

Universidade Federal de Alagoas
Instituto de Computação



Dissertação de Mestrado

**Modelo para Classificação de Nódulos Pulmonares
Pequenos usando Descritores *Radiomics***

Ailton Felix de Lima Filho
aflf@ic.ufal.com

Orientador:

Prof. Dr. Marcelo Costa Oliveira

Co-orientador:

Prof. Dr. Aydano Pamponet Machado

Ailton Felix de Lima Filho

Modelo para Classificação de Nódulos Pulmonares Pequenos usando Descritores *Radiomics*

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Prof. Dr. Marcelo Costa Oliveira

Co-orientador:

Prof. Dr. Aydano Pamponet Machado

Catálogo na fonte
Universidade Federal de Alagoas
Biblioteca Central

Bibliotecária Responsável: Helena Cristina Pimentel do Vale

- L732m Lima Filho, Ailton Felix de.
Modelo para classificação de nódulos pulmonares pequenos usando
descritores radiomics / Ailton Felix de Lima Filho. – 2017.
84 f.: il.
- Orientador: Marcelo Costa Oliveira.
Coorientador: Aydano Pamponet Machado.
Dissertação (mestrado em Informática) – Universidade Federal de Alagoas.
Instituto de Computação. Programa de Pós-Graduação em Informática. Maceió,
2017.
- Bibliografia: f. 54-60.
Apêndice: f. 61-84.
1. Tecnologia da Informação. 2. Informática médica. 3. Diagnóstico auxiliado
por computador. 4. Nódulo pulmonar. 5. Câncer de pulmão. 6. Diagnóstico
por imagem. 7. Radiomics. I. Título.

CDU: 004.891.3:61

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação

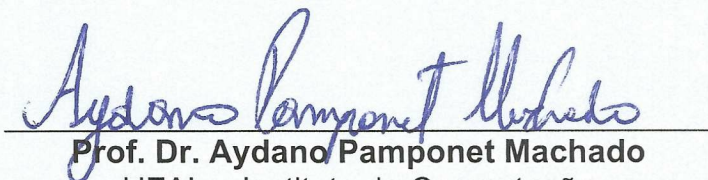
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Membros da Comissão Julgadora da Dissertação de Mestrado de Ailton Felix de Lima Filho, intitulada: “Modelo para Classificação de Nódulos Pulmonares Pequenos usando Descritores Radiomics”, apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 31 de julho de 2017, às 09h00min, na Sala de Videoconferência do CEPETEC, no Instituto de Computação da UFAL.

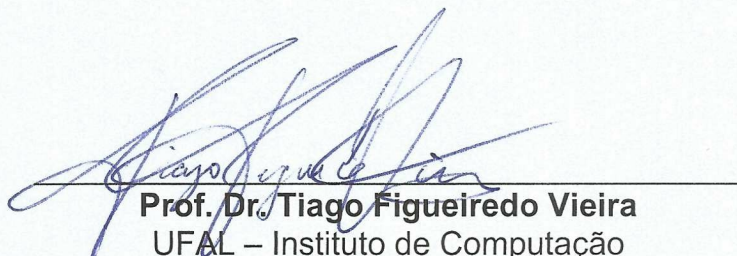
COMISSÃO JULGADORA



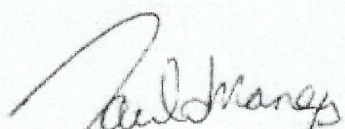
Prof. Dr. Marcelo Costa Oliveira
UFAL – Instituto de Computação
Orientador



Prof. Dr. Aydano Pamponet Machado
UFAL – Instituto de Computação
Coorientador



Prof. Dr. Tiago Figueiredo Vieira
UFAL – Instituto de Computação
Examinador



Prof. Dr. Paulo Mazzoncini de Azevedo Marques
USP – Universidade de São Paulo
Examinador

Agradecimentos

Primeiramente, quero agradecer ao senhor nosso Deus, que por mais dificuldades que eu tenha encontrado durante minha formação, com a força que Ele me forneceu, consegui realizar mais essa grande conquista na minha vida.

Agradeço a minha família, meus irmãos, Allan Cristian e Arnold de Almeida por estarem sempre do meu lado, e principalmente aos meus pais, Ailton Felix e Ana Cristina, por todo o apoio, incentivo e dedicação a fim de me proporcionar uma educação de qualidade. Sem eles este momento não teria chegado até aqui, e este momento não seria possível de acontecer. Agradeço também a minha namorada e companheira, Amanda Regina, por todo o seu apoio e compreensão nos momentos mais difíceis, sempre me fazendo acreditar em meu potencial.

Quero agradecer também aos meus colegas de laboratório, Lucas Lins e José Raniery, colegas com os quais convivi durante minha formação acadêmica e com quem compartilhei mutuamente conhecimentos e experiências. Agradeço a eles por serem bastante prestativos, me ajudando sempre que possível a entender certos problemas envolvidos no cenário deste trabalho. A todos os colegas que cursaram as disciplinas da grade curricular e que ingressaram no mestrado comigo.

O meu muito obrigado ao meu orientador, o professor Dr. Marcelo Oliveira, por ter ajudado em meu trabalho sempre que precisei, um orientador acessível e dedicado a seu papel como orientador. Agradeço a ele por tantas outras coisas fundamentais na minha formação. Agraço também ao meu co-orientador, o professor Dr. Aydano Machado e ao Professor Dr. Tiago Vieira, pelo auxílio durante meus momentos de dúvidas na implementação deste trabalho. Quero agradecer também a todos os professores do Programa de Pós-Graduação em Informática (PPGI) do Instituto de Computação (IC), por toda atenção e auxílio durante minha formação.

Agradeço também ao Dr. Marcel Koenigkam por o seu tempo dedicado a avaliação dos resultados, bem como por sua significativa sugestão de escopo do nosso trabalho, ajudando em muito a resolver os problemas encontrados no projeto.

Agradeço a todo o pessoal da secretaria do IC, em especial o secretário Marcelo e à secretária Floripes (Flor), por toda ajuda e disponibilidade no tratamento dos assuntos burocráticos do programa, sempre me atendendo com extrema boa vontade.

Por fim, quero agradecer a todos que contribuíram, de maneira arbitrária ou não, significativamente para minha formação. A todos o meu muito obrigado.

Resumo

O câncer de pulmão é uma doença caracterizada como crescimento anormal de células que invadem e destroem tecidos vizinhos, sendo responsável por muitas mortes ao redor do mundo. Um diagnóstico precoce da doença, geralmente realizado com base em informações qualitativas e semi-quantitativas extraídas de imagens de Tomografia Computadorizada (TC), traz maiores chances de cura e opções de tratamento para o paciente, porém, devido aos desafios no processo de interpretação de imagens médicas, principalmente no tocante a nódulos pulmonares pequenos (<10mm), o diagnóstico se torna clinicamente difícil, tornando a decisão clínica complexa. Devido à variabilidade e complexidade do diagnóstico de nódulos pulmonares pequenos, ferramentas de Auxílio ao Diagnóstico por Computador (CAD) baseadas em atributos de imagem, providenciam ajuda ao radiologista visando alcançar uma melhor acurácia da classificação de nódulo (provável maligno ou benigno), por agir como uma segunda opinião ao especialista. O uso de atributos *radiomics* permite um diagnóstico quantitativo mais objetivo se comparado às abordagens qualitativas ou semi-quantitativas mais comumente utilizadas na avaliação do câncer, diminuindo significativamente o problema da variabilidade no diagnóstico. Porém, ainda existe uma necessidade de descobrir conteúdos/atributos relevantes a fim de melhorar o desempenho de sistemas CAD. O objetivo deste trabalho foi desenvolver um modelo para classificação de nódulos pulmonares pequenos usando atributos *radiomics* extraídos da região de microambiente do nódulo. Foi avaliado também o teste de hipótese de que considerar a região de parênquima no entorno do nódulo, permite uma melhora de desempenho na classificação de nódulos pulmonares pequenos. O modelo para classificação desenvolvido obteve como melhor resultado uma área sob a curva ROC (AUC) média de 0.875 ± 0.048 com o algoritmo Perceptron de Múltiplas Camadas (MLP) com uma validação cruzada 10-*fold* na classificação de 214 nódulos pulmonares com diâmetros entre 5 e 10mm. Os resultados mostraram a relevância da utilização de atributos *radiomics* para classificação de nódulos pulmonares pequenos. A utilização da região do parênquima pulmonar melhorou o desempenho do modelo, comprovando o teste de hipótese. A classificação de nódulos pulmonares é uma área desafiadora mesmo para os médicos especialistas, devido à natureza complexa desse tipo de achado radiológico, porém, crítica para sobrevivência dos pacientes diagnosticados com câncer. Logo, avanços nessa área são de extrema importância.

Palavras-chaves: Câncer de Pulmão; Nódulo Pulmonar Pequeno; Auxílio ao Diagnóstico por Computador; *Radiomics*; Atributo de Imagem; Classificação.

Abstract

Lung cancer is a disease characterized as abnormal cells growth that invade and destroy neighboring tissues, accounting for many deaths around the world. An early diagnosis, usually performed based on qualitative information extracted from CT images, brings greater chances of cure and treatment options for the patient, however, due to the challenges in the medical image interpretation process, mainly for small pulmonary nodules (<10mm), the diagnosis becomes clinically difficult, making the clinical decision complex. Due to the variability and complexity of the diagnosis of small pulmonary nodules, Computer-Aided Diagnosis (CAD) tools based on image features, provides assistance to the radiologist in order to achieve a better accuracy of nodule classification (probable malignant or benign), by acting as a second opinion to the specialist. The use of radiomics features allows a quantitative diagnosis when compared to the recent qualitative strategies of cancer evaluation, significantly reducing the problem of variability in diagnosis. However, discovering the relevant content/features is still a necessity in order to improve the CAD systems performances. The aim of this study was to develop a classification model for small pulmonary nodules using radiomics features extracted from the nodule microenvironment. It was also evaluated the hypotheses test that considering the parenchyma region around the nodule allows an improvement in the small pulmonary nodules classification. The developed classification model obtained the best Area Under the ROC curve (AUC) of 0.875 ± 0.048 with the Multilayer Perceptron (MLP) algorithm with a 10-fold cross-validation in the classification of 214 pulmonary nodules with diameters between 5 and 10mm. The results showed the relevance of radiomics features for the classification of small pulmonary nodules. The use of the pulmonary parenchyma region improved the model performance, proving the hypothesis test. The nodules classification is a challenging area for specialists due to the natural complexity of diagnosis lesions, however, critical for patient survival diagnosed with cancer. Therefore, advances in this area are extremely important.

Keywords: Lung Cancer; Small Pulmonary Nodule; Computer-Aided Diagnosis; *Radiomics*; Image Features; Classification.

Lista de Figuras

1.1	Exemplos de nódulos pulmonares	3
2.1	Representação gráfica de um aparelho de TC tradicional	7
2.2	Representação de um espaço <i>voxel</i>	7
2.3	Aparência dos nódulos em uma imagem de TC quanto a intensidade	8
2.4	Ilustração do processo <i>radiomics</i> no auxílio a decisão	9
2.5	Diagrama de extração de um vetor de atributos	10
2.6	Imagem ROI de uma TC do pulmão com uma lesão em nível de cinza	11
2.7	Exemplos de tipos de forma dos nódulos	11
2.8	Exemplos de texturas	12
2.9	Ilustração do cálculo da matriz de coocorrência	13
2.10	Linha ortogonal traçada sobre a borda do nódulo pulmonar	15
2.11	Imagens geradas a partir da análise de nitidez de borda	16
2.12	Exemplo de mutação	18
2.13	Exemplo de cruzamento de um ponto	18
2.14	Hierarquia do aprendizado	19
2.15	Ilustração das etapas do algoritmo k-NN	20
2.16	Neurônio biológico simplificado	21
2.17	Ilustração de um Neurônio Artificial	21
2.18	Exemplo de RNA multicamadas	22
2.19	Ilustração de uma AD	24
2.20	Ilustração de uma AD com atributo contínuo	25
2.21	Espaço ROC	26
2.22	Ilustração de uma validação cruzada <i>3-fold</i>	27
3.1	Esquema geral da metodologia aplicada nesse trabalho	28
3.2	Ilustração do processo de cálculo do diâmetro do nódulo	30
3.3	Ilustração do processo de segmentação do parênquima pulmonar	31
4.1	Avaliação do resultado de segmentação pelo especialista	38
4.2	Resultados obtidos pela segmentação da região do parênquima	39
4.3	Resultados obtidos pela segmentação da região do parênquima	39
4.4	Resultados obtidos pela segmentação da região do parênquima	40
4.5	Resultados obtidos pela segmentação da região do parênquima	40
4.6	Classificação de nódulos pequenos usando o k-NN	42
4.7	Classificação de nódulos pequenos usando o k-NN com atributos selecionados pelo algoritmo genético	43
4.8	Classificação de nódulos pequenos usando o RF	44
4.9	Classificação de nódulos pequenos usando o MLP	45

4.10 Comparação da curva ROC entre os melhores resultados do modelo de classificação	46
--	----

Lista de Tabelas

2.1	Ilustração de uma Matrix de Confusão.	25
3.1	Número de nódulos sólidos entre 5-10mm que serão usados do BNP.	30
3.2	Atributos usados e suas respectivas regiões de extração	32
3.3	Composição de cada conjunto de atributos abordado neste trabalho	36
4.1	Classificação de nódulos pequenos usando o k-NN	41
4.2	Classificação de nódulos pequenos usando o k-NN com atributos selecionados pelo algoritmo genético	41
4.3	Classificação de nódulos pequenos usando o RF	43
4.4	Classificação de nódulos pequenos usando o MLP	44
4.5	Classificação de nódulos pequenos usando o MLP com atributos selecionados pelo algoritmo genético.	45
4.6	Atributos selecionados do conjunto I com o AGE	47

Lista de Abreviaturas

TC	Tomografia Computadorizada
CAD	do inglês <i>Computer-Aided Diagnosis</i>
k-NN	do inglês <i>k-Nearest Neighbor</i>
MLP	do inglês <i>Multilayer Perceptron</i>
RNA	Rede Neural Artificial
AD	Árvore de Decisão
RF	do inglês <i>Random Forest</i>
SVM	do inglês <i>Support Vector Machine</i>
ATI	Atributos de Intensidade
AF	Atributos de Forma
AT	Atributos de Textura
ANB	Atributos de Nitidez de Borda
AGE	Algoritmo Genético Evolutivo
TVP	Taxa de Verdadeiros Positivos
TFP	Taxa de Falsos Positivos
ROC	do inglês <i>Receiving Operating Characteristics</i>
AUC	do inglês <i>Area Under the ROC Curve</i>
LIDC	do inglês <i>Lung Image Database Consortium</i>

BNP	Banco de Nódulos Pulmonares
ROI	do inglês <i>Region of Interest</i>
T+NB	Atributos Textura + Nitidez de Borda
N	Atributos Nódulo
P	Atributos Parênquima
I	Atributos Integrados

Sumário

1	Introdução	1
1.1	Objetivo	5
1.2	Estrutura do Trabalho	5
2	Fundamentação Teórica	6
2.1	Nódulos Pulmonares em Imagens de Tomografia Computadorizada	6
2.2	Intensidade do Nódulo	7
2.3	Diagnóstico Auxiliado por Computador	8
2.4	<i>Radiomics</i>	9
2.5	Descritores de Imagem	10
2.5.1	Atributos de Intensidade	10
2.5.2	Atributos de Forma	11
2.5.3	Atributos de Textura	12
2.5.4	Atributos de Nitidez de Borda	14
2.6	Redução de Dimensionalidade	17
2.6.1	Algoritmo Genético Evolutivo	17
2.7	Aprendizado de Máquina	19
2.7.1	<i>k-Nearest Neighbor</i>	19
2.7.2	Rede Neural Artificial	20
2.7.3	Árvore de Decisão	23
2.7.4	Avaliação de desempenho	25
3	Materiais e Métodos	28
3.1	Banco de Nódulos Pulmonares	29
3.2	Seleção de Nódulos Pequenos	30
3.3	Segmentação	31
3.4	Extração de Atributos	32
3.4.1	Análise de Intensidade 3D	32
3.4.2	Análise de Forma 3D	34
3.4.3	Análise de Textura 3D	35
3.4.4	Análise de Nitidez de Borda 3D	35
3.5	Seleção de Atributos	35
3.6	Classificação	36
4	Resultados e Discussão	38
4.1	Resultados da Segmentação	38
4.2	Desempenho do Modelo para Classificação	41
4.3	Discussão	48

5 Conclusão	51
5.1 Trabalhos Futuros	52
5.2 Contribuições Científicas	52
Referências	60
Apêndice A Documento Modelo para Validação da Segmentação	61

Capítulo 1

Introdução

O câncer é caracterizado como um crescimento desordenado de células que invadem e destroem os tecidos e órgãos, podendo espalhar-se de forma muito rápida para outras regiões do corpo através da corrente sanguínea e do sistema linfático (JEMAL et al., 2014).

