pelo autor.

A Figura 12 sintetiza, de forma esquemática, os três experimentos realizados neste estudo, destacando, em cada cenário, a base de dados empregada, a tarefa de classificação, as arquiteturas avaliadas, a estratégia de *ensemble* adotada e as métricas de desempenho consideradas, oferecendo uma visão comparativa geral que antecede a apresentação dos resultados individuais.

EXPERIMENTO 1: CLASSIFICAÇÃO MULTICLASSE (LNDb)



EXPERIMENTO 2: CLASSIFICAÇÃO BINÁRIA (LIDC-IDRI)



EXPERIMENTO 3: FLUXO SEQUENCIAL (LIDC-IDRI)

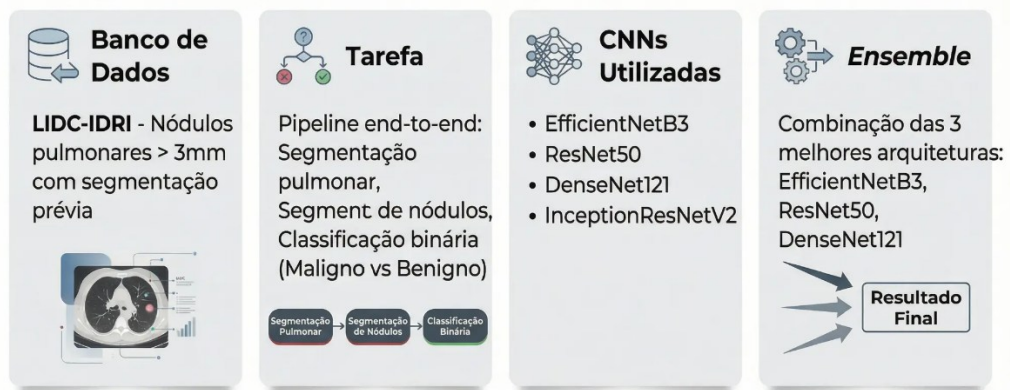


Figura 12. Síntese esquemática dos três experimentos realizados. Fonte: Elaborado pelo autor.

2.2.1 Bases de dados

Neste projeto foram utilizadas duas bases de dados públicas de tomografia computadorizada de tórax, previamente rotuladas e anonimizadas de acordo com padrões éticos internacionais. Para experimentos preliminares, foi empregada a base de dados *Lung Nodule Database* (LNDb), que contém 294 exames de tomografia computadorizada adquiridos no Centro Hospitalar e Universitário de São João, no Porto, Portugal³². Essa base foi utilizada principalmente para validação inicial do *pipeline* e comparação exploratória entre arquiteturas.

As bases de dados utilizadas disponibilizam informações clínicas limitadas. A base LNDb fornece dados sobre localização, tamanho e características morfológicas dos nódulos, mas não apresenta, de forma sistemática, informações detalhadas sobre idade, sexo ou subtipos histológicos. De modo semelhante, a base *Lung Image Database Consortium and Image Database Resource Initiative* (LIDC-IDRI) inclui anotações radiológicas realizadas por múltiplos especialistas, com *scores* subjetivos de malignidade e características morfológicas, porém não contempla, de forma padronizada, a confirmação histopatológica nem a distinção entre subtipos de câncer de pulmão, como carcinoma de pequenas células e não pequenas células³³. Dessa forma, a análise desenvolvida neste trabalho concentra-se em padrões radiológicos e desempenho preditivo, reconhecendo como limitação a ausência de correlação direta com dados clínico-patológicos.

A etapa principal do estudo foi conduzida utilizando a base (LIDC-IDRI), acessada por meio do portal *The Cancer Imaging Archive* (TCIA), a qual foi adotada como base de dados de referência neste trabalho. A LIDC-IDRI é composta por 1.018 exames de tomografia computadorizada provenientes de 1.010 pacientes, abrangendo uma ampla variedade de protocolos de aquisição, espessuras de corte e resoluções espaciais, o que reflete a heterogeneidade encontrada em cenários clínicos reais.

Um diferencial da LIDC-IDRI é o processo de anotação estruturado em duas fases, realizado por até quatro médicos radiologistas especialistas. Na primeira fase, cada radiologista realiza anotações de forma independente, identificando e delimitando nódulos pulmonares em diferentes níveis de confiança. Na segunda fase, os radiologistas revisam anonimamente as anotações dos demais especialistas, podendo manter ou ajustar suas marcações iniciais. Esse procedimento resulta em

múltiplas segmentações e escalas de malignidade por nódulo, permitindo a análise da variabilidade interobservador e a construção de máscaras consensuais.

A análise exploratória de ambas as bases foi conduzida em ambiente computacional *Python*. Foram utilizadas as bibliotecas *NumPy* para operações numéricas, *pandas* para organização e manipulação de metadados, *Matplotlib* para visualização gráfica e *scikit-learn* para particionamento estratificado dos dados e cálculo de métricas de desempenho. Para a LIDC-IDRI, o pacote *pylidc* foi empregado para consulta estruturada aos exames, recuperação das anotações dos radiologistas, reconstrução volumétrica consistente e conversão das intensidades para unidades HU, assegurando padronização no pré-processamento dos dados.

2.2.2 Preparação e curadoria dos dados

A análise exploratória de dados foi conduzida de forma sistemática. Os exames apresentaram dimensões variáveis, com matriz típica de 512×512 *pixels* e número de cortes no eixo z (longitudinal) dependente do protocolo de aquisição e da extensão anatômica de cada indivíduo. A distribuição de nódulos por paciente mostrou-se heterogênea, e casos sem nódulos anotados foram identificados e mantidos como negativos verdadeiros, refletindo cenários clínicos reais.

Para garantir a independência entre os conjuntos de treinamento, validação e teste, e assim evitar o risco de *data leakage* (vazamento de dados), a divisão dos dados foi realizada no nível do paciente. Isso assegura que todos os cortes tomográficos de um mesmo indivíduo pertençam exclusivamente a um único conjunto, prevenindo que o modelo seja avaliado com dados correlacionados aos que utilizou no treinamento. Essa prática é fundamental para uma estimativa não enviesada do desempenho de generalização do modelo, conforme destacado na literatura sobre o tema.

A inspeção das anotações confirmou que cada nódulo poderia apresentar até quatro delineamentos independentes, além de uma escala de malignidade variando de 1 a 5. A Figura 13 apresenta a distribuição em histogramas das características radiológicas atribuídas pelos especialistas.

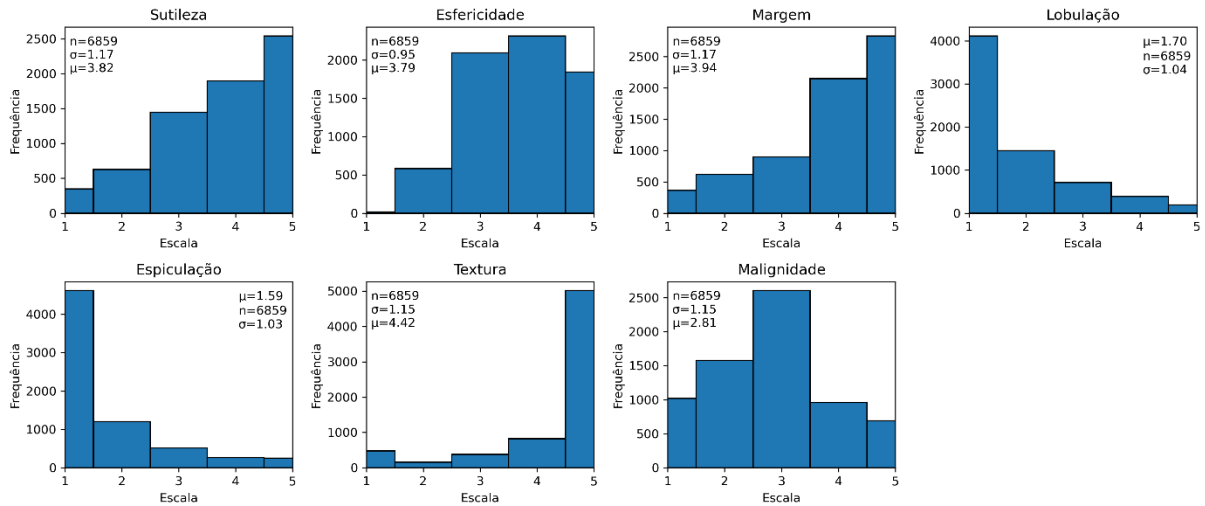


Figura 13. Distribuições de características anotadas pelos radiologistas na base de dados LIDC-IDRI. Fonte: Elaborado pelo autor.

Com o objetivo de ampliar a diversidade espacial dos dados de entrada e reduzir o risco de sobreajuste durante o treinamento dos modelos, foram exploradas técnicas de *data augmentation* aplicadas diretamente aos cortes de tomografia computadorizada³⁴. As transformações incluíram rotações, espelhamentos, escala (*zoom*) e deslocamentos, preservando a estrutura anatômica essencial dos nódulos³⁵. A Figura 14 ilustra exemplos dessas transformações aplicadas a uma mesma imagem, ilustrando como diferentes variações geométricas podem ser geradas a partir de um único corte original.