Segundo a Organização Mundial da Saúde (OMS), é inquestionável que o câncer é um problema de saúde pública, especialmente entre os países em desenvolvimento. A estimativa mundial, realizada em 2012, pelo projeto Globocan/Iarc, apontou que, dos 14 milhões de casos novos estimados, mais de 60% ocorreram em países em desenvolvimento. Nestes países a mortalidade por câncer de pulmão é outro agravante, pois dos 8 milhões de óbitos previstos em 2012, 70% ocorreram nesses mesmos países (INCA, 2016).

No Brasil as doenças crônicas não transmissíveis representam um importante ônus para a sociedade, sobretudo as neoplasias malignas, que são responsáveis por mais de 15% de todas as mortes no país. Diante da estimativa de mais de 600 mil novos casos estimados para 2016, o câncer compreende a segunda causa de morte da população brasileira, tirando a vida de cerca 225 mil indivíduos anualmente (JEMAL et al., 2014).

Dentre os tipos de câncer existentes, o câncer de pulmão é o diagnosticado com mais frequência, responsável por praticamente uma em cada cinco mortes no mundo, e conta com um maior número de mortes relacionadas a este tipo de doença (HOWLADER et al., 2016). Estimou-se um quantitativo de 1,8 milhões de novos casos diagnosticados de câncer de pulmão em 2012 (13% de todos os novos casos de câncer), e foi responsável por cerca de 22 mil mortes só no Brasil em 2011 (JEMAL et al., 2014; INCA, 2017).

O tabagismo ainda é o principal responsável pelo desenvolvimento da doença. Porém, outros fatores também oferecem risco: exposição a alguns tipos de metais, à poluição do ar relacionada principalmente à exaustão de motor a diesel, e a emissão da combustão derivada do carvão (INCA, 2016).

O diagnóstico do câncer de pulmão é feito principalmente com o uso de imagens de Tomografia Computadorizada (TC), pois é considerada na literatura como a principal ferramenta de visualização para detecção de nódulos pulmonares (BUSHBERG et al., 2011; DICHIOTTI et

al., 2010). Desde a sua aplicação na investigação do câncer, uma nova perspectiva foi aberta se comparada ao diagnóstico do câncer de pulmão por meio do Raio X digital, pois a TC fornece imagens 3D, formatação multiplanar (axial, coronal e sagital) além de uma alta resolução, permitindo uma análise com maior precisão de alguns dados acerca dos nódulos identificados, tais como: tamanho (altura, largura e volume), localização (lóbulo e a distância da pleura), calcificação benigna (presente ou ausente), forma (arredondada ou não) e borda (suavizada ou não), possibilitando um aumento da sensibilidade na detecção de nódulos (HENSCHKE et al., 1999). Além das informações clínicas, a análise destas características serão consideradas pelo radiologista no seu diagnóstico com o objetivo de definir se o paciente precisará ou não a fazer exames invasivos, como a biopsia.

O câncer de pulmão é uma doença agressiva, onde geralmente a presença de sintomas indica um sinal de prognóstico ruim (WU et al., 2013). Na maioria das vezes, quando a doença já se encontra em fases avançadas, o diagnóstico tardio impede o tratamento curativo (NOVAES et al., 2008). A taxa de sobrevivência do paciente com câncer de pulmão analisada em cinco anos é de apenas 15%, e se a doença for identificada em estágios iniciais, a taxa de sobrevivência salta para 49% (AGGARWAL; SARDANA; VIG, 2014). Portanto, a identificação de nódulos precoces tornou-se muito significativa no estudo do câncer de pulmão, pois, assim que são encontrados, são mais curáveis e opções de tratamento mais simples com menos custos podem estar disponíveis (REEVES; XIE; JIRAPATNAKUL, 2015). Além disso, a classificação da malignidade do nódulo pelo radiologista depende de aspectos temporais como taxa de crescimento e mudança de tamanho da lesão identificada entre dois ou três exames de TC realizados em tempos diferentes (REEVES et al., 2006). Assim, além de aumentar a probabilidade de sobrevivência do paciente, a identificação precoce dos nódulos pulmonares reduz a angústia do indivíduo, pois o paciente não necessita aguardar dias ou até mesmo meses para medir a mudança de tamanho, forma ou textura do nódulo, bem como reduz a quantidade de radiação à qual o paciente é exposto a medida que realiza os exames de TC para o acompanhamento do desenvolvimento da lesão.

Geralmente, nódulos são visualmente avaliados e verbalmente caracterizados com uma linguagem de atributos radiológicos e termos que são semi-quantitativos mas subjetivamente avaliados, como por exemplo: espiculado, suave, plano, esférico e outros. Estes atributos podem ter uma grande quantidade de variabilidade subjetiva, experimental e perceptual (BARTHOLMAI et al., 2015). O processo de interpretação da imagem médica tem mostrado significativa variação inter-observador em numerosos estudos devido a vários aspectos, como por exemplo: limitações de tempo, erro de percepção do leitor das imagens, falta de treino ou até mesmo fadiga (SIEGLE et al., 1998; AKGÜL et al., 2010).

Além dos desafios na interpretação de imagens médicas já mencionados, nódulos pulmonares pequenos (aqueles menores que 10mm em diâmetro) detectados em imagens TC, fazem o diagnóstico clinicamente difícil e podem tornar a decisão clínica complexa (HUA et al., 2015; YANKELEVITZ et al., 1999). Visto que, além de outros fatores, nódulos pequenos têm baixo

contraste em relação ao tecido do pulmão e podem estar anexados a estruturas complexas deste órgão (Figura 1.1) (ALILOU et al., 2014). Logo, o diagnóstico de nódulos pulmonares pequenos é uma tarefa desafiadora para os especialistas, porém muito importante para a sobrevivência do paciente.

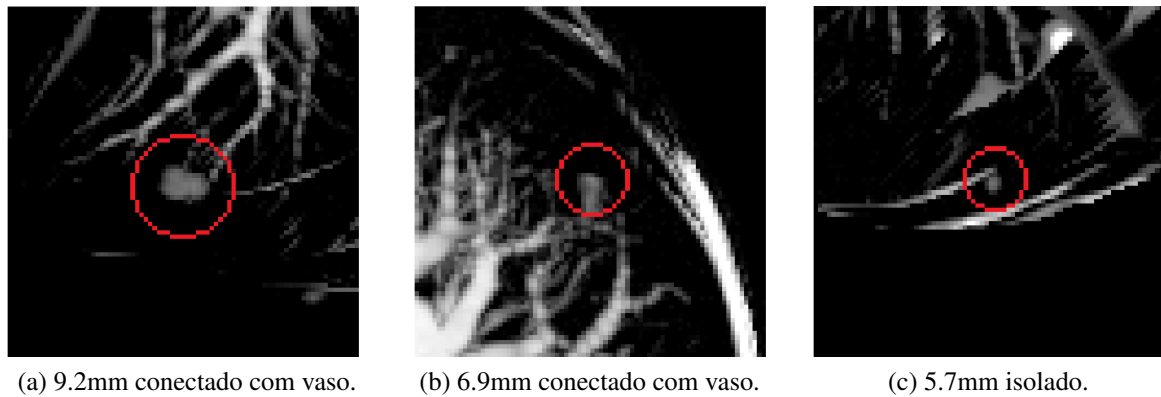


Figura 1.1: Exemplos de nódulos pulmonares (destacados em vermelho) com seus respectivos tamanhos e estrutura anatômica conexas a ele. Fonte: Imagem extraída de (ALILOU et al., 2014).

Ferramentas de Diagnóstico Auxiliado por Computador (do inglês *Computer-Aided Diagnosis - CAD*) têm sido desenvolvidas para auxiliar especialistas a interpretar achados radiológicos e identificar precocemente doenças, especialmente tumores de mama e pulmão (ERICKSON et al., 2017; GILLIES; KINAHAN; HRICAK, 2016; GIGER, 2010; DOI, 2007). O objetivo do CAD é melhorar a acurácia e consistência da interpretação e diagnóstico por imagem, mediante o uso da sugestão de resposta fornecida por técnicas de visão computacional e aprendizado de máquina (DOI, 2007; AZEVEDO-MARQUES, 2001). Tradicionalmente, sistemas CAD fornecem uma única resposta ao especialista, agindo como uma segunda opinião (DILGER et al., 2015). Apesar dos muitos sistemas propostos, ainda existem problemas a serem resolvidos, como a alta taxa de resultados falso-positivos e falso-negativos detectados nestes sistemas. Assim, ainda é preciso o desenvolvimento de técnicas de análise de imagens capazes de avançar o estado-da-arte, com o objetivo de melhorar o desempenho no auxílio ao diagnóstico do câncer de pulmão (AL-ABSI; SAMIR; SULAIMAN, 2014). A melhora na taxa de acerto desse tipo de sistema poderá, em um futuro próximo, ajudar o especialista em um ambiente clínico no diagnóstico precoce e aumentar as chances de sobrevivência dos pacientes com câncer.

Outro desafio da área de análise de imagens, é que os nódulos pequenos possuem uma quantidade de *voxels* muito limitada em imagens de TC, dificultando o uso de ferramentas de auxílio ao diagnóstico (DILGER et al., 2015). Segundo um exemplo dado por Reeves (REEVES; XIE; JIRAPATNAKUL, 2015), um nódulo de 2mm se estende em uma ordem de 8 pixels, um nódulo de 3mm a 27 pixels, 4mm a 64 pixels e um nódulo de 5mm a 620 pixels. Contudo, no microambiente do nódulo (habitat) existe uma interação significativa entre o parênquima do pulmão e o nódulo, porém, a maioria dos atributos usados na literatura para classificar o nódulo pul-

monar são derivados apenas do nódulo, mesmo existindo trabalhos ressaltando a importância em usar o microambiente na classificação de nódulos pulmonares, principalmente no tocante a nódulo pequeno, onde a limitação baixa de tamanho/informação é um desafio (DILGER et al., 2015; DILGER; JUDISCH; UTHOFF, 2015). Portanto, levar em conta o microambiente do nódulo, como a região do parênquima do pulmão, pode ser relevante para classificação de nódulos pequenos.

Um conceito importante encontrado na literatura que representa um importante avanço na análise quantitativa de imagem, chama-se *radiomics*. *Radiomics* é a extração e análise em larga escala de atributos quantitativos de imagens médicas usando avançados algoritmos matemáticos visando revelar características das lesões que podem passar despercebidas se analisadas a olho nu (ZHANG et al., 2015; YIP; AERTS, 2016). O uso de uma abordagem *radiomics* permite avaliar e diagnosticar o câncer de forma mais objetiva/quantitativa, permitindo reduzir significativamente o problema da variabilidade intra/inter-observador e melhorando a acurácia no diagnóstico se comparado as recentes estratégias qualitativas de avaliação do câncer (KUMAR et al., 2015). Além disso, o uso de *radiomics* na detecção precoce tem o potencial de auxiliar os médicos no diagnóstico, ajudando a selecionar o melhor tipo de tratamento para os pacientes individualmente, podendo então aumentar as chances de tratamento e cura da doença (YIP; AERTS, 2016).

Vários atributos quantitativos têm sido usados para caracterizar nódulos pulmonares, como: textura, forma, densidade, etc (W. J. Choi and T. S. Choi, 2014; OLIVEIRA; FERREIRA, 2013; Y.-X. J. Wang, J.-S. Gong, K. Suzuki, and S. K. Morcos, 2014). Em particular, alguns estudos têm usado esses atributos quantitativos na caracterização de nódulos pulmonares pequenos (REEVES; XIE; JIRAPATNAKUL, 2015). Mas a questão se estes parâmetros quantitativos são aptos a conferir uma vantagem ou não na classificação entre nódulos benigno e maligno ainda permanece em aberto (DILGER et al., 2015). Logo, existe uma necessidade de descobrir conteúdos relevantes de imagem a fim de melhorar a performance dos algoritmos de classificação.

A Classificação é uma das etapas de um sistema CAD e está relacionada a construção de modelos baseados em métricas sobre os atributos recuperados na etapa de caracterização a fim de entregar uma probabilidade de malignância do nódulo (REEVES; KOSTIS, 2000). Aprendizado de máquina é um conceito que unifica ideias relacionadas a problemas com métodos estatísticos de tomada de decisão, tem tido uma ampla aplicação na computação moderna, como: negócios inteligentes, detecção de e-mail spam e fraude em cartões de créditos. Já partindo para o campo de imagens médicas, podemos dizer que as técnicas de aprendizado de máquina são adotadas de forma mais contida, apesar de ser aplicada neste campo há décadas, principalmente em ferramentas CAD. Porém, o avanço do poder computacional vem aumentando cada vez mais o interesse da comunidade científica em empregar o uso de algoritmos avançados a fim de facilitar o uso de imagens médicas (SUZUKI, 2013; WERNICK et al., 2010). Muitos algoritmos de aprendizado de máquina vêm sendo aplicados tanto na classificação como detecção de

nódulos pulmonares: *k-Nearest Neighbor* (k-NN) (NAMIN et al., 2010), Rede Neural Artificial (RNA) (KURUVILLA; GUNAVATHI, 2014), *Random Forest* (RF) (TARTAR; KILIC; AKAN, 2013), Máquina de Vetor de Suporte (do inglês *Support Vector Machine* - SVM) (NUZHAYAYA et al., 2011).

1.1 Objetivo

O objetivo principal deste trabalho foi desenvolver um modelo para classificação de nódulos pulmonares pequenos usando atributos *radiomics* extraídos da região de microambiente do nódulo.

Como objetivo secundário, podemos citar o teste de hipótese de que considerar o parênquima do pulmão na extração dos atributos permite aumentar a acurácia na caracterização de nódulos pequenos, melhorando o desempenho da classificação.

1.2 Estrutura do Trabalho

A estrutura deste trabalho é dada da seguinte forma:

Capítulo 2 - Fundamentação Teórica. Este capítulo descreve todo cenário de contextualização do trabalho, trazendo todos os conceitos necessários para o entendimento do modelo desenvolvido;

Capítulo 3 - Materiais e Métodos. Este capítulo descreve toda metodologia aplicada para o desenvolvimento do modelo para classificação de nódulos pequenos, explicitando todos os passos necessários que foram realizados neste trabalho;

Capítulo 4 - Resultados e Discussão. Este capítulo apresenta alguns dos resultados obtidos;

Capítulo 5 - Conclusão. Este capítulo apresenta as conclusões e discute algumas direções futuras.