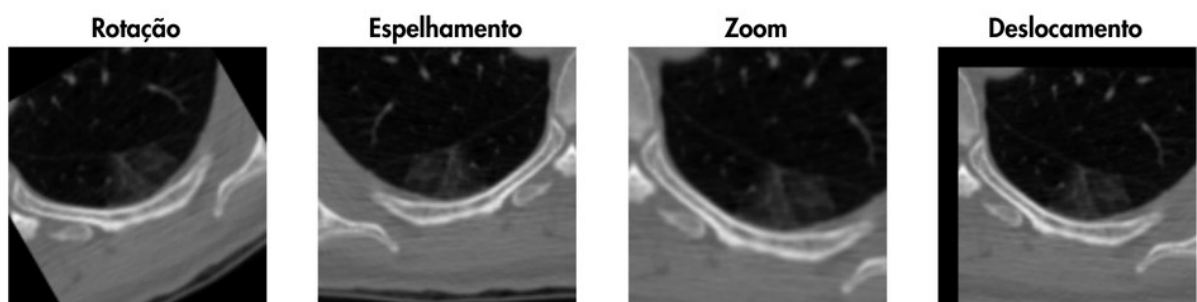


Figura 14. Exemplos de técnicas de *data augmentation* aplicadas a cortes de tomografia computadorizada de tórax. Fonte: Elaborado pelo autor.

A anotação e a rotulagem dos nódulos seguiram um consenso operacional. Máscaras binárias por nódulo foram geradas a partir dos contornos individuais realizados independentemente pelos radiologistas. Adotou-se um critério de votação majoritária com limiar de 50%, no qual nódulos delineados por pelo menos dois dos quatro observadores eram incluídos na máscara final³⁶. Esse procedimento replica

práticas consolidadas na literatura para uso da base LIDC-IDRI, em que volumes de interesse e máscaras consensuais são derivados por interseção majoritária. As etapas incluíram extração dos contornos, combinação por votação, remoção de máscaras muito pequenas (área inferior a oito *pixels*²) e inspeção de qualidade para exclusão de anotações inconsistentes. A Figura 15 ilustra a sobreposição das anotações individuais e a máscara consensual resultante.

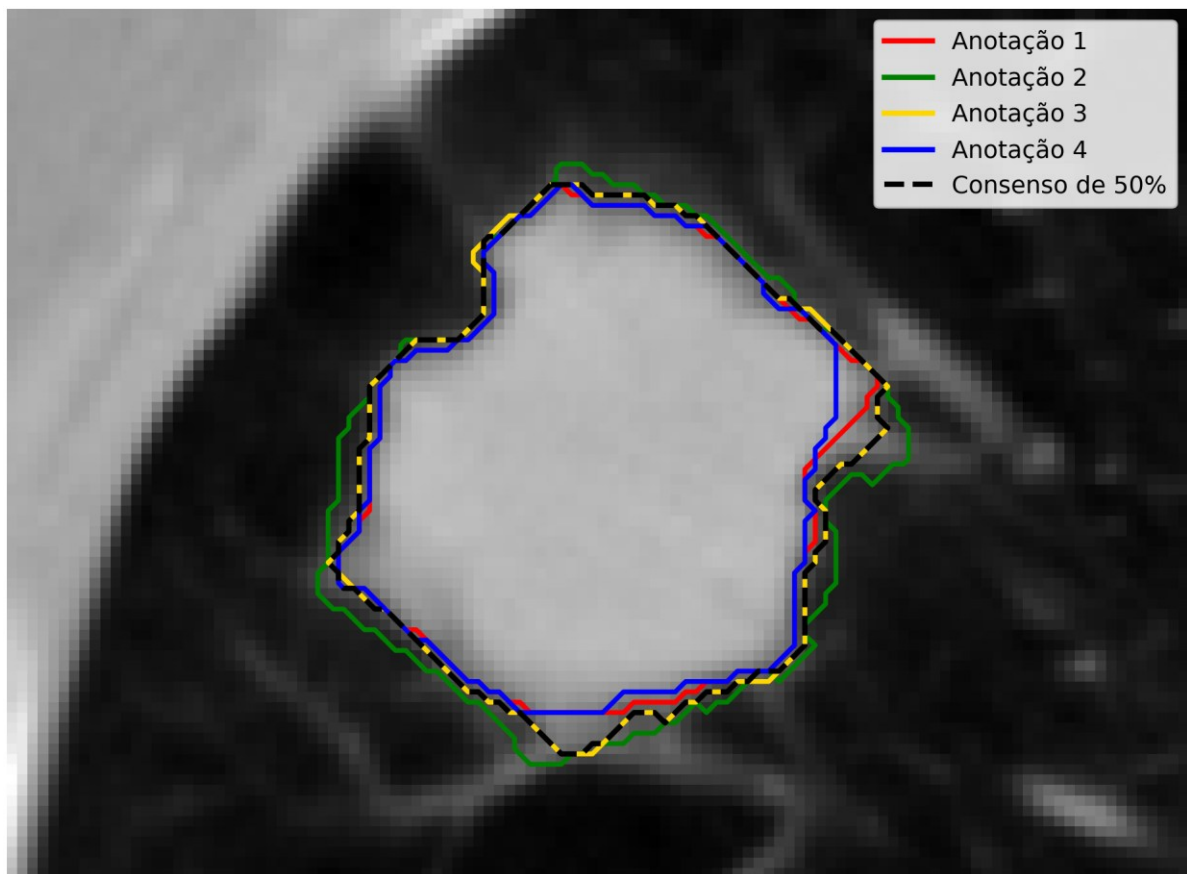


Figura 15. Sobreposição das quatro anotações de um mesmo nódulo e máscara consensual de 50%.
Fonte: Elaborado pelo autor.

Além disso, realizamos uma análise visual sistemática dos recortes, na qual observamos ampla variabilidade morfológica dos nódulos pulmonares, incluindo diferenças em lobulação, espiculação, esfericidade, margens, sutileza e textura interna. No esquema de anotação da base LIDC-IDRI, esses descritores semânticos são atribuídos por radiologistas em escalas ordinais e refletem aspectos distintos da aparência radiológica dos nódulos. A lobulação descreve a presença de saliências arredondadas ao longo do contorno, enquanto a espiculação quantifica a ocorrência de prolongamentos radiais em forma de espinho, característica frequentemente

associada a maior suspeição de malignidade. A esfericidade avalia o quão próxima a forma tridimensional do nódulo é de uma geometria esférica, e a margem descreve o grau de definição do contorno, variando de bordas mal definidas a nitidamente delimitadas. A sutileza está relacionada à conspicuidade do nódulo em relação ao parênquima pulmonar adjacente, refletindo a dificuldade de detecção visual, enquanto a textura interna caracteriza a composição do nódulo, abrangendo padrões sólidos, parcialmente sólidos e de vidro fosco, os quais impactam diretamente a percepção visual e a delimitação das lesões³⁷.

Exemplos dessas características são apresentados na Figura 16, evidenciando a heterogeneidade morfológica dos nódulos e a complexidade de sua caracterização radiológica. Essas variações estruturais, frequentemente associadas a diferentes níveis de suspeição clínica na base LIDC-IDRI, reforçam os desafios da análise visual e motivam o uso de métodos computacionais capazes de aprender representações discriminativas diretamente dos dados de imagem.

Para a tarefa de classificação, a binarização dos rótulos considerou como malignos ou de alta suspeição os nódulos com escala de malignidade igual a 4 ou 5, e como benignos ou indeterminados aqueles com escala entre 1 e 3. Essa definição segue prática amplamente adotada em estudos baseados na base LIDC-IDRI, nos quais escores elevados apresentam maior concordância interobservador e maior probabilidade clínica de malignidade, enquanto escores intermediários refletem maior incerteza diagnóstica. Essa estratégia prioriza alta suspeição clínica durante o treinamento, contribuindo para a redução de falsos positivos nas fases iniciais de ajuste dos modelos.

Para garantir reprodutibilidade, foram fixadas sementes aleatórias (*random seeds*) em todas as fontes de aleatoriedade (Python, NumPy, scikit-learn e framework de deep learning). A partição final do conjunto de dados manteve 80% dos pacientes para treinamento, 10% para validação e 10% para teste, configuração amplamente empregada em estudos de aprendizado profundo quando há número suficiente de casos. A divisão foi realizada em nível de paciente, com estratificação para preservação da proporção entre as classes, assegurando avaliação imparcial do desempenho final dos modelos.

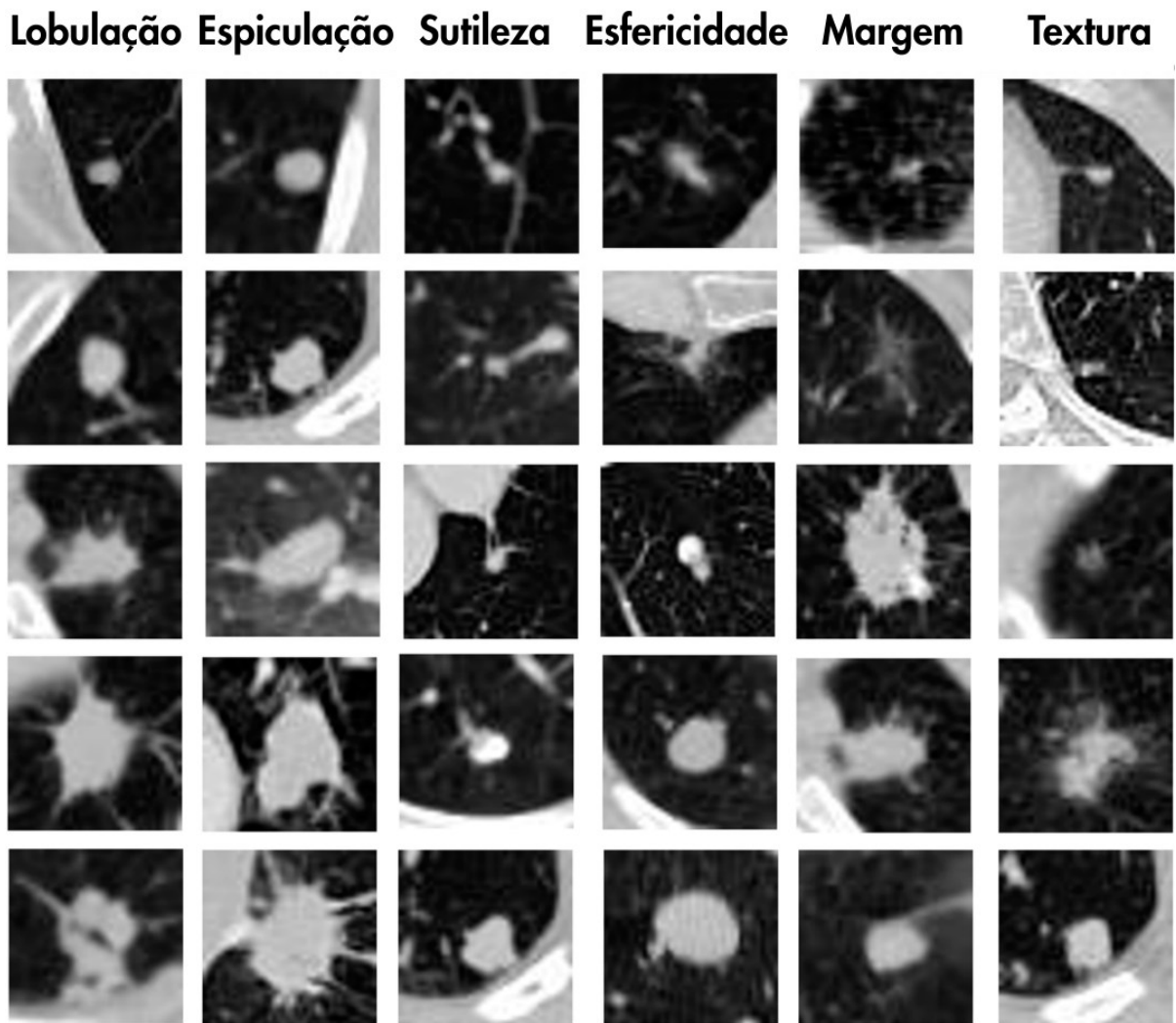


Figura 16. Exemplos da variabilidade morfológica dos nódulos pulmonares presentes na base de dados LIDC-IDRI. Fonte: Elaborado pelo autor.

2.2.3 Pré-processamento e segmentação

Dois *pipelines* de pré-processamento foram empregados para atender a diferentes necessidades computacionais ao longo do desenvolvimento do estudo. O *pipeline* de alta resolução produziu imagens finais com 512×512 *pixels* por corte tomográfico e foi utilizado no treinamento dos modelos de maior capacidade, preservando detalhes anatômicos finos relevantes para a segmentação e a classificação, conforme ilustrado nos exemplos com sobreposição de máscaras. Em contraste, o *pipeline* de resolução reduzida produziu imagens com 16×16 *pixels* por corte e foi empregado exclusivamente em etapas de prototipagem e testes iniciais, com o objetivo de reduzir a carga computacional e viabilizar experimentações rápidas, mantendo apenas a morfologia essencial dos nódulos. A Figura 17 apresenta um

exemplo do *pipeline* em resolução reduzida ($16 \times 16 \text{ pixels}$), incluindo a imagem original, a máscara segmentada do nódulo e a sobreposição correspondente.

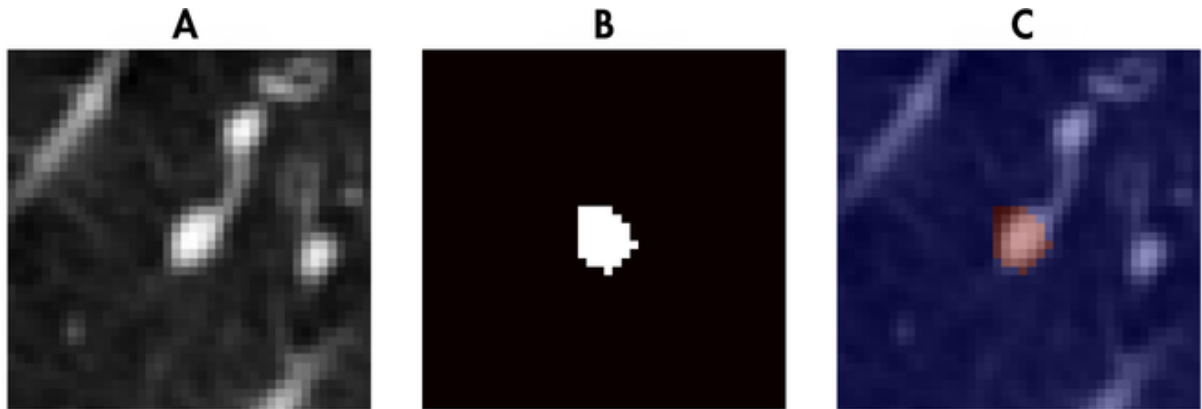


Figura 17. Exemplo do pipeline em resolução reduzida. (A) Imagem axial original, (B) máscara binária e (C) sobreposição. Fonte: Elaborado pelo autor.

A Figura 18 ilustra o *pipeline end-to-end* em alta resolução, com segmentação prévia do pulmão e, posteriormente, do nódulo, seguindo o mesmo padrão de representação adotado na figura anterior.

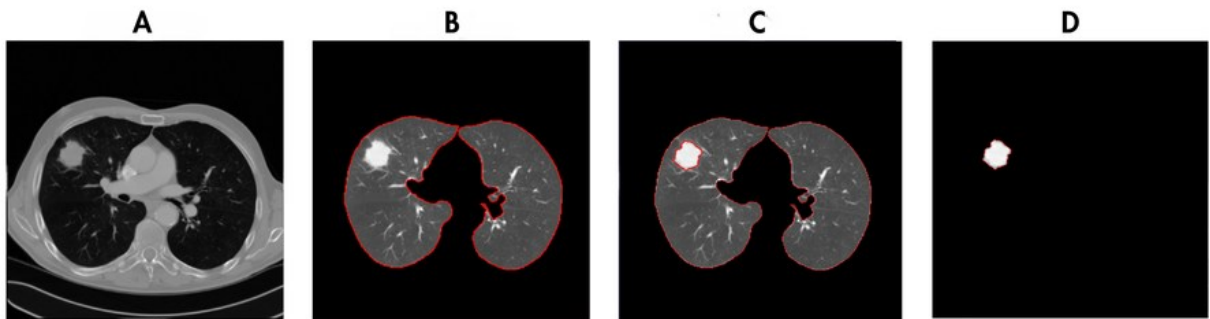


Figura 18. Exemplo do pipeline de segmentação aplicado a uma tomografia computadorizada de tórax. (A) Imagem axial original da tomografia computadorizada. (B) Segmentação com contorno dos pulmões sobrepostos. (C) Pulmões segmentados com sobreposição do nódulo pulmonar. (D) Máscara isolada do nódulo pulmonar correspondente. Fonte: Elaborado pelo autor.

A padronização geométrica foi realizada por meio de operações de preenchimento (*padding*) e redimensionamento, de modo a manter proporções espaciais consistentes entre os exames, independentemente das variações de protocolo de aquisição. As imagens resultantes foram convertidas em arranjos *NumPy* e armazenadas no formato *.npy*, permitindo carregamento eficiente e integração direta com os frameworks de *deep learning* utilizados no treinamento dos modelos.

A normalização de intensidade considerou as características específicas da tomografia computadorizada de tórax. As intensidades em unidades HU foram janeladas utilizando parâmetros apropriados para a avaliação pulmonar, adotando-se janela de pulmão com largura aproximada de 1.600 HU e nível em torno de -600 HU, configuração amplamente empregada para realçar o parênquima pulmonar e facilitar a visualização e a segmentação de nódulos. O fluxo incluiu o recorte de valores fora da janela de interesse e a conversão das intensidades para a escala de 0 a 255 quando necessário, de modo a garantir compatibilidade com modelos pré-treinados originalmente desenvolvidos para imagens em 8 *bits*.

Em cenários que exigiram estabilização adicional do treinamento, aplicou-se normalização *z-score* e, quando indicado, normalização *min-max* para conversão das intensidades para o intervalo [0,1]. A reamostragem foi utilizada para padronizar o espaçamento físico entre exames adquiridos com protocolos distintos. O processo considerou o espaçamento original entre *pixels* e entre cortes no eixo z, aplicando reamostragem isotrópica por interpolação trilinear nas imagens volumétricas. Para as máscaras de segmentação, empregou-se interpolação pelo método do vizinho mais próximo, de modo a preservar a natureza binária das máscaras e evitar suavização indevida das bordas.