Capítulo 2

Fundamentação Teórica

2.1 Nódulos Pulmonares em Imagens de Tomografia Computadorizada

Até a década passada os exames de citologia do escarro e a radiografia de tórax (do inglês *chest X-ray* - CXR) foram usados para detectar o câncer de pulmão em sua fase inicial. Contudo, estes exames possuem eficácia limitada quando se observa a taxa de sobrevivência do paciente, possivelmente devido a limitação desses métodos em identificar os diferentes subtipos da doença (CUI et al., 2015). Atualmente a tomografia computadorizada de baixa dosagem (do inglês *low-dose CT* - LDCT) tem sido sugerida pela comunidade científica como o exame fundamental para realizar o rastreamento do câncer de pulmão. A técnica apresentou resultados efetivos na redução da mortalidade pela doença devido a sua capacidade de diagnosticar precocemente o câncer, quando a probabilidade de cura é maior (PATZ et al., 2014; LIANG et al., 2016).

A Tomografia computadorizada (TC) pode ser descrita como um processo no qual um anel de detectores envolve um paciente, e uma fonte emite raios X concêntricos com o anel detector, girando em torno do paciente. Os raios emitidos pela fonte de raios X interagem com os tecidos dos órgãos do paciente, são atenuados de forma diferenciada e são coletados no lado oposto ao da emissão pelos detectores correspondentes no anel. Este processo é repetido conforme o emissor gira (Figura 2.1). A aplicação de algoritmos sobre os dados capturados permite a construção de uma imagem que representa uma fatia do paciente, e o conjunto de todas estas fatias forma um volume de imagens (GONZALEZ; WOODS, 2008).

Exames de TC são comumente adquiridos em forma de volume de fatias paralelas e uniformemente espaçadas. O empilhamento destas imagens, mantendo o espaço original entre elas, pode ser idealizado de forma que cada *pixel* represente o volume de um *voxel*. O *voxel* tem como característica principal a área em torno da malha de pontos com o mesmo valor. Portanto, um *voxel* é uma área hexaédrica de valor constante em torno de uma malha de pontos central (OLIVEIRA, 2002). O conjunto de *voxels* forma uma representação digital de uma região cúbica em estudo no paciente e é denominado espaço *voxel*, onde cada *voxel* tem normalmente

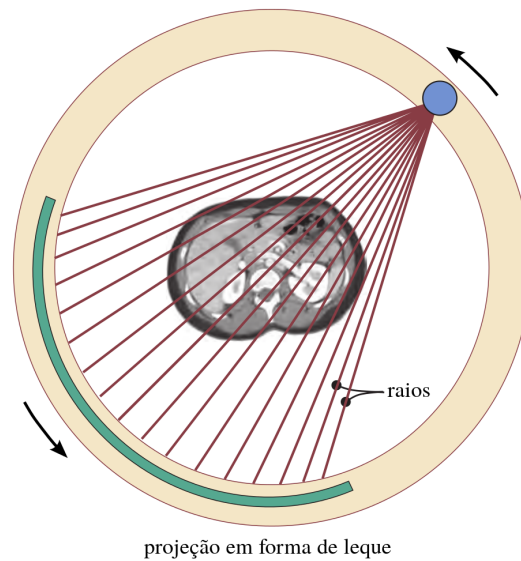


Figura 2.1: Representação gráfica de um aparelho de TC tradicional. O tubo de raios X gira em torno do paciente, emitindo um feixe colimado de raios X em forma de leque, direcionado ao anel estacionário de detectores. A imagem é formada de acordo com as variações de densidades dos tecidos atingidos pela emissão dos raios. Fonte: (BUSHBERG et al., 2011).

associado um número inteiro proporcional ao tom de cinza do *pixel* na imagem correspondente (Figura 2.2).

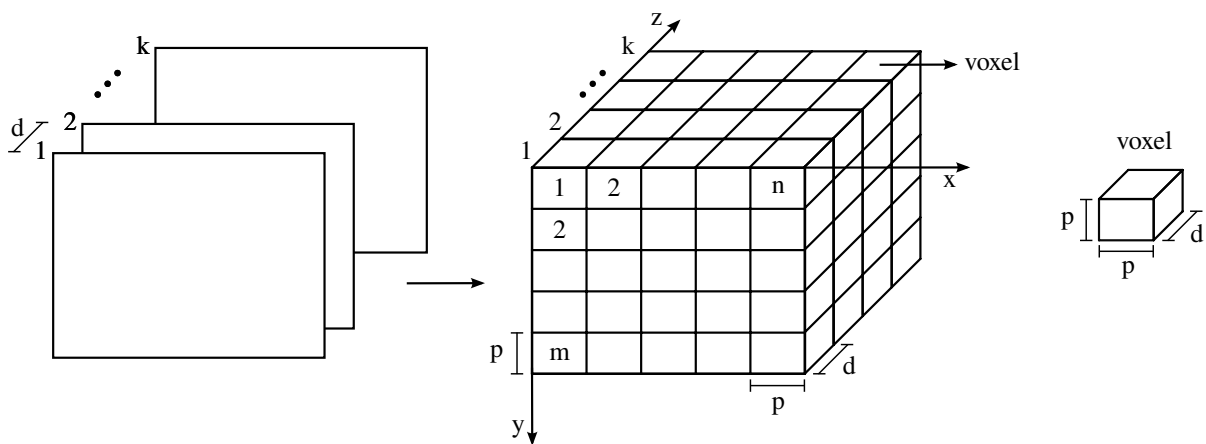


Figura 2.2: Representação de um espaço *voxel*. Fonte: elaborado pelo autor.

2.2 Intensidade do Nódulo

Um nódulo pode ser visto em uma imagem de TC em formas diferentes quanto a sua intensidade, e então ser caracterizado como sólido, semi-sólido ou até mesmo como não sólido. Nódulos classificados como sólidos são densos e possuem alta intensidade, o que os leva a ter uma aparência sólida (Figura 2.3a).

Em alguns casos, os nódulos podem possuir uma parte sólida com uma parte nebulosa em seu entorno. Aos nódulos com essa característica, damos o nome de nódulos semi-sólidos. Estes tipos de nódulos são menos densos e possuem uma intensidade média (Figura 2.3b).

Nódulos que possuem uma aparência nebulosa, com partes não sólidas, são chamados de não sólidos ou nódulos vidro-fosco (do inglês *ground glass*), pois possuem uma aparência de vidro-fosco. É possível ver através deles, mas ainda permanecem visíveis em imagens de TC, pois, apesar de possuírem uma intensidade muito menor, se comparados a nódulos sólidos, eles possuem uma densidade maior que o restante do tecido pulmonar (Figura 2.3c).

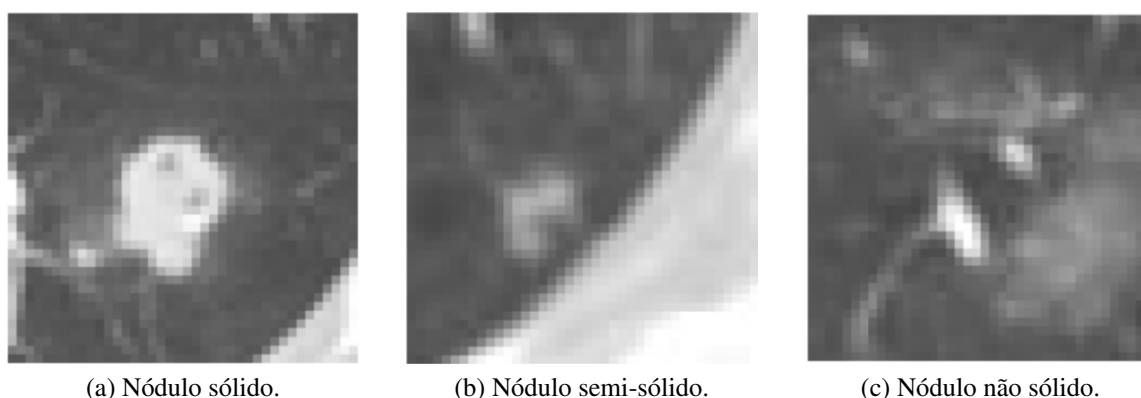


Figura 2.3: Aparência dos nódulos em uma imagem de TC quanto a intensidade. Fonte: Imagem extraída de (NITHILA; KUMAR, 2016).

É importa mencionar que, nódulos dos três tipos podem ser malignos, porém a consistência pode servir de grande ajuda para o especialista decidir como proceder no acompanhamento do paciente (NITHILA; KUMAR, 2016).

2.3 Diagnóstico Auxiliado por Computador

Com o objetivo de auxiliar o radiologista na interpretação de imagens médicas, CAD fornece uma maneira de reduzir o tempo de interpretação, aumentar a sensibilidade da detecção e melhorar a acurácia do diagnóstico. Além disso, ferramentas CAD aptas a identificar nódulos pequenos com baixo risco de malignidade, podem minimizar repetições de aquisição de imagens de TC e, conseqüentemente, diminuir a exposição do paciente à radiação, bem como evitar aumentar seu estresse emocional (WANG; RONALD, 2012; DILGER et al., 2015).

Ferramentas CAD servem como uma segunda opinião ao especialista, que visa guiá-lo na sua tomada de decisão acerca do diagnóstico de doenças. Vale ressaltar que, o computador deve ser entendido como uma ferramenta de auxílio e, portanto, não deve ter seu desempenho comparado ao de um ser humano, mas deve promover um aumento da precisão através do efeito sinérgico das competências do médico com a capacidade de processamento de informações do computador (DOI, 2007; AZEVEDO-MARQUES, 2001).

CAD pode ser definido como uma categoria de sistema que permite que o diagnóstico médico seja feito com o auxílio de informações extraídas usando-se análises quantitativas realizadas pelo computador. Tais informações são obtidas por meio de descritores/atributos de imagem que são relevantes a análise de similaridade, e o desempenho de sistemas CAD depende diretamente da escolha dos descritores e da técnica de similaridade.

2.4 Radiomics

A conversão de imagens médicas digitais em dados de alta dimensionalidade é motivada pelo conceito de que imagens possuem informações que refletem fisiopatologias difíceis de serem identificadas, e que essa relação pode ser revelada por meio de uma análise quantitativa de imagem (GILLIES; KINAHAN; HRICAK, 2016).

Radiomics é um prática que vem sendo muito bem desenvolvida na oncologia, sendo justamente definida como um processo designado para extração em larga escala de atributos quantitativos a partir de imagens digitais, para subsequente análise destes dados para geração de hipóteses. Estes atributos, os quais podem ser baseados em intensidade, forma, tamanho ou volume, textura, combinados com outras informações como: histológicas, genéticas e até mesmo dados clínicos, podem ser usadas para auxílio na tomada de decisão clínica baseada em evidências (Figura 2.4). *Radiomics* surge para o fornecimento quase ilimitado de biomarcadores de imagem, os quais possuem o potencial de auxiliar na detecção, diagnóstico, monitoramento do câncer, entre outros benefícios (GILLIES; KINAHAN; HRICAK, 2016).

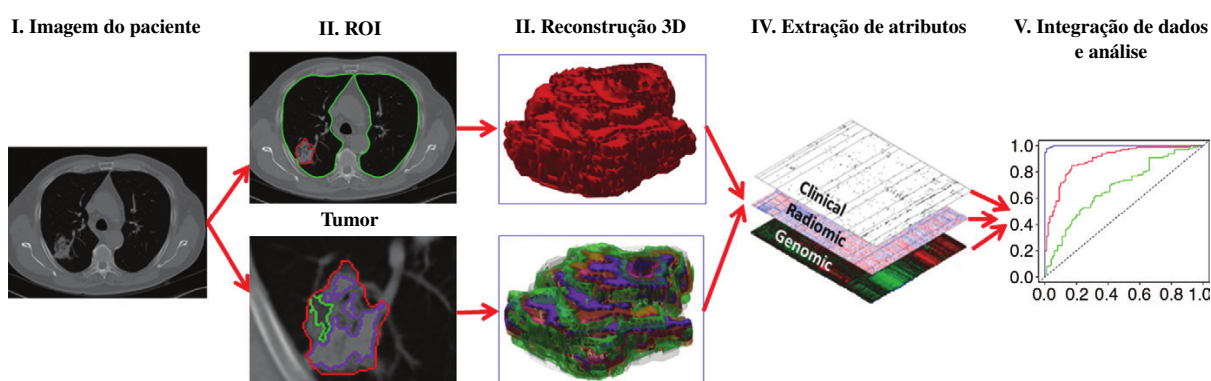


Figura 2.4: Ilustração do processo *radiomics* no auxílio a decisão. O processo se inicia com a aquisição de imagens de alta qualidade (I). Uma ROI contendo o tumor ou uma sub-região (microambiente) é identificada a partir destas imagens (II), para em seguida ser segmentada e reconstruída em 3D (III). Atributos quantitativos são extraídos destes volumes renderizados e integrados com outros dados, tais como dados clínicos e genéticos (IV). Estes dados são então correlacionados para o desenvolvimento de modelos preditivos (V). Fonte: Imagem extraída de (GILLIES; KINAHAN; HRICAK, 2016).

Radiomics é uma extensão natural de sistemas CAD, embora significativamente haja diferenças entre eles. Sistemas CAD possuem o papel de entregar ao especialista uma simples

resposta ao fim do processo, esta resposta pode referir-se a presença de uma lesão ou câncer, por exemplo. Já *radiomics* está relacionado a correlação de dados baseado em características quantitativas (atributos de imagem) bem como até características qualitativas (GILLIES; KINAHAN; HRICAK, 2016; AERTS et al., 2014; WU et al., 2016).

2.5 Descritores de Imagem

O processo de extração de atributos de imagem consiste na obtenção de valores numéricos que representam o conteúdo visual da imagem (descritores de imagem) através da implementação de algoritmos, como reconhecimento e classificação de texturas, formas e contornos, estimativas de área e volume (SILVA, 2009). Após a extração destes descritores, os atributos são armazenados em um vetor de atributos (Figura 2.5).

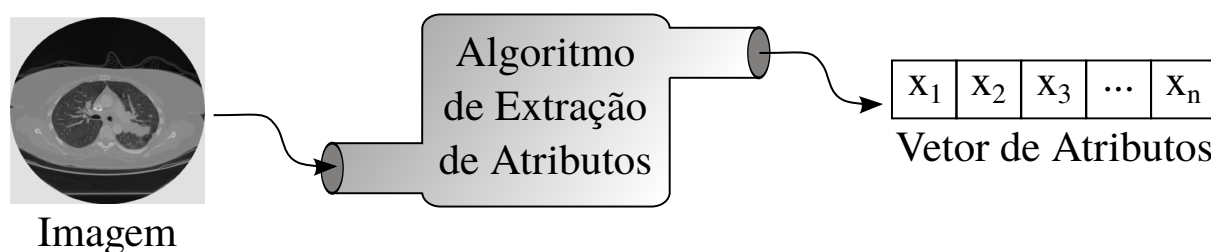


Figura 2.5: Diagrama de extração de um vetor de atributos. Fonte: elaborado pelo autor.

Os especialistas utilizam informações visuais, textuais, conhecimentos teóricos, suas próprias experiências e até mesmo consultas a outros especialistas para a tomada de decisão, enquanto a máquina faz uso de representações matemáticas para modelar os objetos de interesse e, a partir delas, tomar decisões acerca dos nódulos apresentados. Portanto, existem diferenças nos parâmetros utilizados pelo homem e pela máquina. Esta disparidade é conhecida na literatura como *gap* semântico de sistemas CAD (BEDO et al., 2016).