A segmentação pulmonar foi considerada etapa fundamental para reduzir o campo de visão e excluir regiões de fundo não relevantes à análise. Quando necessário, especialmente em exames sem delimitação anatômica explícita, empregaram-se rotinas baseadas em limiarização no espaço de unidades HU, operações morfológicas e preenchimento de cavidades para obtenção de máscaras pulmonares coerentes. Na base LIDC-IDRI, a própria organização do conjunto de dados e as funções disponibilizadas pelo pacote *pylidc* facilitaram o foco nas regiões anatômicas de interesse, como o parênquima pulmonar, com exclusão de estruturas extra-pulmonares.

A segmentação dos nódulos pulmonares foi derivada dos contornos manuais disponibilizados na base LIDC-IDRI, os quais são fornecidos individualmente por até quatro radiologistas especialistas. Adicionalmente, aplicou-se um limiar mínimo de área para remoção de máscaras muito pequenas, e inspeções pontuais foram realizadas para descartar inconsistências. As máscaras resultantes foram utilizadas como padrão-ouro (*ground truth*) operacional tanto para o treinamento dos

modelos de segmentação quanto para a confirmação da localização dos nódulos durante a etapa de classificação.

2.2.4 Treinamento e *ensemble* de modelos

A eficácia de um *ensemble* não reside apenas no desempenho individual dos seus componentes, mas também na diversidade entre eles. Modelos diversos são aqueles que cometem erros em diferentes amostras, permitindo que a combinação de suas predições compense as falhas individuais. Neste trabalho, a diversidade foi promovida pela inclusão de arquiteturas de CNN com diferentes designs e profundidades (e.g., VGG, ResNet, DenseNet). Essas arquiteturas possuem vieses indutivos distintos, capturando diferentes níveis de abstração e padrões nas imagens. A seleção dos modelos para compor o *ensemble* final não foi aleatória; foram escolhidos os modelos com melhor desempenho individual no conjunto de validação, garantindo que apenas os "especialistas" mais competentes contribuíssem para a decisão final. A combinação de modelos de alto desempenho e com diversidade arquitetural visa criar um *ensemble* que seja, ao mesmo tempo, acurado e robusto.

O treinamento utilizou o otimizador Adam em todos os experimentos, em função de sua estabilidade e convergência consistente em tarefas de classificação de imagens médicas, especialmente em cenários de *fine-tuning* de modelos pré-treinados. Embora variantes como o AdamW incorporem estratégias de regularização mais explícitas por meio do desacoplamento do decaimento de peso, optou-se pelo Adam devido ao bom desempenho empírico observado nos experimentos preliminares e à presença de outras formas de regularização no treinamento, como parada antecipada (*early stopping*) e *data augmentation*^{21,22}. A função de perda adotada foi a *binary cross-entropy* (BCE) para classificação binária e a *categorical cross-entropy* (CCE) nos cenários multiclasse. Empregou-se parada antecipada, com monitoramento da menor perda de validação e salvamento do melhor ponto de treinamento. Os modelos foram inicializados com pesos pré-treinados no *ImageNet* e adaptados à tomografia computadorizada torácica por meio de *fine-tuning* específico^{38,39}.

Inicialmente, ajustou-se a entrada de imagens em canal único, expressas em unidades HU, para compatibilidade com as arquiteturas convolucionais, seja por meio da replicação do canal para três entradas, seja pela modificação do primeiro

bloco convolucional para aceitar um único canal. Em seguida, a cabeça de classificação foi reinicializada, e apenas essa porção do modelo foi treinada nas primeiras épocas, com o *backbone* congelado. Posteriormente, o *backbone* foi descongelado de forma progressiva, aplicando-se taxas de aprendizado diferenciadas, maiores na cabeça e nas camadas finais e menores nas camadas iniciais. As camadas de *batch normalization* tiveram suas estatísticas recalibradas com os dados médicos. A regularização incluiu *dropout* nas camadas finais e regularização L2 dos pesos. O cronograma de aprendizado iniciou-se com um aquecimento curto da taxa de aprendizado, partindo de zero e sendo aumentada de forma linear até o valor alvo nas primeiras épocas, com o objetivo de estabilizar os gradientes e favorecer a convergência inicial.

Em seguida, foi aplicada a estratégia de redução da taxa de aprendizado em platô (*ReduceLROnPlateau*). Sempre que a métrica de validação deixava de apresentar melhora por um número consecutivo de épocas, a taxa de aprendizado era reduzida por um fator fixo, até atingir um limite mínimo predefinido. A parada antecipada (*early stopping*) foi empregada para interromper o treinamento quando não houve melhora adicional da perda de validação por um número estipulado de épocas, utilizando como referência uma partição estratificada em nível de paciente, de modo a evitar sobreajuste e garantir avaliação imparcial do desempenho dos modelos.

2.2.5 Avaliação experimental

O estudo foi conduzido em três ciclos experimentais com progressão metodológica, conforme descrito anteriormente. Inicialmente, foi utilizada a base LNDb para demonstrar a viabilidade do *pipeline* proposto e realizar a calibração de hiperparâmetros. Em seguida, a base LIDC-IDRI foi adotada como conjunto de referência para validação em maior escala e em um cenário mais heterogêneo. As arquiteturas avaliadas incluíram modelos das famílias ResNet²⁶, DenseNet²⁷, InceptionResNet⁴⁰, Xception⁴¹, EfficientNet⁴², ConvNeXt⁴³ e VGG (Visual Geometry Group)⁴⁴. As predições foram combinadas por meio de estratégias de *ensemble* baseadas em média ponderada das probabilidades dos melhores modelos individuais, abordagem escolhida por sua robustez e por permitir estimativas indiretas de incerteza associadas às predições.

Os resultados quantitativos obtidos em cada fase experimental, bem como as comparações entre os diferentes cenários avaliados, são apresentados de forma detalhada na seção de Resultados.

2.3 RESULTADOS

2.3.1 Classificação multiclasse de nódulos pulmonares na base de dados LNDb (Experimento 1)

Inicialmente, avaliou-se o desempenho de diferentes arquiteturas de CNNs na tarefa de classificação multiclasse de nódulos pulmonares na base de dados LNDb. Esse experimento teve como objetivo estabelecer uma linha de base de desempenho em um cenário mais controlado, permitindo comparar modelos individuais e investigar o potencial de estratégias de *ensemble* antes da aplicação em contextos mais complexos.

As arquiteturas analisadas incluíram VGG19, InceptionResNetV2, InceptionV3, Xception, ResNet50V2, ResNet101V2, ResNet152V2, DenseNet121, DenseNet169 e DenseNet201.

O desempenho em termos de acurácia obtido por cada modelo individual é apresentado na Figura 19, permitindo a comparação direta entre as diferentes arquiteturas avaliadas.

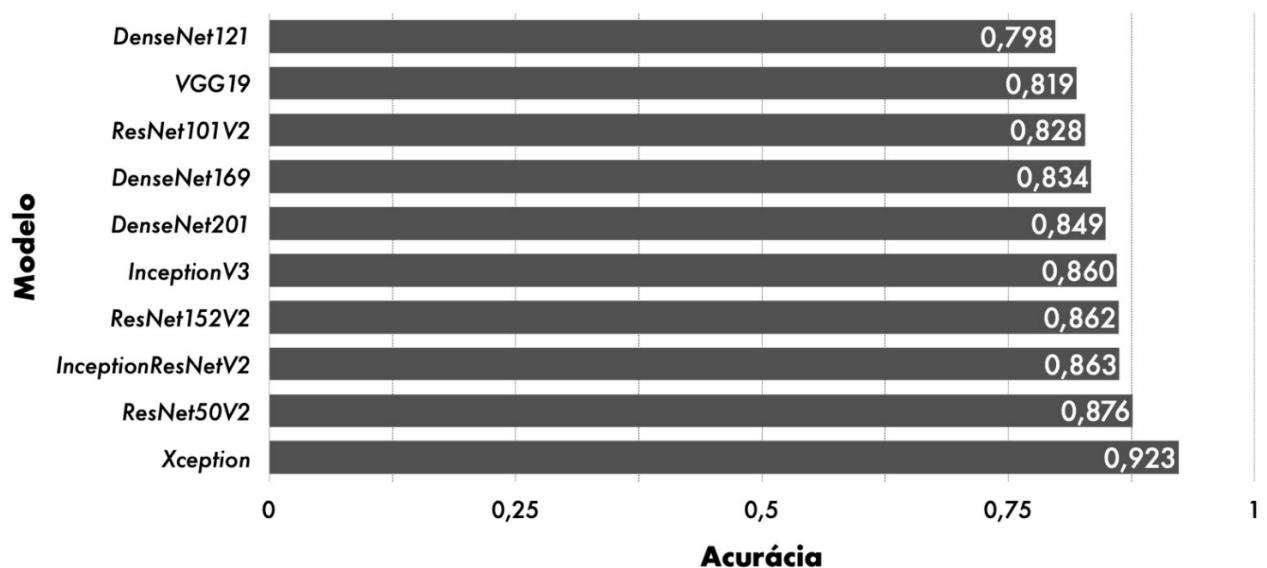


Figura 19. Desempenho em acurácia dos modelos individuais de CNN avaliados na classificação multiclasse de nódulos pulmonares na base LNDb. Fonte: Elaborado pelo autor.

Com base nesses resultados, foi construído um *ensemble* por média ponderada das probabilidades preditas por esses três modelos. O desempenho detalhado por classe, incluindo precisão, *recall* e *F1-score* dos modelos selecionados e do *ensemble* resultante, é apresentado na Tabela 2.

Tabela 2. Desempenho dos modelos individuais e do ensemble na classificação multiclasse de nódulos pulmonares na base de dados LNDb.

Modelo	Classe	Precisão	<i>Recall</i>	<i>F1-Score</i>
ResNet50V2	GGO	0,970	0,800	0,880
	Parcial	0,770	0,670	0,720
	Sólido	0,810	0,980	0,880
InceptionResNetV2	GGO	0,940	0,840	0,890
	Parcial	0,800	0,780	0,790
	Sólido	0,860	0,930	0,890
Xception	GGO	0,960	0,900	0,920
	Parcial	0,850	0,930	0,890
	Sólido	0,960	0,940	0,880
Ensemble	GGO	0,970	0,910	0,940
	Parcial	0,860	0,890	0,870
	Sólido	0,970	0,940	0,950

Fonte: Elaborado pelo autor.

2.3.2 Classificação binária de nódulos pulmonares na base de dados LIDC-IDRI (Experimento 2)

Na sequência, o *pipeline* foi aplicado a base de dados LIDC-IDRI em uma tarefa de classificação binária entre nódulos malignos e benignos. Esse experimento introduz maior complexidade ao problema, em razão da heterogeneidade anatômica,

da variabilidade de protocolos de aquisição e da incerteza associada às anotações radiológicas, permitindo avaliar a capacidade de generalização dos modelos e da estratégia de *ensemble*

Foram avaliadas nove arquiteturas de CNNs, a saber: EfficientNetB7, EfficientNetB4, EfficientNetV2-L, DenseNet121, DenseNet201, ConvNeXtLarge, InceptionResNetV2, ResNet101 e ResNet50.

A Figura 20 apresenta a comparação da acurácia obtida pelos principais modelos individuais e pelo *ensemble*.

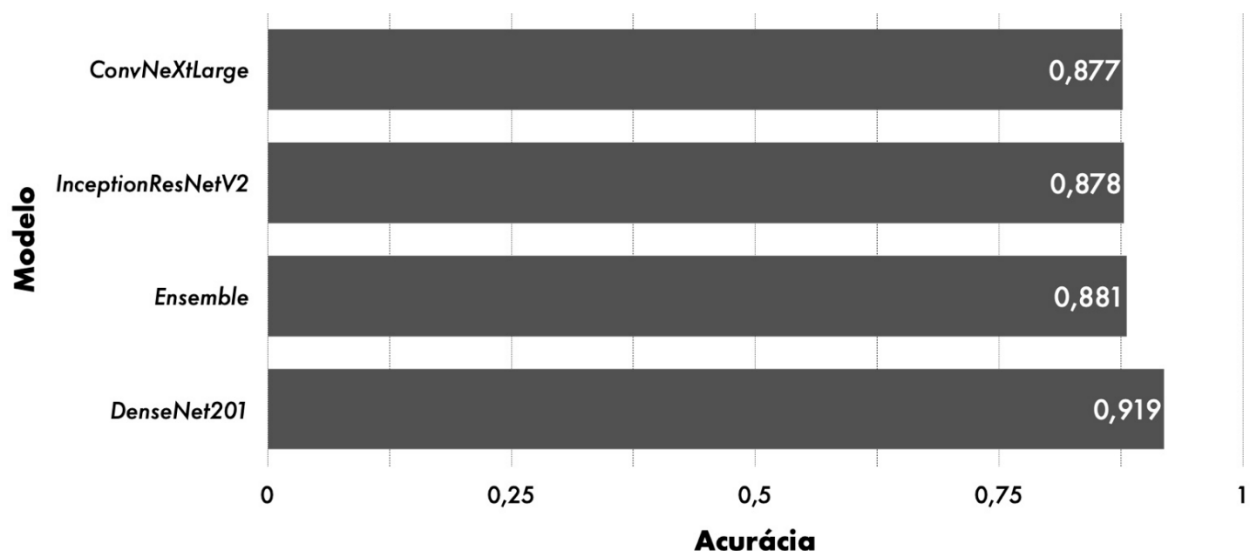


Figura 20. Acurácia dos principais modelos individuais e do ensemble na classificação binária de nódulos pulmonares na base de dados LIDC-IDRI. Fonte: Elaborado pelo autor.

Observa-se que o melhor desempenho individual foi alcançado pelo modelo DenseNet201, com acurácia de 0,919. O *ensemble*, construído a partir da combinação ponderada dos modelos mais performáticos, atingiu acurácia de 0,881, superando os resultados individuais de InceptionResNetV2 e ConvNeXtLarge, embora sem ultrapassar o melhor modelo isolado.

Esses resultados indicam que, neste cenário específico, a estratégia de *ensemble* contribuiu para maior estabilidade em relação a arquiteturas individuais intermediárias, mas não superou o desempenho máximo obtido por um único modelo altamente especializado.

2.3.3 Pipeline end-to-end com segmentação prévia na base de dados LIDC-IDRI (Experimento 3)

Por fim, foi avaliado um *pipeline end-to-end* que incorpora a segmentação pulmonar e nodular como etapa prévia à classificação, considerando apenas lesões com diâmetro superior a 3 mm. Esse experimento aproxima o fluxo de processamento de um cenário clínico realista, permitindo analisar o impacto da segmentação prévia na etapa de classificação e seus reflexos nas métricas de desempenho.

Quatro arquiteturas de CNNs foram avaliadas na tarefa de classificação binária entre nódulos malignos e benignos: EfficientNetB3, ResNet50, DenseNet121 e InceptionResNetV2. O *ensemble* foi construído a partir da combinação das três arquiteturas com melhor desempenho individual.

A Tabela 3 apresenta os resultados obtidos em termos de acurácia, precisão, *recall* e *F1-score*.

Tabela 3. Desempenho dos modelos no pipeline end-to-end com segmentação prévia de nódulos pulmonares na base de dados LIDC-IDRI.

Modelo	Acurácia	Precisão	Recall	F1-Score
EfficientNetB3	0,716	0,510	0,720	0,600
ResNet50	0,806	0,510	0,720	0,600
DenseNet121	0,826	0,760	0,760	0,760
InceptionResNetV2	0,824	0,770	0,780	0,770
Ensemble	0,799	0,780	0,790	0,760

Fonte: Elaborado pelo autor.

Essa redução de acurácia no *pipeline end-to-end* é compatível com efeitos de cascata entre etapas (segmentação → classificação) e com a restrição de contexto imposta pelas regiões de interesse (ROIs), especialmente em casos limítrofes.

2.3.4 Síntese comparativa e análise integrada dos experimentos

Os três experimentos realizados permitiram avaliar progressivamente o impacto do aumento da complexidade do problema, da mudança de base de dados e da introdução da segmentação prévia no desempenho dos modelos. Enquanto o Experimento 1 foi conduzido em um cenário multiclasse controlado na base de dados LNDb, os Experimentos 2 e 3 exploraram a base de dados LIDC-IDRI em tarefas binárias, sendo o último configurado como um *pipeline end-to-end* com segmentação pulmonar e de nódulos.

A Tabela 4 apresenta uma síntese comparativa dos três experimentos, destacando o melhor modelo individual e o desempenho do *ensemble* em cada cenário. Observa-se que o desempenho global é superior no Experimento 1 (LNDb), enquanto, nos experimentos conduzidos na base LIDC-IDRI, há redução mais evidente nas métricas agregadas, especialmente ao avançar para o *pipeline end-to-end* com segmentação prévia, refletindo maior variabilidade anatômica, heterogeneidade de protocolos e incerteza nas anotações.

Tabela 4. Comparação do melhor desempenho individual e do ensemble nos três experimentos

Experimento	Base de dados	Melhor modelo individual	Acurácia	Acurácia do <i>ensemble</i>
1	LNDb	Xception	0,920	0,940
2	LIDC-IDRI	DenseNet201	0,919	0,881
3	LIDC-IDRI	DenseNet121	0,826	0,799

Fonte: Elaborado pelo autor.

A Figura 21 compara as principais métricas entre LNDb (Experimento 1) e LIDC-IDRI (Experimento 2), contrastando o melhor modelo individual e o *ensemble* em cada base. Na base de dados LNDb, o *ensemble* iguala a acurácia do melhor modelo individual, porém apresenta redução em *F1-score* e *recall*. Na base LIDC-IDRI, observa-se queda de acurácia e *F1-score* com o *ensemble*, ao passo que há um

ganho discreto em *recall*, sugerindo aumento de sensibilidade à custa de desempenho global.