2.5.1 Atributos de Intensidade

Atributos de Intensidade (ATI) incluem informações a respeito do histograma da imagem (DILGER; JUDISCH; UTHOFF, 2015). O histograma reflete a ocorrência do número de *pixels* em cada diferente valor de intensidade encontrado numa imagem. Por exemplo, em uma imagem 8-bit em escala de cinza, existem 256 intensidades diferentes possíveis, e então, o histograma exibirá graficamente 256 valores com a distribuição dos *pixels* juntamente com seus respectivos valores em escala de cinza (RAMLI et al., 2008; LU et al., 2015) (Figura 2.6).

ATI fazem parte de uma categoria de atributos fácil de se calcular, em contrapartida, fornecem informação discriminatória à partir da diferença de intensidade dos nódulos benignos e malignos devido as diferenças entre os tecidos (SHEWAYE; MEKONNEN, 2016).

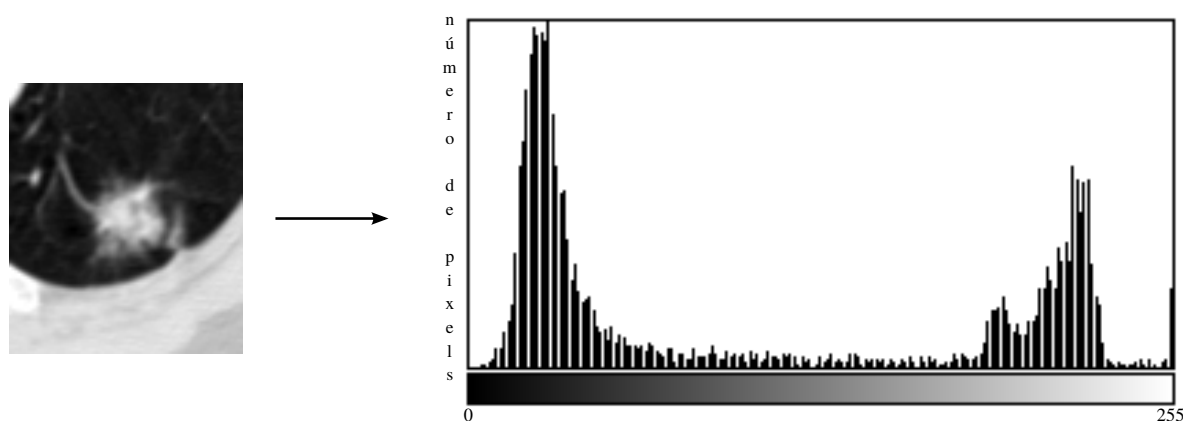


Figura 2.6: Imagem ROI de uma TC do pulmão com uma lesão em nível de cinza e seu respectivo histograma de intensidade. Fonte: elaborado pelo autor.

2.5.2 Atributos de Forma

O termo forma refere-se a informação que pode ser inferida da imagem e que não pode ser representada por cor e textura. Uma maneira de representar Atributos de Forma (AF) é através de características geométricas extraídas da imagem ou do objeto segmentado, e.g. bordas, contornos, junções, curvas e regiões poligonais (AKGÜL et al., 2010).

Existem muitos casos em que é possível caracterizar nódulos como maligno ou benigno analisando apenas sua forma. Um nódulo arredondado ou que tenha uma forma bem definida, é provavelmente benigno. Por outro lado, um nódulo espiculado ou que tenha uma configuração irregular, é um provável maligno (SILVA; CARVALHO; GATTASS, 2005). A Figura 2.7 mostra alguns exemplos de tais características.

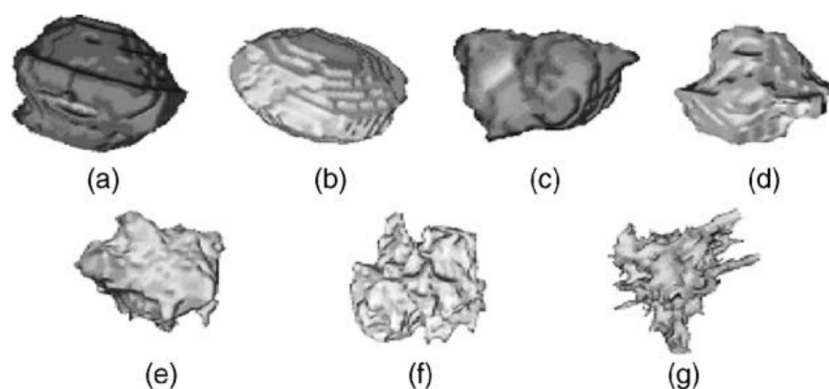
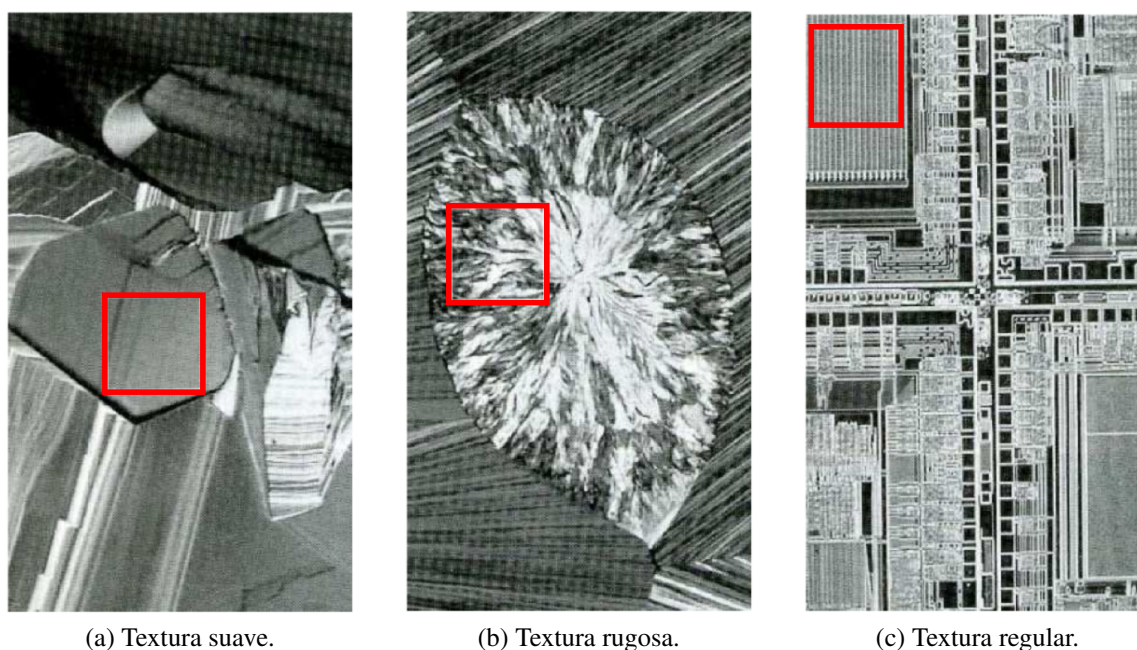


Figura 2.7: Exemplos de tipos de forma dos nódulos. Na figura, temos exemplos de nódulos benignos, com uma forma bem definida (a-d) bem como de nódulos malignos, com uma forma mais espiculada (e-g). Fonte: Imagem extraída de Corrêa (SILVA; CARVALHO; GATTASS, 2005).

2.5.3 Atributos de Textura

Apesar de não haver um consenso sobre a definição formal de textura, Parker afirma que a maior característica de textura é a repetição de padrões sobre uma região (J. R. Parker, 1997). Este padrão pode ser repetido exatamente igual, ou como um conjunto de pequenas variações no tema, possivelmente uma função da posição. Alguns aspectos da textura podem variar entre as regiões e não devem ser ignorados, como por exemplo: tamanho, forma, cor e orientação dos elementos do padrão (Figura 2.8).



(a) Textura suave.

(b) Textura rugosa.

(c) Textura regular.

Figura 2.8: Exemplos de texturas delimitadas em vermelho. Fonte: Imagem extraída de (GONZALEZ; WOODS, 2010).

No cenário de câncer, Atributos de Textura (AT) se tornaram importantes devido a sua capacidade de refletir detalhes contidos dentro de uma lesão em uma imagem (AKGÜL et al., 2010). Alguns aspectos da textura podem variar entre as regiões e não devem ser ignorados. A variação nos padrões de textura dos nódulos pulmonares fornece fortes indicadores da sua natureza maligna ou benigna. Por exemplo, a presença de gordura ou calcificação são fortes indicadores de um tumor benigno e resulta em uma distribuição irregular de textura. Por outro lado, nódulos malignos possuem textura uniforme produzida pela presença de necrose (ERASMUS et al., 2000; TAKASHIMA et al., 2003).

Um dos métodos de análise textural mais utilizados em recuperação de nódulos pulmonares foi proposto em (HARALICK; SHANMUGAM; DINSTEN, 1973), usando a análise estatística. Os descritores de textura estatísticos de Haralick et al. são classificados como de segunda ordem. Características de textura de segunda ordem ou matriz de coocorrência obtém informações sobre o posicionamento dos *pixels* (OLIVEIRA; CIRNE; MARQUES, 2007). A matriz de coocorrência é a representação da distribuição de probabilidade de ocorrência de pares de *pi-*

xels com determinada intensidade e distância em uma direção. Em outras palavras, a matriz de coocorrência envolve a estimativa de uma função de probabilidade de segunda ordem discreta, que representa a probabilidade de ocorrência de um par de *pixels* com níveis de cinza i e j , dada uma distância d e uma orientação θ entre os *pixels* nas dimensões x e y , respectivamente.

O cálculo da matriz de coocorrência é ilustrado na Figura 2.9. Onde, por exemplo, se desejássemos obter o valor do índice [1,4] da matriz de coocorrência (Figura 2.9b) deveríamos realizar a contagem dos pares de *pixels* iguais a [1,4] na matriz da imagem (Figura 2.9a), que tenham distância entre os índices igual a 1 e orientação em 0° . O mesmo procedimento é realizado para as orientações de 45° , 90° e 135° e os seus respectivos ângulos simétricos (OLIVEIRA, 2006).

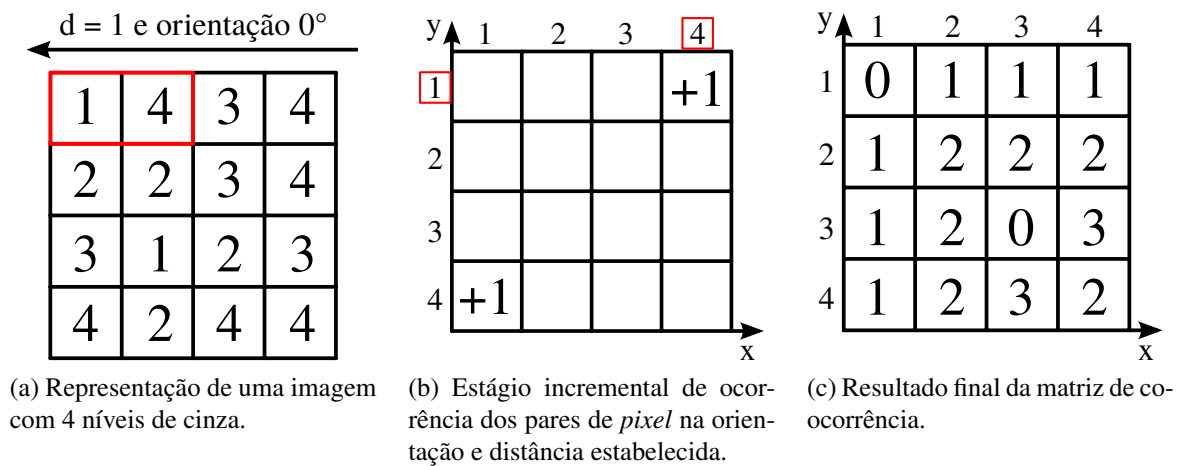


Figura 2.9: Ilustração do cálculo da matriz de coocorrência com orientação 0° (incluindo 180°) e distância 1. Fonte: Elaborado pelo autor.

O cálculo da matriz de coocorrência também pode ser realizado em um volume de imagens. A matriz de coocorrência 3D estende a avaliação da função de probabilidade para o eixo z , com o objetivo de examinar a probabilidade de ocorrência de pares de *pixels* entre fatias de um volume de imagens (MAHMOUD-GHONEIM et al., 2003).

Funções estatísticas de segunda ordem são aplicadas na matriz de coocorrência produzindo os atributos de textura. Alguns dos atributos de textura propostos por Haralick et al. são:

$$\text{Energia} = \sum_{i,j} C^2(i,j), \quad (2.1)$$

$$\text{Entropia} = - \sum_{i,j} C(i,j) \log C(i,j), \quad (2.2)$$

$$\text{Momento da diferença inverso (MDI)} = \sum_{i,j} \frac{C(i,j)}{1 + (i-j)^2}, \quad (2.3)$$

$$\text{Contraste} = \sum_{i,j} (i - j)^2 C(i, j), \quad (2.4)$$

$$\text{Variância} = \sum_{i,j} (i - \mu)^2 C(i, j), \quad (2.5)$$

$$\text{Matiz} = \sum_{i,j} (i + j - \mu_x - \mu_y)^3 C(i, j), \quad (2.6)$$

$$\text{Proeminência} = \sum_{i,j} (i + j - \mu_x - \mu_y)^4 C(i, j), \quad (2.7)$$

$$\text{Correlação} = - \sum_{i,j} \frac{(i - \mu_x)(j - \mu_y)}{\sqrt{\sigma_x \sigma_y}} C(i, j), \quad (2.8)$$

$$\text{Homogeneidade} = \sum_{i,j} \frac{C(i, j)}{(1 + |i - j|)}, \quad (2.9)$$

onde $C(i, j)$ são elementos $[i, j]$ da matriz de co-ocorrência, μ_x, μ_y são médias e σ_x e σ_y são desvios padrões, obtidos segundo as equações abaixo:

$$\mu_x = \sum_i i C_x(i), \quad (2.10)$$

$$\mu_y = \sum_j j C_y(j), \quad (2.11)$$

$$\sigma_x = \sum_i (i - \mu_x)^2 \cdot \sum_j C(i, j), \quad (2.12)$$

$$\sigma_y = \sum_j (j - \mu_y)^2 \cdot \sum_i C(i, j), \quad (2.13)$$

$$C_x(i) = \sum_j C(i, j), \quad (2.14)$$

$$C_y(j) = \sum_i C(i, j). \quad (2.15)$$

2.5.4 Atributos de Nitidez de Borda

Atributos de Nitidez de Borda (ANB) são importantes para diferenciar lesões em termos de potencial de malignidade por causa que tumores de câncer crescem em tecidos vizinhos aos nódulos (LEVMAN; MARTEL, 2011). Uma margem mais nítida, por exemplo, terá uma transição mais abrupta, além de poder ter uma alta diferença de intensidade observando o lado

de dentro e o lado de fora da lesão, enquanto que uma margem borrada terá uma transição mais suave e pode ter uma pequena diferença de intensidade (XU et al., 2012).

Xu (XU et al., 2012) definiu a nitidez de borda de nódulos pulmonares em duas características. A primeira mede a diferença de intensidade entre os *pixels* do tecido do pulmão e do tecido do nódulo. A segunda mede a transição de intensidade dos *pixels* sobre a borda do nódulo. As duas características são extraídas de uma função sigmoide aplicada em um vetor de intensidade formado pelos *pixels* presentes em linhas ortogonais traçadas automaticamente na borda do nódulo (Figura 2.10).