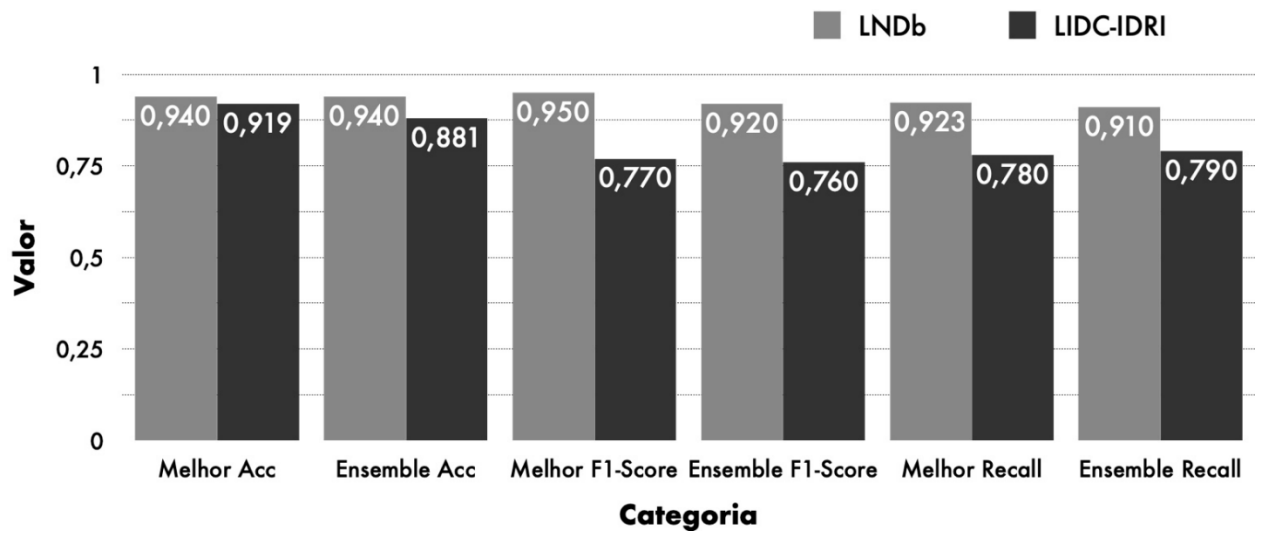


Figura 21. Comparação das métricas de desempenho entre os Experimentos 1 e 2. Fonte: Elaborado pelo autor.

A Figura 22 apresenta a diferença entre o desempenho do *ensemble* e o melhor modelo individual em cada métrica, considerando apenas os Experimentos 1 (LNDb) e 2 (LIDC-IDRI), definida como $\Delta = M_{ensemble} - M_{ref}$, em que M_{ref} corresponde ao maior valor obtido por um modelo individual naquela métrica. Valores positivos indicam melhoria proporcionada pelo *ensemble*, enquanto valores negativos indicam desempenho inferior ao melhor classificador isolado. Para acurácia na base LNDb, $\Delta = 0$ (empate entre *ensemble* e melhor modelo individual).

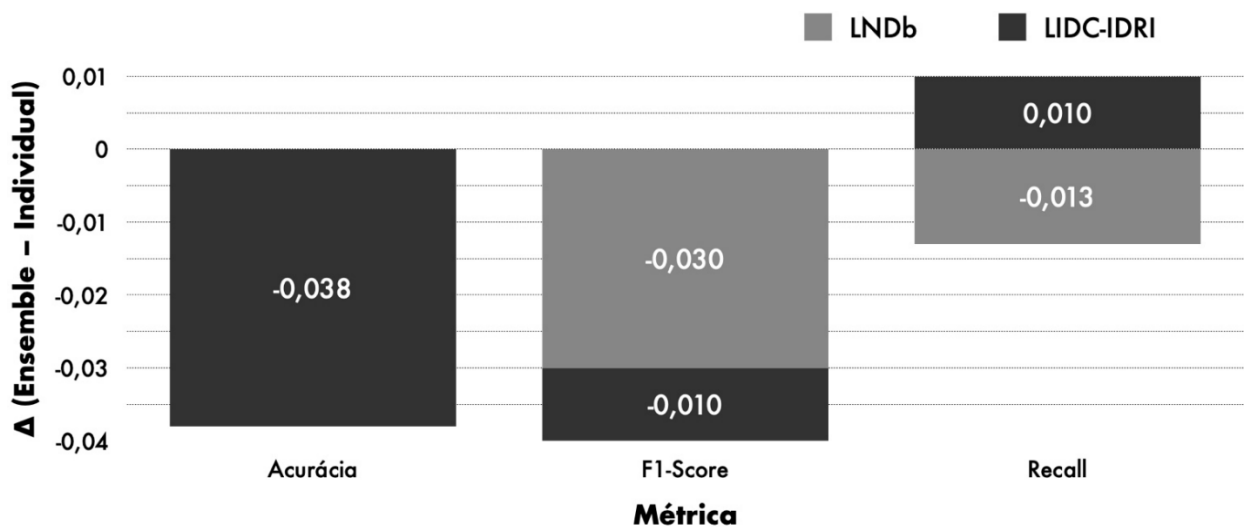


Figura 22. Diferença de desempenho entre o ensemble e o melhor modelo individual nas métricas avaliadas nos Experimentos 1 e 2. Fonte: Elaborado pelo autor.

Na base de dados LNDb, o *ensemble* empata em acurácia com o melhor modelo individual, mas apresenta redução em *F1-score* e *recall*. Na base de dados LIDC-IDRI, observa-se redução de acurácia e *F1-score* com o *ensemble*, ao passo que há ganho discreto em *recall*. Esses resultados indicam que, em cenários mais complexos e heterogêneos, como na base LIDC-IDRI, o *ensemble* pode aumentar a sensibilidade (*recall*) em detrimento do desempenho global, refletindo o compromisso entre a detecção de casos positivos e o desempenho global.

2.4 DISCUSSÃO

Os resultados obtidos ao longo dos três experimentos permitem uma análise integrada do desempenho de arquiteturas modernas de *deep learning* aplicadas à classificação de nódulos pulmonares em tomografia computadorizada, considerando diferentes níveis de complexidade do problema e heterogeneidade das bases de dados. A progressão experimental adotada — do cenário multiclasse na base de dados LNDb até um *pipeline end-to-end* com segmentação prévia na base de dados LIDC-IDRI — possibilitou avaliar não apenas o desempenho absoluto dos modelos, mas também sua estabilidade e capacidade de generalização.

Embora os modelos tenham apresentado bom desempenho em termos de acurácia, uma análise futura poderia investigar a calibração das probabilidades de saída. Um modelo bem calibrado é aquele em que a confiança em sua predição (e.g., uma probabilidade de 80%) corresponde à probabilidade real de acerto. Técnicas como a construção de curvas de calibração (ou diagramas de confiabilidade) e o uso de métodos de re-calibração, como a escala de Platt ou a regressão isotônica, poderiam ser empregadas para avaliar e, se necessário, corrigir as saídas do modelo, tornando-as mais interpretáveis e confiáveis para o uso clínico.

É importante ponderar que a combinação de múltiplos modelos, especialmente arquiteturas profundas, resulta em um *ensemble* com maior custo computacional tanto para treinamento quanto para inferência. Em um cenário de aplicação clínica em tempo real, o "peso" do *ensemble* pode ser um fator limitante. Futuros trabalhos poderiam explorar técnicas de poda (*pruning*) ou destilação de conhecimento (*knowledge distillation*) para transferir o desempenho do *ensemble* para um modelo único e mais leve, otimizando a eficiência sem sacrificar significativamente a acurácia. Adicionalmente, a seleção dos melhores modelos para o *ensemble*, aqui

baseada no desempenho de validação, poderia ser refinada com critérios estatísticos, como a análise da correlação dos erros entre os modelos, buscando maximizar a diversidade e, conseqüentemente, o ganho de desempenho do *ensemble*.

No Experimento 1, conduzido em um ambiente mais controlado, observou-se que diferentes arquiteturas apresentaram desempenhos individuais elevados e relativamente próximos, refletindo a capacidade das CNNs em capturar padrões morfológicos associados aos diferentes tipos de nódulos pulmonares. Esse comportamento é consistente com achados prévios em bases menos heterogêneas, nas quais o aprendizado supervisionado tende a convergir para soluções similares entre arquiteturas profundas consolidadas. Nesse contexto, a estratégia de *ensemble* contribuiu para reduzir a variância entre modelos e estabilizar as decisões, efeito amplamente descrito na literatura de aprendizado profundo aplicado à saúde.

Ao transpor o *pipeline* para a base de dados LIDC-IDRI no Experimento 2, em uma tarefa binária e com maior variabilidade anatômica e protocolar, verificou-se uma redução global do desempenho, tanto para modelos individuais quanto para o *ensemble*. Essa queda é compatível com características conhecidas dessa base, como a diversidade de protocolos de aquisição, a variabilidade populacional e a incerteza associada às anotações radiológicas. Nesses cenários, diferenças numéricas pequenas entre modelos — da ordem de poucos pontos percentuais — devem ser interpretadas com cautela, uma vez que podem não representar ganhos estatisticamente ou clinicamente relevantes, especialmente na ausência de validação estatística formal.

A análise da diferença relativa entre o desempenho do *ensemble* e do melhor modelo individual evidenciou que, em cenários mais complexos, a combinação de modelos nem sempre resulta em melhoria global das métricas. Ainda assim, observa-se um comportamento consistente de preservação ou aumento do *recall*, mesmo quando há redução em acurácia ou *F1-score*. Esse compromisso entre sensibilidade e especificidade é particularmente relevante em aplicações clínicas, nas quais a priorização da detecção de casos potencialmente malignos pode ser desejável, sobretudo em contextos de triagem.