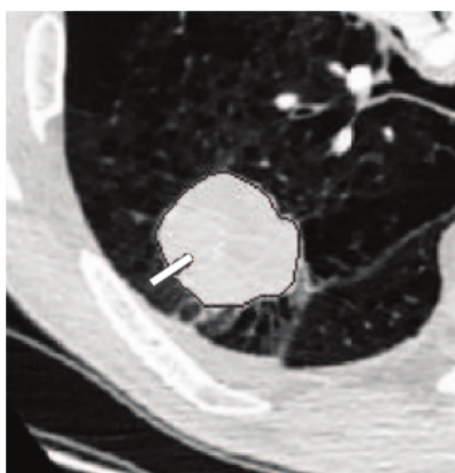


Figura 2.10: Linha ortogonal traçada sobre a borda do nódulo pulmonar. Fonte: Imagem extraída de (XU et al., 2012).

Ferreira Junior (JUNIOR; AZEVEDO-MARQUES; OLIVEIRA, 2016) realizou a extração de ANB 3D usando um método semelhante ao de Xu, no qual linhas perpendiculares a borda do nódulo foram desenhadas em todas as fatias, usando-se as imagens originais de TC. A estratégia usada foi como segue: 20 pontos de controle foram automaticamente selecionados a partir das marcações nas arestas da lesão, iniciando pelo primeiro ponto da marcação realizada pelo radiologista (Figura 2.11a). Se a borda possui p *pixels*, então um ponto de controle é marcada a cada $\frac{p}{20}$ *pixels*. Linhas normais foram então desenhadas em cada um dos 20 pontos de controle em relação a borda do nódulo (Figura 2.11b). Uma máscara foi criada usando a aplicação de uma técnica de limiarização com o objetivo de eliminar os segmentos que atravessam a parede do pulmão, a fim de se evitar a introdução de informações que não pertenciam ao nódulo ou até mesmo a tecidos do pulmão (Figura 2.11c). Por fim, os segmentos de linhas normais que não pertenciam ao pulmão foram excluídos por meio da aplicação da máscara (Figura 2.11d).

Os segmentos de linha resultantes (intensidades dos *pixels*) foram armazenados em simples *arrays* ordenados. Então uma análise estatística desses dados foi realizada pela extração de atributos estatísticos a partir dos *pixels* de intensidade armazenados nestes *arrays* ordenados. Tais atributos são listados abaixo:

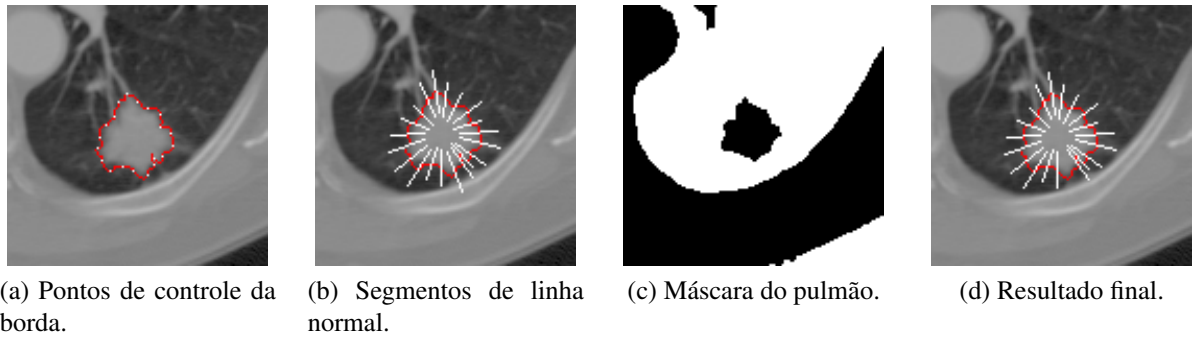


Figura 2.11: Imagens geradas a partir da análise de nitidez de borda. Fonte: Imagem extraída de (JUNIOR; OLIVEIRA, 2015).

$$\text{Diferença entre os extremos} = x_n - x_1, \quad (2.16)$$

$$\text{Soma dos valores} = \sum_{i=1}^n x_i, \quad (2.17)$$

$$\text{Soma dos quadrados} = \sum_{i=1}^n x_i^2, \quad (2.18)$$

$$\text{Soma dos logs} = \sum_{i=1}^n \log x_i, \quad (2.19)$$

$$\text{Média aritmética } (\mu) = \frac{1}{n} \sum_{i=1}^n x_i, \quad (2.20)$$

$$\text{Média geométrica} = \sqrt[n]{\prod_{i=1}^n x_i}, \quad (2.21)$$

$$\text{Variância da população} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2, \quad (2.22)$$

$$\text{Variância da amostra } (v) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2, \quad (2.23)$$

$$\text{Desvio padrão } (s) = \sqrt{v}, \quad (2.24)$$

$$\text{Medida de kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^4}{s^4}, \quad (2.25)$$

$$\text{Medida de skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^3}{s^3}, \quad (2.26)$$

$$\text{Segundo momento central} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}{s^2}, \quad (2.27)$$

em que x é o valor de intensidade do array de *pixels* de tamanho n ; x_1 é o valor de intensidade do *pixel* externo a região do nódulo e x_n o valor de intensidade do *pixel* na região interna ao nódulo.

2.6 Redução de Dimensionalidade

Uma quantidade muito grande de atributos em um algoritmo de aprendizagem de máquina pode levar a uma alta probabilidade de existência de ruídos ou atributos irrelevantes. Analisemos o seguinte caso, se cada atributo for visto como uma coordenada em um espaço d -dimensional, em que d é o número de atributos, o hipervolume que representa esse espaço cresce exponencialmente com a adição de novos atributos, adicionando um número maior de possibilidades de elementos nesse espaço, e como consequência, uma probabilidade maior de existência de conteúdos irrelevantes, o que termina dificultando o processo de aprendizagem. Este problema é conhecido como a maldição da dimensionalidade (MASON et al., 2010; FACELI et al., 2011).

A seleção de atributos nos ajuda a identificar atributos importantes, eliminando atributos irrelevantes, reduzindo ruídos e diminuindo a dimensão do vetor dos atributos.

2.6.1 Algoritmo Genético Evolutivo

O Algoritmo Genético Evolutivo (AGE) é baseado na genética e teoria evolucionária, onde os organismos mais adaptados ao ambiente são os mais aptos a terem seus atributos reproduzidos em uma nova geração. Segundo Holland (HOLLAND, 1992), AGE tem-se mostrado muito eficiente na busca de soluções ótimas, ou soluções próximas das ótimas, em uma grande variedade de problemas.

No AGE, inicialmente, é gerada uma população formada por um conjunto aleatório de indivíduos (ou cromossomos), que a princípio podem ser considerados como solução para o problema. Tradicionalmente, um indivíduo é representado por um vetor binário, onde cada elemento de um vetor representa a presença (1) ou ausência (0) de uma determinada característica (atributo).

Durante o processo evolutivo, a população é avaliada: para cada indivíduo é dada uma nota por meio de uma função de aptidão, refletindo sua habilidade de adaptação a determinado ambi-

ente, ou seja, fornecendo o quão boa é a solução codificada por um indivíduo. Uma porcentagem dos mais aptos é mantida, enquanto os outros são descartados (seleção natural).

Vários métodos de seleção têm sido propostos, dentre eles, temos o método de seleção por torneio. Neste método, n indivíduos da população são escolhidos aleatoriamente, com a mesma probabilidade. O cromossomo com a maior aptidão dentre os escolhidos é selecionado para população intermediária. O processo se repete até que a população intermediária seja preenchida.

Os membros mantidos pela seleção na população intermediária podem sofrer modificações em suas características fundamentais por meio de mutações (Figura 2.12) e cruzamento (*crossover*) (Figura 2.13), gerando descendentes totalmente novos para a próxima geração, mas que possuam de alguma forma, características de seus pais. Esse processo, ao qual dá-se o nome de reprodução, é repetido até que uma solução satisfatória seja encontrada (REZENDE, 2003).

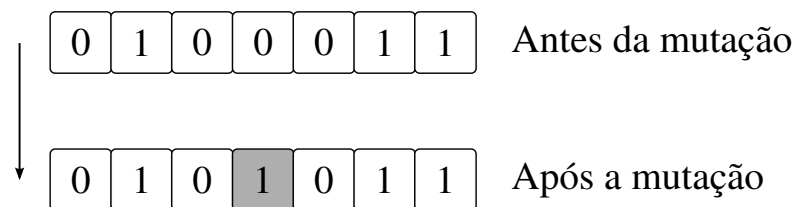


Figura 2.12: Exemplo de mutação, onde acontece a alteração aleatória em um dos componentes do indivíduo. Fonte: elaborado pelo autor.

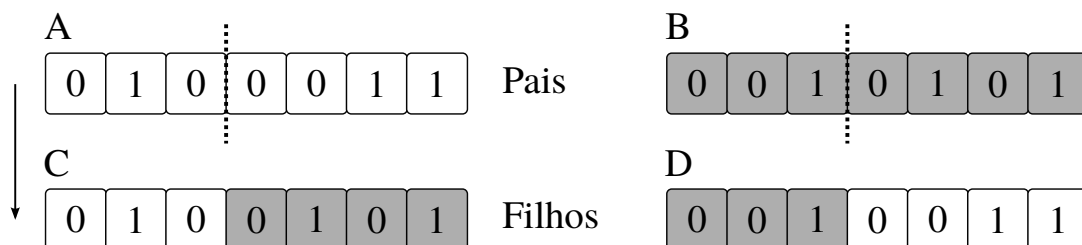


Figura 2.13: Exemplo de cruzamento de um ponto, onde é escolhido um ponto de cruzamento e a partir dele as informações genéticas dos pais são trocadas. Fonte: elaborado pelo autor.

Alguma das vantagens do algoritmos genéticos está no fato de que eles realizam buscas simultâneas em várias regiões do espaço de solução, permitindo assim que tais algoritmos encontrem um conjunto bastante variado de soluções, o que os tornam um método de busca global (FACELI et al., 2011). Embora seja um algoritmo aleatório, pode-se classificar o AGE como um algoritmo de busca de solução direcionada à busca de pontos de “alta aptidão”, ou seja, pontos onde são esperados melhores desempenhos (REZENDE, 2003).

2.7 Aprendizado de Máquina

Aprendizado de Máquina é a área da Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o processo de aprendizado, e está relacionada com a questão de como construir programas de computador que melhorem automaticamente seus desempenhos através da experiência (BISHOP, 2007; MITCHELL, 1997).

Dentre os tipos de Aprendizado de Máquina existentes, vale destacar a Aprendizagem Indutiva, que é a forma de inferência lógica que permite obter conclusões a partir de um conjunto de exemplos (generalização).

Aprendizagem Supervisionada é uma técnica de Aprendizagem Indutiva que cria uma função para um conjunto de dados de treinamento, fazendo-se uso do mecanismo de inferência indutiva (Figura 2.14). Tais conjuntos, consistem de pares de entrada de objetos x (vetor de atributos) e a saída desejada $f(x)$ (label). A saída dessa função pode ser dada de duas formas: a primeira, como um valor contínuo, e nesse caso estaremos tratando de um problema de *regressão*; a segunda, como um valor discreto, e nesse caso estaremos tratando de um problema de *classificação*. A tarefa do computador, nesse caso, é prever o valor de saída para qualquer objeto de entrada tendo visto um número de exemplos de treinamento. Ou seja, de maneira geral, o objetivo é estimar a função f , dado um conjunto de pontos na forma $(x, f(x))$.

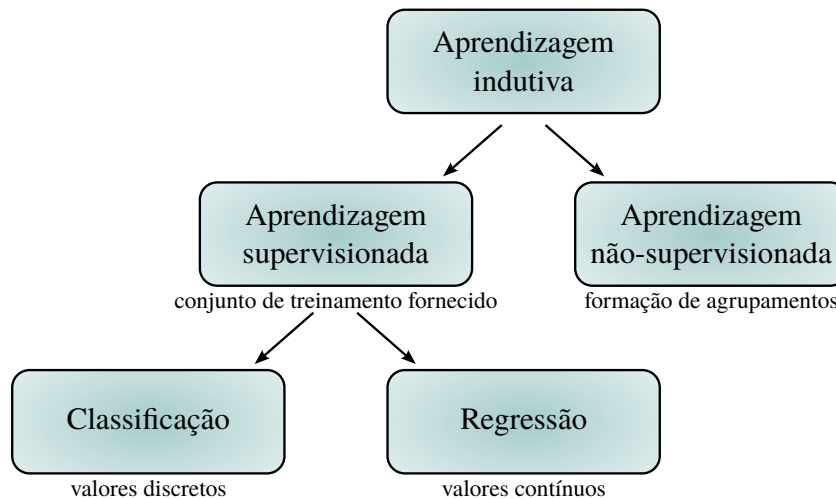


Figura 2.14: Hierarquia do aprendizado. Fonte: elaborado pelo autor.

Como exemplo de algoritmos de Aprendizagem Supervisionada, podemos citar: k-Nearest Neighbor (k-NN), Rede Neural Artificial (RNA) e Árvores de Decisão (AD).

2.7.1 k-Nearest Neighbor

k-NN é amplamente discutido e aplicado em padrões de reconhecimento e aprendizagem de máquina supervisionada (HU; YU; XIE, 2008). É um dos algoritmos mais simples de classificação usado para classificar objetos com base em exemplos de treinamento que estão mais próxi-

mos no espaço de características, classificando um novo objeto em função da classe majoritária dentre os k -vizinhos mais próximos a ele com base nos exemplos do conjunto de treinamento (FACELI et al., 2011).

O k -NN é baseado em três principais componentes:

1. Um conjunto de exemplos de treinamento empregado para futuras classificações (Figura 2.15a);
2. Uma função de similaridade, ou seja, uma métrica para calcular a distância entre os exemplos de treinamento e os exemplos a serem classificados. Como exemplo de funções de similaridade, temos: Distância Euclidiana e Distância Manhattan (LIMA; JUNIOR; OLIVEIRA, 2014) (Figura 2.15b);
3. Uma função de classificação, que tem como entrada um valor inteiro k , que representa o número de vizinhos mais próximos a ser considerado pelo algoritmo para classificar o exemplo solicitado. Como exemplo, a classe majoritária dentre os k vizinhos mais próximos pode ser usada como função de classificação (Figura 2.15c).