O Experimento 3 introduziu um *pipeline end-to-end* com segmentação pulmonar e nodular prévia, restringindo a análise a nódulos com diâmetro superior a 3 mm. Neste experimento, os dados de entrada foram as sequências completas dos pacientes, ao invés dos nódulos já identificados como no experimento 1 e 2. Isto torna

o problema mais próximo da prática clínica, e dificulta a tarefa em si, pois demanda uma correta segmentação prévia ao processo de classificação, ocasionando uma redução global do desempenho dos modelos. Esse comportamento sugere que a segmentação prévia, apesar de conceitualmente vantajosa por restringir o campo de visão, introduz fontes adicionais de erro que podem se propagar para a etapa de classificação. Ainda assim, o equilíbrio observado entre precisão e *recall* no *ensemble* reforça o potencial dessa estratégia para aplicações de triagem e priorização visual.

Quando a comparação é restrita a estudos que utilizam a base LIDC-IDRI, observa-se que estratégias baseadas em *ensembles* e representações multiescala tendem a favorecer métricas sensíveis à detecção. Xu et al.⁴⁵ propuseram o método MSCS-DeepLN, baseado em um *ensemble* de CNNs 3D multiescala com função de perda *cost-sensitive*, reportando acurácia superior a 90%, com ganhos em *recall* e *F1-score* em relação a modelos individuais. De forma complementar, Xiao et al.⁴⁶ utilizaram um *ensemble* híbrido combinando CNNs profundas e atributos morfológicos, alcançando acurácia de 93,1%, precisão de 83,9%, *recall* de 81,8% e *F1-score* de 82,8% na base LIDC-IDRI.

Resultados mais elevados têm sido reportados em pipelines altamente integrados. Dhayalini e Revathi⁴⁷ alcançaram acurácia de 99,4%, precisão de 99,0%, *recall* de 99,2% e *F1-score* de 99,1% na base LIDC-IDRI ao empregar segmentação dedicada e classificação com *Multiscale Adaptive Swin Transformer*, embora tais abordagens dependam de segmentações altamente estáveis. Para comparação externa, Pedrosa et al.³² reportaram, na base de dados LNDb, acurácia em torno de 94%, com *F1-score* e *recall* superiores aos observados em cenários equivalentes na base LIDC-IDRI, reforçando o impacto da heterogeneidade entre bases. Em conjunto, esses resultados indicam que a redução de desempenho observada no Experimento 3 é compatível com a literatura quando consideradas a propagação de erro da segmentação e a heterogeneidade da base LIDC-IDRI, enquanto o *ensemble* mantém um equilíbrio favorável entre precisão e sensibilidade.

As escolhas arquiteturais realizadas neste estudo refletem tanto o estado da arte no período de desenvolvimento quanto limitações práticas de custo computacional. Arquiteturas como ResNet, DenseNet, InceptionResNet e ConvNeXt representam famílias consolidadas, com diferentes mecanismos de conectividade e profundidade efetiva, permitindo avaliar a complementaridade entre modelos sem recorrer a *ensembles* excessivamente grandes. Embora estudos recentes explorem

combinações com maior número de arquiteturas, não há consenso na literatura quanto ao número ideal de modelos em *ensembles*, e ganhos adicionais tendem a apresentar retornos decrescentes.

Do ponto de vista clínico, os resultados devem ser interpretados à luz do objetivo principal do estudo: apoiar a tomada de decisão e não substituir o especialista. Em ambientes com alto volume de exames e recursos humanos limitados, sistemas capazes de priorizar regiões suspeitas ou maximizar a sensibilidade podem contribuir para reduzir a carga cognitiva do radiologista e diminuir a probabilidade de falhas diagnósticas. Mesmo em cenários nos quais o *ensemble* não supera o melhor modelo individual em todas as métricas, ganhos em *recall* podem representar benefícios relevantes na prática clínica.

Por fim, destaca-se que a ausência de validação estatística formal entre diferenças pequenas de desempenho e a limitação de informações clínico-patológicas nas bases utilizadas constituem restrições importantes deste trabalho. Estratégias futuras podem incluir validação cruzada interbases, análise sistemática de intervalos de confiança para todas as métricas e integração com dados clínicos adicionais, de modo a fundamentar escolhas mais robustas de modelos e estratégias de *ensemble*.

3. CONCLUSÃO

Este trabalho avaliou o uso de arquiteturas modernas de *deep learning*, individuais e em *ensemble*, para a classificação de nódulos pulmonares em tomografia computadorizada, com ênfase na maximização do *recall* como critério clinicamente relevante. Ao longo de diferentes cenários experimentais, observou-se que, embora modelos individuais possam apresentar desempenho competitivo, estratégias de *ensemble* tendem a produzir previsões mais estáveis e a preservar a sensibilidade, especialmente em contextos de maior complexidade e heterogeneidade dos dados.

A análise integrada dos experimentos evidenciou que o aumento da complexidade do *pipeline*, particularmente com a introdução da segmentação prévia, impacta negativamente as métricas globais, em parte devido à propagação de erros entre estágios. Ainda assim, o comportamento do *ensemble* mostrou-se consistente com o objetivo de priorização de casos suspeitos, característica desejável em aplicações de triagem.

Os resultados reforçam o potencial de sistemas computacionais como ferramentas de apoio à decisão clínica, capazes de auxiliar o radiologista na priorização visual de exames, sem substituir o julgamento especializado. Como perspectivas futuras, destacam-se a validação cruzada entre bases independentes, a integração de informações clínico-demográficas e a incorporação de análises estatísticas formais das métricas, visando aumentar a robustez e a aplicabilidade clínica das abordagens propostas.

REFERÊNCIAS

1. International Agency for Research on Cancer (IARC). Global Cancer Observatory (GCO) – Cancer Today. GLOBOCAN 2022. Lyon: IARC; 2024. (Accessed 2024 Feb 8; Version 1.1).
2. Santos MO, Lima FCS, Martins LFL, Oliveira JFP, Almeida LM, Cancela MC. Estimativa de incidência de câncer no Brasil, 2023-2025. *Rev Bras Cancerol*. 2023;69(1). doi:10.32635/2176-9745.RBC.2023v69n1.3700.
3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al; National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395–409.
4. American College of Radiology (ACR). Lung CT Screening Reporting & Data System (Lung-RADS). Reston (VA): American College of Radiology; 2022.
5. Gould MK, Tang T, Liu ILA, Lee J, Zheng C, Danforth KN, et al. Recent trends in the identification of incidental pulmonary nodules. *Am J Respir Crit Care Med*. 2015;192(10):1208–14.
6. van Riel SJ, Sánchez CI, Bankier AA, Naidich DP, Verschakelen J, Scholten ET, et al. Observer variability for Lung-RADS categories in CT lung cancer screening: impact on patient management. *Radiology*. 2019;292(2):421–9.
7. Wormanns D, Beyer F, Diederich S. Diagnostic performance of CAD for pulmonary nodules in CT: systematic aspects and clinical value as a second reader. *Eur Radiol*. 2004;14(11):2106–15.
8. Li Q. Computer-aided diagnosis of pulmonary nodules in CT: a survey and future prospects. *Expert Syst Appl*. 2014;41(5):177–88.
9. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.
10. Kalender WA. *Computed Tomography: Fundamentals, System Technology, Image Quality, Applications*. 3rd ed. Erlangen: Publicis; 2011.
11. Hounsfield GN. Computerized transverse axial scanning (tomography): Part 1. Description of system. *Br J Radiol*. 1973;46(552):1016–22.
12. Seeram E. *Computed Tomography: Physical Principles, Clinical Applications, and Quality Control*. 4th ed. St. Louis: Elsevier; 2016.

13. McCollough CH, Bruesewitz MR, Kofler JM Jr. CT dose reduction and dose management tools: overview of available options. *Radiographics*. 2006;26(2):503–12.
14. Fleischner Society. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology*. 2017;284(1):228–43.
15. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. *Radiology*. 2008;246(3):697–722.
16. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society: International multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol*. 2011;6(2):244–85.
17. McNitt-Gray MF. Lung nodules and cancer screening: radiologist variability and imaging protocol effects. *Radiology*. 2019;292(2):430–1.
18. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441–6.
19. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge (MA): MIT Press; 2016.
20. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *DLMIA*. 2017.
21. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*. 2014.
22. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: *International Conference on Learning Representations (ICLR)*. 2019.
23. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *ICML*. 2015.
24. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*. 2014;15:1929–58.

25. Aresta G, Jacobs C, Araújo T, et al. Cross-dataset evaluation of pulmonary nodule analysis in CT: effects of heterogeneity and protocol. *Med Image Anal.* 2020;
26. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *CVPR*. 2016. doi:10.1109/CVPR.2016.90.
27. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: *CVPR*. 2017. doi:10.1109/CVPR.2017.243.
28. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *MICCAI*. 2015. doi:10.1007/978-3-319-24574-4_28.
29. Dietterich TG. Ensemble methods in machine learning. In: *Multiple Classifier Systems*. 2000. p. 1–15.
30. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen.* 1936;7(2):179–88.
31. van Rijsbergen, C.J. *Information Retrieval*. 2nd ed. London: Butterworths; 1979.
32. Pedrosa J, Aresta G, Ferreira C, Araujo T, Costa P, Pereira S, et al. LNDb: A Lung Nodule Database on Computed Tomography. arXiv:1911.08434. 2019.
33. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys.* 2011;38(2):915–31.
34. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *J Big Data.* 2019;6:60.
35. Perez L, Wang J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv:1712.04621. 2017.
36. Reeves AP, Xie Y, Jirapatnakul AC, et al. LIDC-IDRI nodule segmentation consistency and consensus strategies. *Med Phys.* 2012;39(10):6171–82.
37. Opulencia P, Channin DS, Raicu DS, Furst JD. Mapping LIDC, RadLex™, and lung nodule image features. *J Digit Imaging.* 2011;24(2):256–270. doi:10.1007/s10278-010-9285-6.
38. Prechelt L. Early stopping — but when? In: *Neural Networks: Tricks of the Trade*. Springer; 1998.
39. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: *CVPR*. 2009.
40. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261. 2016.

41. Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
42. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). 2019. arXiv:1905.11946.
43. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. doi:10.1109/CVPR52688.2022.01198.
44. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Proceedings of the International Conference on Learning Representations (ICLR). 2015. arXiv:1409.1556.
45. Xu X, Wang C, Guo J, Yi Z. MSCS-DeepLN: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. *Med Image Anal.* 2020;65:101772. doi:10.1016/j.media.2020.101772.
46. Xiao N, Ding J, Liu X, et al. Ensemble classification for predicting the malignancy level of pulmonary nodules on chest computed tomography images. *PLoS One.* 2020;15(6):e0234388. doi:10.1371/journal.pone.0234388.
47. Dhayalini M, Revathi alias Ponmozhi B. Multi-phase deep learning framework with Multiscale Adaptive Swin Transformer and embedding attention for precision lung nodule detection and classification. *Sci Rep.* 2026;16:1652. doi:10.1038/s41598-025-31147-2.

DADOS CURRICULARES

IDENTIFICAÇÃO	
Nome	Nycolas Mariotto
Data de Nascimento	10/03/2000
Nacionalidade	Brasileira
Nome em citações bibliográficas:	Mariotto, Nycolas Mariotto, N.
Currículo Lattes	http://lattes.cnpq.br/9283079499512657
ORCID	https://orcid.org/0009-0000-3586-2020
outro identificador	https://br.linkedin.com/in/nycolas-mariotto-314b87220
FORMAÇÃO ACADÊMICA	
686/2023	Bacharel em Física Médica. (2018 - 2023) - Universidade Estadual Paulista "Júlio de Mesquita Filho".
PRODUÇÃO BIBLIOGRÁFICA	
Artigos completos publicados em periódicos: Mariotto, N., Andrade dos Santos, G., Ramos Fioschi, G., Cristini Marata Higino, A., Fernando Pereira Paseti, L., Omia Bueno Pereira, T., ... Felipe Fattori Alves, A. (2025). Desenvolvimento de uma Plataforma Automatizada de Segmentação de Imagens para Ensaios de Migração Celular Baseada em Inteligência Artificial. Revista Brasileira De Física Médica, 19, 853. https://doi.org/10.29384/rbfm.2025.v19.19849001853	
Resumos expandidos publicados em anais de congressos: MARIOTTO, N.; GONCALVES, F. H. B. ; ALVES, A. F. F. . DEEP ENSEMBLE MODEL FOR SEGMENTATION AND CLASSIFICATION OF LUNG NODULES IN COMPUTED TOMOGRAPHY SCANS. In: 5th European Congress of Medical Physics, 2024, Munich. European Congress of Medical Physics, 2024. v. 125. https://doi.org/10.1016/j.ejimp.2024.103601	
Resumos publicados em anais de congressos: MARIOTTO, N.; ALVES, A. F. F. ; PEREIRA, T. O. B. ; SANDRIM, V. C. . SEGMENTATION OF CELL SCRATCH ASSAY EXPERIMENTS USING ARTIFICIAL INTELLIGENCE. In: 3rd European Congress on Biomedical Science and Engineering & Human Genetics, 2025, Vienna. Biomedical Science and Engineering & Human Genetics, 2025.	

RAMOS FIOSCHI, GEOVANA ; MARIOTTO, N. ; ANDRADE DOS SANTOS, GEOVANA ; CRISTINI MARATA HIGINO, ANDRIA ; ALVES, A. F. F. . Avaliação de Técnicas de Inteligência Artificial para Detecção e Segmentação em Scratch Assay de Células Endoteliais. In: 29º Congresso Brasileiro de Física Médica, 2025, São Paulo. Anais 29º Congresso Brasileiro de Física Médica, 2025.

CRISTINI MARATA HIGINO, ANDRIA ; MARIOTTO, N. ; ANDRADE DOS SANTOS, GEOVANA ; RAMOS FIOSCHI, GEOVANA ; ALVES, A. F. F. . Detecção de nanopartícula em células no roboflow. In: 29º Congresso Brasileiro de Física Médica, 2025, São Paulo. Anais 29º Congresso Brasileiro de Física Médica, 2025.

GONCALVES, L. G. ; MARIOTTO, N. ; ALVES, A. F. F. . Métodos de visão computacional para classificação de radiografias de pacientes com e sem tuberculose. In: 29º Congresso Brasileiro de Física Médica, 2025, São Paulo. Anais 29º Congresso Brasileiro de Física Médica, 2025.

MARIOTTO, N.; ALVES, A. F. F. . Deep Ensemble Model for Lung Nodule Classification in Computed Tomography Scans. In: XXIX CONGRESSO BRASILEIRO DE ENGENHARIA BIOMÉDICA, 2024, Ribeirão Preto. INTEGRANDO CIÊNCIA, TECNOLOGIA E INOVAÇÃO para uma prática clínica, acessível e sustentável, 2024.

PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS

Trabalhos apresentados pelo autor:

MARIOTTO, N.; ALVES, A. F. F. . Deep ensemble para segmentação e classificação de nódulos pulmonares em tomografia computadorizada. 2025. (Apresentação de Trabalho/Simpósio).

MARIOTTO, NYCOLAS; ANDRADE DOS SANTOS, GEOVANA ; RAMOS FIOSCHI, GEOVANA ; CRISTINI MARATA HIGINO, ANDRIA ; OMIA BUENO PEREIRA, THAINÁ ; PASSETI, L. F. P. ; MORELLI ZAMPIERI, GABRIELA ; SANDRIM, V. C. ; ALVES, A. F. F. . Desenvolvimento de uma Plataforma Automatizada de Segmentação de Imagens para Ensaios de Migração Celular Baseada em Inteligência Artificial. 2025. (Apresentação de Trabalho/Congresso).

Trabalhos apresentados por coautores com participação do autor

ALVES, A. F. F. ; MARIOTTO, NYCOLAS ; PEREIRA, T. O. B. ; SANDRIM, V. C. . SEGMENTATION OF CELL SCRATCH ASSAY EXPERIMENTS USING ARTIFICIAL INTELLIGENCE. 2025. (Apresentação de Trabalho/Congresso).

CRISTINI MARATA HIGINO, ANDRIA ; MARIOTTO, NYCOLAS ; ANDRADE DOS SANTOS, GEOVANA ; RAMOS FIOSCHI, GEOVANA ; ALVES, A. F. F. . Detecção de nanopartícula em células no roboflow. 2025. (Apresentação de Trabalho/Congresso).

RAMOS FIOSCHI, GEOVANA ; MARIOTTO, NYCOLAS ; ANDRADE DOS SANTOS, GEOVANA ; CRISTINI MARATA HIGINO, ANDRIA ; ALVES, A. F. F. . Avaliação de Técnicas de Inteligência Artificial para Detecção e Segmentação em Scratch Assay de Células Endoteliais. 2025. (Apresentação de Trabalho/Congresso).

GONCALVES, L. G. ; MARIOTTO, NYCOLAS ; ALVES, A. F. F. . Métodos de visão computacional para classificação de radiografias de pacientes com e sem tuberculose. 2025. (Apresentação de Trabalho/Congresso).

RAMOS FIOSCHI, GEOVANA ; MARIOTTO, N. ; ANDRADE DOS SANTOS, GEOVANA ; CRISTINI MARATA HIGINO, ANDRIA ; ALVES, A. F. F. . Detecção de Processos de Internalização de Nanopartículas em Culturas de Células Tumoriais Através de Métodos Computacionais. 2025. (Apresentação de Trabalho/Congresso).

ANDRADE, J. R. A. ; MARIOTTO, N. ; FERREIRA, J. G. ; ALVES, A. F. F. . Classificação de nódulos pulmonares em exames de tomografia computadorizada tridimensional através de redes neurais convolucionais. 2024. (Apresentação de Trabalho/Congresso).

FERREIRA, J. G. ; ANDRADE, J. R. A. ; MARIOTTO, N. ; ALVES, A. F. F. . CARACTERIZAÇÃO DE NÓDULOS PULMONARES UTILIZANDO TÉCNICAS DE SEGMENTAÇÃO PARA REDES NEURAI CONVOLUCIONAIS. 2024. (Apresentação de Trabalho/Congresso).

ALVES, A. F. F. ; MARIOTTO, N. ; GONCALVES, F. H. B. . Deep ensemble model for segmentation and classification of lung nodules in computed tomography scans. 2024. (Apresentação de Trabalho/Congresso).

GONCALVES, F. H. B. ; ALMEIDA, C. B. ; PICELLI, L. ; MARIOTTO, N. ; ALVES, A. F. F. . Radiômica na classificação de texturas de nódulos pulmonares: avaliação de desempenho diagnóstico. 2023. (Apresentação de Trabalho/Congresso).

Formação complementar:

Formação complementar robusta em Machine Learning, Deep Learning (Keras e PyTorch) e Python para Data Science, totalizando mais de 180 horas de capacitação avançada, incluindo visão computacional, NLP, embeddings semânticos, redes convolucionais e recorrentes, e análise de dados aplicada à saúde.