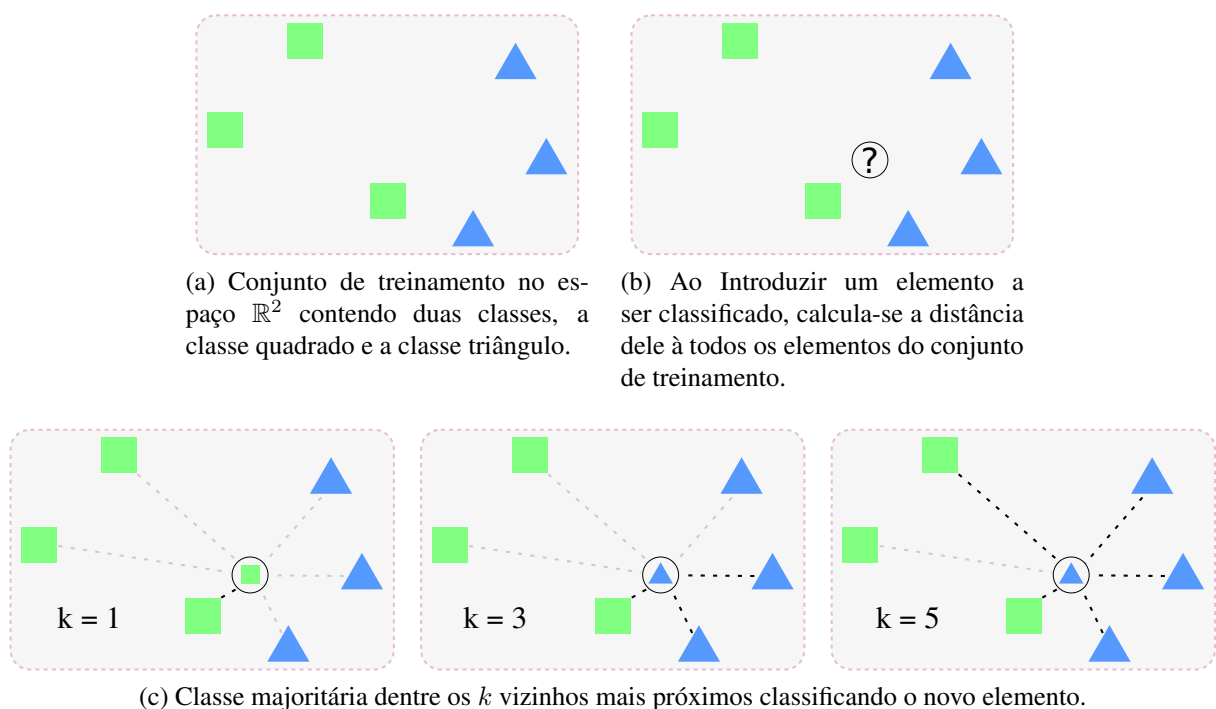


Figura 2.15: Ilustração das etapas do algoritmo k -NN. Fonte: elaborado pelo autor.

2.7.2 Rede Neural Artificial

As RNAs são modelos matemáticos que se assemelham às estruturas neurais biológicas (Figura 2.16) e que têm a capacidade computacional adquirida por meio do aprendizado e gene-

realização (BRAGA; CARVALHO; LUDERMIR, 2007; HAYKIN, 1999). O processo de aprendizado é interativo, e por meio dele a RNA deve melhorar o seu desempenho gradativamente à medida que interage com o meio exterior (REZENDE, 2003).

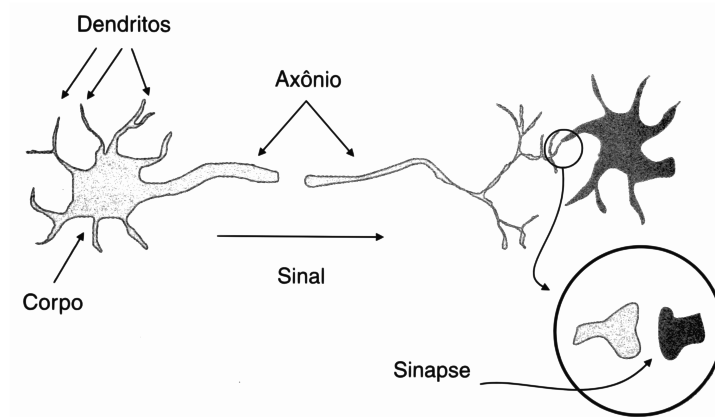


Figura 2.16: Neurônio biológico simplificado. Os Dendritos são responsáveis pela recepção de estímulos nervosos que são transmitidos para o Corpo celular, onde as informações são combinadas e processadas, e a depender da intensidade e frequência desses estímulos, o Corpo celular gera um novo impulso que é enviado pelo Axônio até o Dendrito de outro neurônio através da Sinapse (HAYKIN, 1999). Fonte: Imagem extraída de (FACELI et al., 2011).

Uma RNA pode ser descrita como um grafo direcionado cujo os nós correspondem aos neurônios e as arestas as ligações entre eles, que simulam as sinapses biológicas. Cada neurônio recebe como entrada uma soma ponderada das saídas de neurônios conectados as suas arestas de entrada, e os pesos têm seus valores ajustados em um processo de aprendizagem e codificam o conhecimento adquirido pela rede (SHALEV-SHWARTZ; BEN-DAVID, 2014; FACELI et al., 2011; BRAGA; CARVALHO; LUDERMIR, 2007).

A Figura 2.17 apresenta um modelo simples de neurônio artificial. Cada entrada do neurônio recebe um valor, que são ponderados na forma apresentada pela Equação 2.28, e então combinados por uma função matemática, a qual daremos o nome de função de ativação $f_a(u)$, equivalendo ao processamento realizado pela soma. A saída da função é a resposta do neurônio para a entrada. Várias funções diferentes pode ser utilizadas: linear, sigmoide, etc.

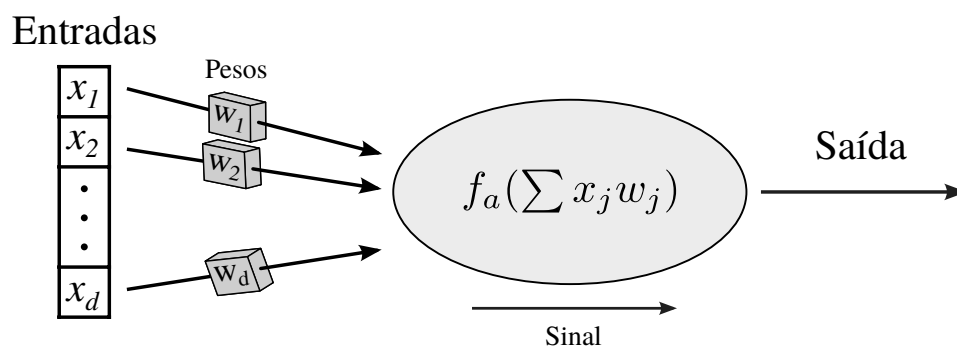


Figura 2.17: Ilustração de um Neurônio Artificial. Fonte: elaborado pelo autor.

$$u = \sum_{j=1}^d x_j w_j \quad (2.28)$$

Em uma RNA, os neurônios podem estar dispostos em uma ou mais camadas. A Figura 2.18 ilustra um exemplo de uma RNA com três camadas, que recebe como entrada valores de dois atributos e gera dois valores em sua camada de saída. As demais camadas são denominadas de camadas intermediárias/ocultas. É importante observar no exemplo que, todos os neurônios estão conectados a todos os neurônios da camada anterior e/ou seguinte, a esse tipo de rede, dá-se o nome de rede neural completamente conectada (FACELI et al., 2011).

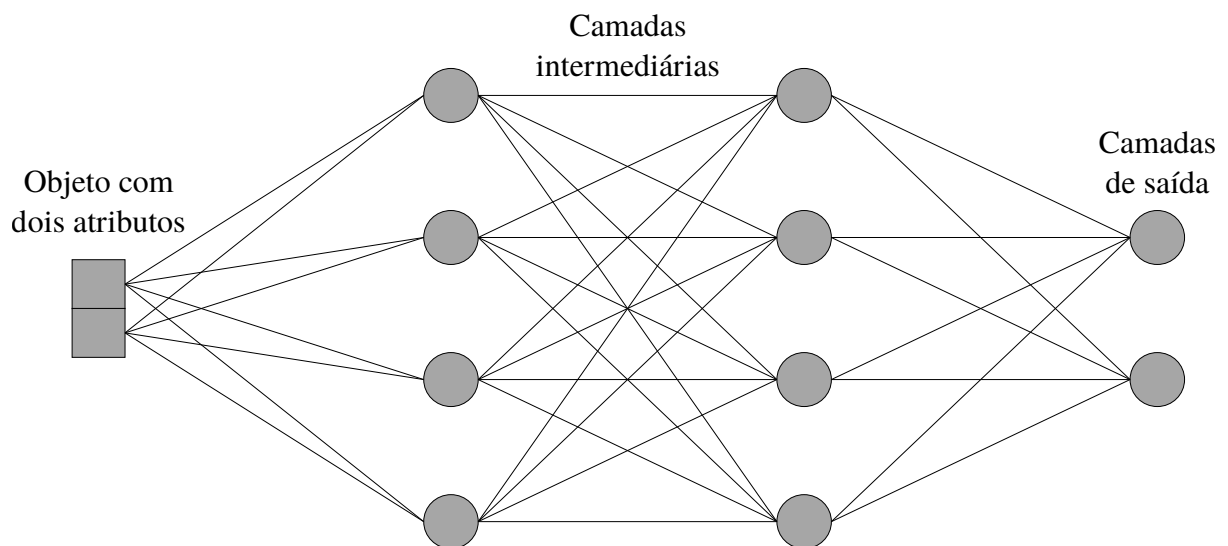


Figura 2.18: Exemplo de RNA multicamadas. Fonte: elaborado pelo autor.

A informação em uma rede neural geralmente flui da camada de entrada da rede para os neurônios da camada de saída. Para redes multicamadas, esse fluxo ocorre camada a camada. Redes sem conexões de retropropagação são denominadas RNAs *feedforward*.

O que foi discutido até agora está relacionado a arquitetura de um RNA, adiante, será mostrado o processo de aprendizado de uma RNA, que diz respeito às regras de ajuste de pesos.

Algoritmo *Back-propagation*

Um grande obstáculo que havia para utilização de redes multicamadas era a ausência de um algoritmo para treinamento, que foi vencido com a proposta do algoritmo denominado *back-propagation*. O algoritmo é constituído da iteração de duas fases, uma fase para a frente (*forward*) e uma fase para trás (*backward*). Na fase *forward*, cada objeto de entrada é apresentado à rede. O objeto é primeiramente recebido para cada um dos neurônios da primeira camada intermediária da rede, quando é ponderado pelo peso associado a suas conexões de entrada correspondentes. Cada neurônio nessa camada aplica a função de ativação a sua entrada total e

produz um valor de saída, que é utilizado como valor de entrada pelo neurônios da camada seguinte. Esse processo continua até que os neurônios da camada de saída produzam cada um seu valor de saída, que é então comparado ao valor desejado para saída desse neurônio. A diferença entre os valores de saída produzidos e desejados para cada neurônio da camada de saída indica o erro cometido pela rede para o objeto apresentado. Esse valor de erro é então utilizado na fase *backward* para ajustar seus pesos de entrada. O ajuste prossegue da camada de saída até a primeira camada intermediária. A Equação 2.29 ilustra como é feito o ajuste de pesos:

$$w_j(t + 1) = w_j(t) + \eta x_j \delta, \quad (2.29)$$

sendo w_j o peso do j -ésimo atributo de entrada do neurônio, δ indica o erro associado ao neurônio, x_j indica a j -ésima entrada recebida por esse neurônio e η o valor da taxa de aprendizado.

O valor de η tem forte influência no tempo de convergência da rede. Um valor muito pequeno, pode implicar em um tempo maior para se alcançar um bom modelo. Já um valor de taxa de aprendizagem elevado pode provocar oscilações que dificultam a convergência. Como alternativa para lidar com esse problema, tem-se o uso do termo *momentum* (α), que quantifica o grau de importância da variação do peso do ciclo anterior ao ciclo atual, tornando o processo de aprendizado mais estável e acelerando a convergência da rede (FACELI et al., 2011; RUMELHART, 1986). Com essa abordagem, o ajuste de pesos da rede utiliza a Equação 2.30.

$$w_j(t + 1) = w_j(t) + \eta x_j \delta + \alpha(w_j(t) - w_j(t - 1)) \quad (2.30)$$

Como os valores dos erros são conhecidos apenas para os neurônios da camada de saída, o erro para o neurônio da camada intermediária é estimado como a soma dos erros dos neurônios da camada seguinte, cujos terminais de entrada estão conectados a ele, ponderados pelo valor do pelo associado a essas conexões.

2.7.3 Árvore de Decisão

Uma AD é uma estrutura de dados definida recursivamente como um nó folha, que corresponde a uma classe, ou um nó de decisão, que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma subárvore. Cada subárvore tem a mesma estratégia que a árvore (REZENDE, 2003). As ADs Podem ser representadas como um conjunto de estruturas se-então, o que permite ter uma alta legibilidade com a linguagem humana.

ADs classificam instâncias selecionando o caminho através da árvore a partir do nó raiz até o nó folha. Cada nó da árvore especifica um teste de algum atributo da instância, e cada ramo descendente a partir desse nó, correspondente a um dos possíveis valores desse atributo. Uma instância é classificada começando do nó raiz da árvore, testando o atributo especificado por

esse nó, e então caminhando para baixo no ramo da árvore correspondente ao valor do atributo em um dado exemplo. Esse processo é repetido para cada subárvore até o novo nó (MITCHELL, 1997) (Figura 2.19).

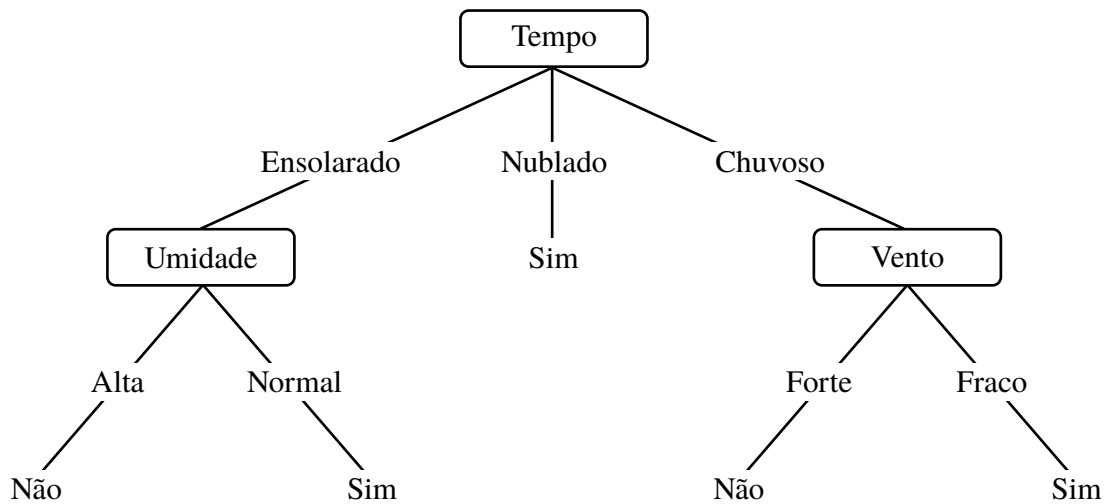


Figura 2.19: Ilustração de uma AD. No exemplo, *Tempo*, *Umidade* e *Vento* representam nós de decisões que contêm testes sobre o valor do respectivo atributo. *Não* e *Sim* representam nós folhas (Classes). Fonte: adaptado de (MITCHELL, 1997).

Como um dos critérios para escolha de cada nó da árvore, temos o ganho de informação, que possui como principal conceito a entropia. A entropia mede a aleatoriedade de uma variável aleatória, que no contexto de uma AD, significa dizer que, entropia mede a dificuldade para prever o atributo alvo. A cada nó de decisão, o atributo que mais reduz a aleatoriedade da variável alvo (aquele que possui máximo ganho de informação) será escolhido para dividir os dados.

Dada uma variável aleatória A cujo domínio é $\{a_1, a_2, \dots, a_v\}$. Suponha que a probabilidade de observar cada valor respectivamente seja p_1, p_2, \dots, p_v . A entropia de A é dada por:

$$H(A) = - \sum_i p_i \cdot \log_2 p_i \quad (2.31)$$

Para o caso particular de um atributo contínuo, ou seja, atributos cujo domínio seja formado por um subconjunto do conjunto dos números reais \mathbb{R} , o teste dividirá os dados em dois subconjuntos: $\text{atributo} > \text{valor}$ e $\text{atributo} \leq \text{valor}$ (Figura 2.20). Com o intuito de se obter um ponto de corte, os valores dos atributos contínuos são primeiro ordenados. O ponto médio entre dois valores consecutivos é um possível ponto corte e é avaliado por um função mérito. O possível ponto de corte que máxima a função mérito é escolhido (FACELI et al., 2011).

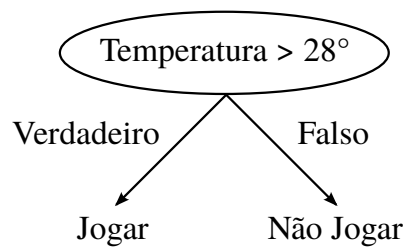


Figura 2.20: Ilustração de uma AD com atributo contínuo. Fonte: elaborado pelo autor.

Random Forest

Random Forest (RF) é um algoritmo que usa muitas ADs para realizar classificação, usando uma combinação de árvores preditoras, de tal modo que, cada árvore depende dos valores de um vetor aleatório, gerado de forma independente, com a mesma distribuição para todas as árvores da floresta. Após a geração de um grande número de árvores, um novo exemplo é classificado em cada árvore de forma independente. Logo em seguida, uma votação da classe mais popular/frequente é escolhida para classificação final do novo exemplo (BREIMAN, 2001).

2.7.4 Avaliação de desempenho

A avaliação de desempenho de um algoritmo de aprendizagem de máquina supervisionado geralmente é realizada por uma análise do desempenho do preditor gerado por ele na rotulação (classificação) de novos objetos, não apresentados em seu treinamento (MONARD; BARANAUSKAS, 2003).

No caso específico de problemas de decisão binária, um classificador rotula exemplos como positivo ou negativo. A decisão realizada pelo classificador na rotulação pode ser representada em uma estrutura conhecida como matriz de confusão ou tabela de contingência (Tabela 2.1). A matriz de confusão possui quatro categorias:

		Valor Predito	
		+	-
Valor Verdadeiro	+	VP	FN
	-	FP	VN

Tabela 2.1: Ilustração de uma Matrix de Confusão.

- Verdadeiro Positivo (VP): Exemplos corretamente classificados como positivos;
- Falso Positivo (FP): exemplos negativos incorretamente classificados como positivos;
- Verdadeiro Negativo (VN): exemplos corretamente classificados como negativos;
- Falso Negativo (FN): exemplos positivos incorretamente classificados como negativos.

Dada uma matriz de confusão, algumas métricas podem ser definidas, como a Taxa de Verdadeiros Positivos (TVP) e a Taxa de Falsos Positivos (TFP) definidas da forma:

$$TVP = \frac{VP}{VP + FN} \quad (2.32)$$

$$TFP = \frac{FP}{FP + VN} \quad (2.33)$$

TVP mede a fração de exemplos positivos rotulados corretamente. Já TFP, mede a fração de exemplos negativos incorretamente rotulados como positivo.

Uma maneira de avaliar classificadores em problemas que possuem apenas duas classes, é com o uso das curvas ROC (do inglês *Receiving Operating Characteristics*), as quais mostram como o número de exemplos positivos corretamente classificados varia com o número de exemplos negativos incorretamente classificados.

O gráfico ROC é bidimensional e plotado em um espaço denominado espaço ROC, com eixos x e y representando as medidas de TFP e TVP, respectivamente (Figura 2.21).

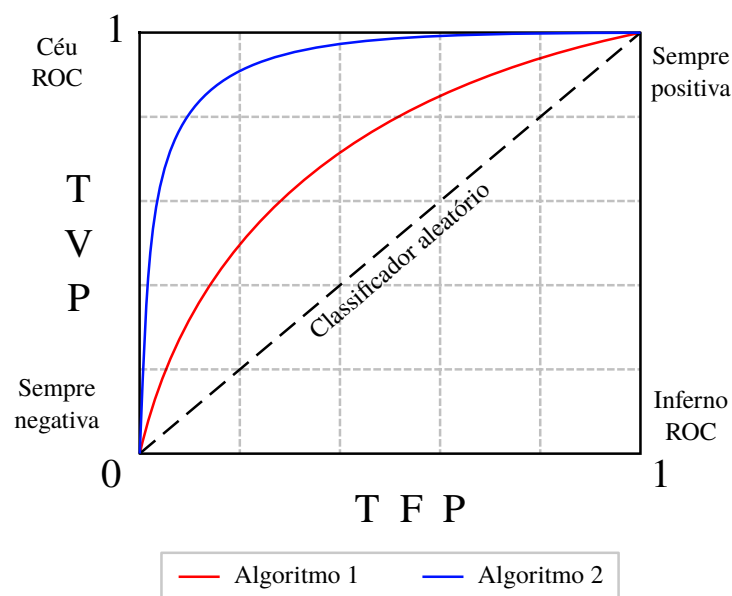


Figura 2.21: Espaço ROC. Fonte: adaptado de (FACELI et al., 2011).

Na Figura 2.21 é possível identificar alguns aspectos do espaço ROC. A linha diagonal representa classificadores que realizam previsões aleatórias. Um classificador abaixo dessa linha pode ser considerado pior que um aleatório. A região que se aproxima do ponto (1,0), damos o nome de inferno ROC. No ponto (0,1) do gráfico temos uma classificação considerada perfeita, onde todos os exemplos positivos e negativos são classificados corretamente (Céu ROC). Já o (0,0) representa classificações sempre negativas e (1,1), classificações sempre positivas.

Ao se comparar duas ou mais curvas, aquela que mais se aproxima do ponto (1,0) é a de melhor desempenho, ou seja, o objetivo de uma curva de um classificador em um espaço ROC é estar o mais próximo possível do canto superior-esquerdo do gráfico (DAVIS; GOADRICH, 2006). No exemplo da Figura 2.21, o Algoritmo 2 pode ser considerado melhor que Algoritmo 1.

Existem casos em que as curvas ROC dos algoritmos possuem interseção, neste caso, cada algoritmo tem uma região em que é melhor que a do outro. Porém, é muito comum comparar o desempenho dos algoritmos em termos de uma medida única extraída de sua curva ROC, a qual dá-se o nome de área sob a curva ROC (do inglês *Area Under the ROC Curve* - AUC). A AUC é uma porção de um quadrado unitário (o espaço ROC), e seu valor irá sempre estar entre 0 e 1 (FAWCETT, 2004). Quanto mais a AUC se aproxima de 1, melhor é considerada a classificação.

A fim de se evitar o erro/acerto aparente em uma avaliação, caso em que usa-se o mesmo conjunto de exemplos no treinamento e avaliação, é aconselhado calcular a AUC em um procedimento de estimativa de desempenho mais confiável, definindo-se subconjuntos de treinamento e de teste. A esses procedimentos, dá-se o nome de métodos de amostragem.

Dentre os métodos de amostragem existentes, vale destacar a validação cruzada *k-fold*, onde o conjunto de exemplos é dividido em k subconjuntos de tamanhos aproximadamente iguais. Os objetos das $k - 1$ partições são utilizados para treino de um preditor, o qual é então testado na partição restante. Esse processo é repetido k vezes, e em cada ciclo uma partição diferente é utilizada para teste. O desempenho final do preditor é dado pela média e desvio padrão dos desempenhos observado sobre cada subconjunto de teste (Figura 2.22) (FACELI et al., 2011).

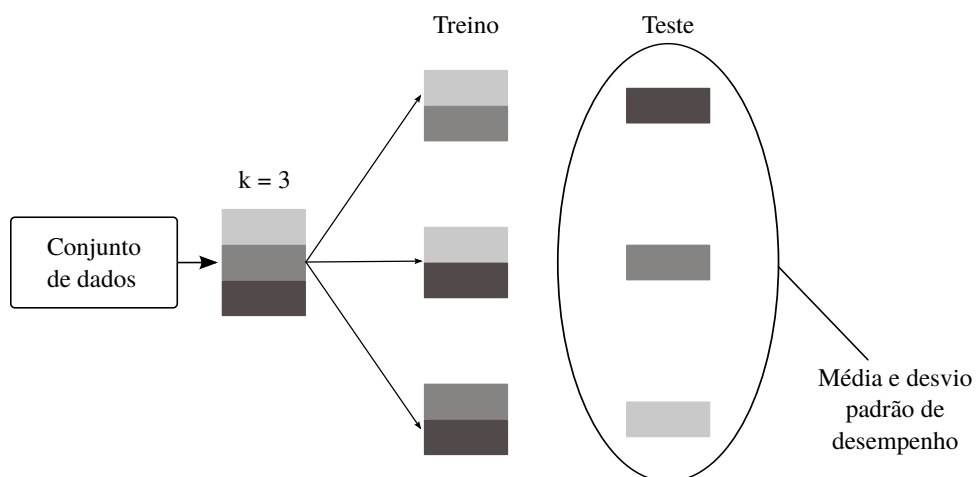


Figura 2.22: Ilustração de uma validação cruzada *3-fold*. Fonte: adaptado de (FACELI et al., 2011).

Capítulo 3

Materiais e Métodos

Uma visão geral da metodologia usada neste trabalho pode ser vista na Figura 3.1. Primeiro usou-se um banco de dados de nódulos pulmonares organizado em nosso laboratório por Ferreira Junior (JUNIOR; OLIVEIRA; AZEVEDO-MARQUES, 2016). O banco contém imagens do *Lung Image Database Consortium* (LIDC) (Seção 3.1). Foi realizada uma etapa de cálculo de tamanho e seleção dos nódulos de interesse (Seção 3.2). Em seguida, desenvolveu-se uma etapa para segmentação da região do parênquima em torno dos nódulos pulmonares (Seção 3.3), para enfim realizar a extração dos atributos de imagens dos nódulos selecionados e suas regiões de parênquima (Seção 3.4). Uma etapa de seleção de atributos foi realizada (Seção 3.5), e finalmente, o treinamento do modelo de classificação e avaliação (Seção 3.6).

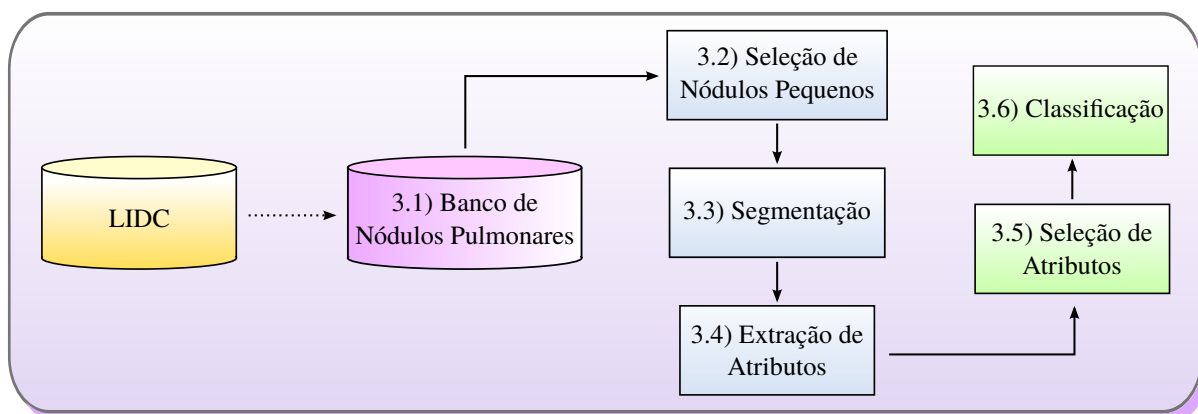


Figura 3.1: Esquema geral da metodologia aplicada nesse trabalho. Fonte: elaborado pelo autor.

Os processos de seleção de atributos e classificação foram realizados usando a ferramenta *RapidMiner Studio* (BOLT; LEONI; AALST, 2016), versão 7.3.000. Os testes foram executados em um PC Intel Core i5, com CPU de 3.10Hz e 8GB RAM com sistema operacional GNU/Linux Ubuntu 14.04 LTS.

3.1 Banco de Nódulos Pulmonares

Neste trabalho usamos o repositório do LIDC (ARMATO et al., 2011) de imagens de TC de câncer de pulmão com lesões identificadas e classificadas por quatro radiologistas em um processo de interpretação de imagem o qual requereu dos especialistas a leitura das TCs e marcação das lesões usando uma interface gráfica. Os nódulos identificados foram classificados pelos radiologistas de acordo com algumas características subjetivas, entre elas a probabilidade de malignidade, de acordo com as seguintes condições:

- Malignidade 1: probabilidade alta para ser benigno;
- Malignidade 2: probabilidade moderada para ser benigno;
- Malignidade 3: probabilidade indeterminada;
- Malignidade 4: probabilidade moderada para ser maligno;
- Malignidade 5: probabilidade alta para ser maligno.

O LIDC é uma coleção não organizada em esquema de banco de dados, assim, não existe correlação entre imagens, dados dos exames e classificação dos nódulos pelos radiologistas. Além disso, o LIDC não contém informações a respeito do tamanho do nódulo ou atributos das imagens. Sendo assim, optaremos por usar o Banco de Nódulos Pulmonares (BNP) organizado por Ferreira Junior et al. (JUNIOR; OLIVEIRA; AZEVEDO-MARQUES, 2016). O BNP usa uma abordagem NoSQL orientado a documentos (STRAUCH, 2011) com o MongoDB (TIWARI, 2011), onde todas as imagens das lesões foram segmentadas usando as marcações realizadas pelos radiologistas e então armazenadas no banco de dados. Isso nos proporcionou fácil acesso as imagens contendo apenas as lesões para etapa de extração de atributos.

O BNP possui 752 exames e 1.944 nódulos pulmonares do LIDC nas cinco classificações de malignidade. Porém, desconsideramos nódulos com probabilidade de malignidade 3, pois possuem probabilidade de malignidade indeterminada, que deixou a nossa base com 1.171 nódulos.

Para este trabalho, nódulos com probabilidades de malignidade 1 e 2 foram considerados benignos, e nódulos com probabilidades de malignidade 4 e 5, malignos.

Devido ao diâmetro dos nódulos escolhido, entre 5 e 10mm, o BNP forneceu uma quantidade de nódulos benignos muito maior do que o de nódulos malignos, o que já era de se esperar devido a maior probabilidade de nódulos pequenos serem benignos (BARTHOLMAI et al., 2015; HUA et al., 2015; Y.-X. J. Wang, J.-S. Gong, K. Suzuki, and S. K. Morcos, 2014). Além disso, devido a complexidade em segmentar o parênquima no entorno de nódulos não sólidos, foi escolhido usar apenas os nódulos sólidos (ver Seção 3.3), o que contribuiu mais ainda para o desbalanceamento do BNP. Com o objetivo de se ter uma base balanceada para o treinamento do modelo de classificação, foram igualadas a quantidade de nódulos benignos e malignos, como pode ser observado na Tabela 3.1.

	Benigno		Maligno		
Probabilidade de Malignidade	1	2	4	5	Total
Número de Nódulos	54	53	87	20	214
Soma	107		107		

Tabela 3.1: Número de nódulos sólidos entre 5-10mm que serão usados do BNP.

3.2 Seleção de Nódulos Pequenos

Antes de selecionarmos os nódulos do BNP em função de seus diâmetros, foi realizado o cálculo dessas medidas manualmente, pois esta informação não constava no BNP. O tamanho de um nódulo pode ser recuperado com uma simples medida 2D do diâmetro de maior valor, que pode ser realizado no plano axial ao longo do eixo do diâmetro mais longo (BARTHOLMAI et al., 2015). Assim, para cada nódulo do BNP, foi calculado o diâmetro para cada fatia como a distância euclidiana entre as coordenadas de mínimo e máximo nos respectivos eixos x e y , a distância de maior valor encontrada foi determinada como o diâmetro do respectivo nódulo (Figura 3.2), e o resultado foi armazenado no BNP e associado ao respectivo nódulos.

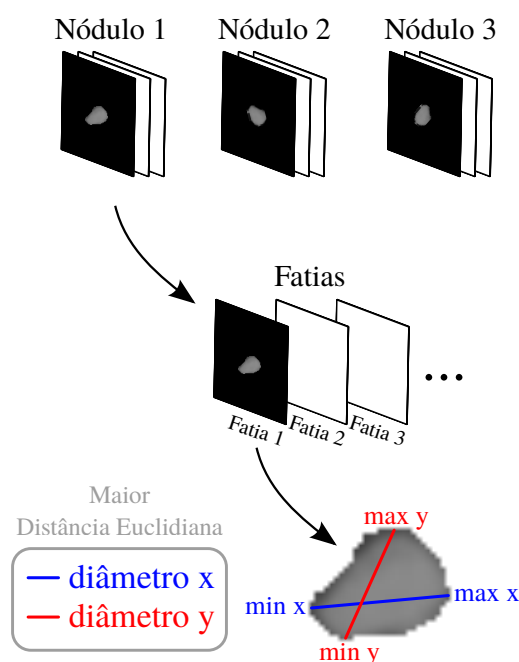


Figura 3.2: Ilustração do processo de cálculo do diâmetro do nódulo. Fonte: elaborado pelo autor.

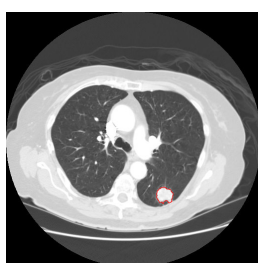
Uma vez que se tenha a informação de diâmetro de todos os nódulos do BNP, foi preciso definir o limiar de tamanho para selecionarmos os nódulos pequenos. Segundo a literatura (YANKELEVITZ et al., 1999; REEVES; KOSTIS, 2000), um nódulo pode ser caracterizado como nódulo pequeno quando possui um diâmetro menor que 10mm. Porém, Bartholmai (BARTHOLMAI et al., 2015) afirma que testes já demonstraram que nódulos com diâmetro menor que 5mm são provavelmente benignos, mesmo para casos de pacientes de alto risco,

o que torna desinteressante trabalharmos com nódulos predispostos a serem benignos. Sendo assim, o limiar escolhido para o trabalho foi 5mm a 10mm.

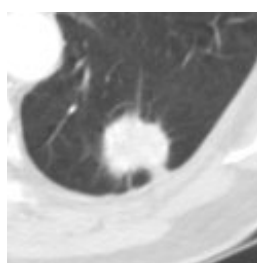
3.3 Segmentação

O objetivo desta etapa foi a obtenção da Região de Interesse (do inglês *Region of Interest* - ROI) da imagem de TC contendo os elementos de interesse deste trabalho, nódulo e parênquima. O BNP possui apenas os nódulos segmentados, logo, foi desenvolvido neste trabalho um algoritmo para segmentar a região do parênquima no entorno do nódulo. Os resultados da segmentação do parênquima foram armazenados no BNP e, então associados aos respectivos nódulos para posterior extração de atributos desta região.

O processo de segmentação foi realizado de forma automática (Figura 3.3), partindo do princípio de que a informação de marcação dos nódulos nas imagens de TC realizadas pelos radiologistas do LIDC consta em nosso BNP (Figura 3.3a). Para cada imagem/fatia, uma ROI foi criada em torno do nódulo usando as coordenadas das marcações feitas pelos radiologistas (Figura 3.3b). Para isto, foi recuperada a informação de máximos e mínimos nas respectivas coordenadas x e y das marcações e a quantidade de parênquima incluída em cada ROI foi proporcional ao tamanho do nódulo, considerando o tamanho de duas vezes o diâmetro do nódulo. Em seguida, foi criada uma máscara contendo a informação de interesse (Figura 3.3c) usando um método de limiarização de imagem com base na minimização das medidas de fuzziness para eliminar nódulo, pleura e vasos respiratórios (HUANG; IUN; WANGT, 1995). Por fim, os resultados (b) e (c) foram combinados para se ter como resultado final uma imagem contendo apenas o tecido do parênquima em torno do nódulo (Figura 3.3d). Esta abordagem foi semelhante à apresentada por Dilger (DILGER et al., 2015), que usou um processo manual para segmentar a região do parênquima.



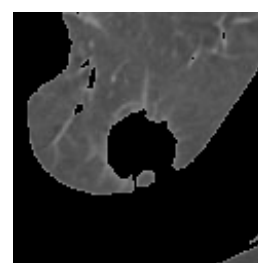
(a) Imagem de TC com um nódulo marcado pelo radiologista.



(b) ROI contendo nódulo, parênquima e outras estruturas.



(c) Máscara do parênquima.



(d) Imagem contendo apenas o parênquima.

Figura 3.3: Ilustração do processo de segmentação do parênquima pulmonar. Fonte: Elaborada pelo autor.

O resultado do processo de segmentação foi submetido a um radiologista especialista em câncer de pulmão, que avaliou o resultado do algoritmo proposto em 20 casos. O especialista

classificou os resultados em péssimo, ruim, moderado, bom ou ótimo (Apêndice A). Devido a complexidade dos nódulos não sólidos que possuem uma intensidade baixa, o especialista sugeriu que o trabalho fosse focado apenas em nódulos sólidos (Seção 2.2), pois, por serem nódulos de alta intensidade, possuem maior contraste em relação a região de parênquima do pulmão, o que facilitou a etapa de limiarização.

3.4 Extração de Atributos

A combinação de diferentes categorias de atributos é bastante comum em aplicações reais para se alcançar uma melhor acurácia na classificação (WU; HE, 2015). Portanto, neste trabalho foram usados quatro categorias de atributos de imagem: Atributos de Intensidade 3D, Forma 3D, Textura 3D e Nitidez de Borda 3D. Tais atributos foram extraídos dos nódulos e do tecido do parênquima pulmonar na forma apresentada na Tabela 3.2 e armazenados no BNP.

	Região Pulmonar	
	Nódulo	Parênquima
Atributos de Intensidade	X	X
Atributos de Forma	X	
Atributos de Textura	X	X
Atributos de Nitidez de Borda	X	

Tabela 3.2: Atributos usados e suas respectivas regiões de extração. Atributos de Nitidez de Borda usam informação tanto do nódulo como do parênquima.

O objetivo de incluir atributos extraídos do parênquima pulmonar no entorno do nódulo, foi aumentar a quantidade de informação disponível além do nódulo. Esta estratégia foi motivada devido ao problema do número/quantidade de *pixels* que os nódulos pequenos abrangem, além disso, Dilger (DILGER et al., 2015) demonstrou que incluir atributos do parênquima pode auxiliar o desenvolvimento de sistemas CAD por conter informações significativas quanto a malignidade dos nódulos.

3.4.1 Análise de Intensidade 3D

Uma vez recuperado a informação de histograma de imagem de cada nódulo, foram extraídos neste trabalho um total de 14 ATI 3D estatísticos sugeridos por Dilger (NOWIK, 2013):

$$\text{Energia} = \sum_i^n x_i^2, \quad (3.1)$$

$$\text{Intensidade média } (\bar{x}) = \frac{1}{n} \sum_{i=1}^n x_i, \quad (3.2)$$

$$\text{Intensidade mediana,} \quad (3.3)$$

$$\text{Intensidade mínima } (I_m), \quad (3.4)$$

$$\text{Intensidade máxima } (I_M), \quad (3.5)$$

$$\text{Entropia} = - \sum_{k=1}^N p(x_k) \log_2(p(x_k)), \quad (3.6)$$

$$\text{Kurtosis} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)s^4}, \quad (3.7)$$

$$\text{Skewness} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)s^3}, \quad (3.8)$$

$$\text{Desvio médio absoluto} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|, \quad (3.9)$$

$$\text{Range} = |I_M - I_m|, \quad (3.10)$$

$$\text{Raiz quadrada média} = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}, \quad (3.11)$$

$$\text{Desvio padrão} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (3.12)$$

$$\text{Uniformidade} = \sum_{k=1}^N p(x_k)^2, \quad (3.13)$$

$$\text{Variância} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (3.14)$$

onde x_i é o i -ésimo valor de intensidade da imagem; s é o desvio padrão da intensidade; n é o número de *pixels* na região; e $p(x_k)$ a probabilidade de ocorrência do k -ésimo valor de intensidade em um conjunto de N intensidades.

Os atributos foram extraídos tanto dos nódulos como da região de parênquima, levando em conta todas as fatias, ou seja, o histograma de cada nódulo foi formado pela união dos histogramas de cada fatia do nódulo, e então realizados os cálculos estatísticos. O mesmo algoritmo foi aplicado no cálculo dos atributos estatísticos para a região do parênquima. Portanto, cada nódulo foi associado a um vetor de ATI com dimensão 28.

3.4.2 Análise de Forma 3D

Neste trabalho foram implementados e adaptados os AF 3D propostos por Aerts (AERTS et al., 2014). Algumas adaptações foram realizadas de forma a simplificar a complexidade de implementação de alguns atributos. Desta forma, os atributos extraídos foram:

$$\text{Compacidade 1} = \frac{V}{\sqrt{\pi A^{\frac{2}{3}}}}, \quad (3.15)$$

$$\text{Compacidade 2} = 36\pi \frac{V^2}{A^3}, \quad (3.16)$$

$$\text{Desproporção esférica} = \frac{A}{4\pi R^2}, \quad (3.17)$$

$$\text{Esfericidade} = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}, \quad (3.18)$$

$$\text{Área (A)} = \sum_{i \in f} p_i, \quad (3.19)$$

$$\text{Área da Superfície} = 4\pi R^2, \quad (3.20)$$

$$\text{Ralação superfície-volume} = \frac{A}{V}, \quad (3.21)$$

$$\text{Volume (V)} = \left(\sum_{f \in F} \sum_{i \in f} p_i \right) \cdot \text{EspessuraDeF}, \quad (3.22)$$

onde f representa o conjunto de *pixels* (p) delimitados pelas coordenadas cartesianas da segmentação do nódulo; F é o conjunto de todas as fatias do nódulo; *EspessuraDeF* corresponde ao valor da espessura das fatias (tamanho do *voxel*); R representa o raio da esfera com mesmo volume que o tumor, definido da forma:

$$\text{Raio (R)} = \sqrt[3]{\frac{3V}{4\pi}}. \quad (3.23)$$

O tamanho do nódulo é um importante indicativo da sua malignância (BRANDMAN SCOTT MD; KO, 2011). Além disso, segundo (DILGER et al., 2015), o atributo tamanho do nódulo tem um papel significativo no diagnóstico de nódulos pulmonares. Logo, foi adicionado o diâmetro do nódulo ao conjunto de AF (cálculo demonstrado na Seção 3.2). Assim, cada nódulo foi associado a um vetor de AF com dimensão 9.

3.4.3 Análise de Textura 3D

Neste trabalho foram usados os AT 3D já desenvolvidos em nosso laboratório. Atributos estes, extraídos dos nódulos (imagens dos nódulos já segmentadas) por Ferreira Junior (JUNIOR; AZEVEDO-MARQUES; OLIVEIRA, 2016) e já armazenados no BNP. (JUNIOR; AZEVEDO-MARQUES; OLIVEIRA, 2016) usou uma matriz de co-ocorrência para obter os AT, a matriz co-ocorrência é uma técnica estatística de segunda ordem, ou seja, obtém informação sobre o posicionamento dos *pixels* da imagem. O trabalho usou alguns dos AT 3D sugeridos por Haralick (HARALICK; SHANMUGAM; DINSTEIN, 1973) (Seção 2.5.3).

O vetor de AT foi obtido pelo cálculo dos nove atributos (Equações 2.1-2.9) aplicados a matriz de co-ocorrência nas orientações 0° , 45° , 90° and 135° , e distância de 1 *voxel*, o que resultou em um conjunto formado por 36 atributos de textura extraídos do nódulo.

Estes mesmos atributos estatísticos e parâmetros foram usados neste trabalho para obtenção de atributos de textura sobre a região do parênquima, onde as imagens resultantes da segmentação da região do parênquima foram dadas como entrada para que o algoritmo realizasse o cálculo dos atributos e armazenasse o vetor de atributos no BNP. Sendo assim, cada nódulo foi associado com um vetor de AT com dimensão 72.

3.4.4 Análise de Nitidez de Borda 3D

Neste trabalho foram usados os ANB 3D já desenvolvidos em nosso laboratório. Tais atributos foram extraídos por Ferreira Junior (JUNIOR; AZEVEDO-MARQUES; OLIVEIRA, 2016) e já armazenados no BNP. (JUNIOR; AZEVEDO-MARQUES; OLIVEIRA, 2016) usou um algoritmo parcialmente proposto por Xu (XU et al., 2012), e realizou a análise de nitidez de borda 3D usando as imagens de TC originais dos exames, adotando a criação de linhas ortogonais a borda do nódulo, para então extrair atributos estatísticos sobre as intensidades dos *pixels* dessas linhas ortogonais em todo volume da imagem (Seção 2.5.4). Portanto, cada nódulo foi caracterizado com um vetor de ANB com dimensão 12.

3.5 Seleção de Atributos

A fim de evitar o problema da maldição de dimensionalidade e selecionar os atributos mais relevantes na classificação de nódulos pulmonares pequenos, foi aplicada uma técnica de seleção de atributos aos conjuntos de atributos denominada Algoritmo Genético Evolutivo (AGE) (ROZSYPAL; KUBAT, 2003) (Seção 2.6.1).

No algoritmo, foi definido um número máximo de 100 gerações, sendo cada geração formada por uma população de 20 indivíduos. Os indivíduos selecionados para reprodução foram escolhidos usando o método de seleção por torneio. Na fase de reprodução, os operados usados foram o cruzamento de um ponto e a mutação com probabilidades de aplicação em cada

indivíduo de 50% e 5%, respectivamente.

3.6 Classificação

Com o objetivo de criar o modelo de classificação para nódulos pequenos, foram usados os algoritmos de aprendizado de máquina: k-NN; um algoritmo de RNA denominado Perceptron de Múltiplas Camadas (*Multilayer Perceptron - MLP*), o qual trata-se de uma RNA *feedforward* treinada com o algoritmo *back-propagation* com função de ativação do tipo sigmoid; e por fim, um algoritmo de Árvore de Decisão denominado *Random Forest*. Estas técnicas de aprendizado de máquina têm sido aplicadas tanto na detecção como na classificação de nódulos pulmonares (NAMIN et al., 2010; KURUVILLA; GUNAVATHI, 2014; TARTAR; KILİÇ; AKAN, 2013).

O modelo foi avaliado com uma validação cruzada *10-fold* com os 214 nódulos pequenos. Quatro conjuntos de atributos foram separadamente usados em cada algoritmo para classificação, aos quais foram dados os nomes: Atributos Textura+Nitidez de Borda (T+NB), Atributos Nódulo (N), Atributos Parênquima (P) e Atributos Integrados (I). A composição dos atributos que cada conjunto possui pode ser vista na Tabela 3.3.

CONJUNTO DE ATRIBUTOS			
T+NB	N	P	I
- Atributos de Textura (Nódulo) - Atributos de Nitidez de Borda	- Atributos de Intensidade - Atributos de Forma - Conjunto T+NB	- Atributos de Intensidade - Atributos de Textura	- Conjunto N - Conjunto P
48 atributos	71 atributos	50 atributos	121 atributos

Tabela 3.3: Composição de cada conjunto de atributos abordado neste trabalho. É importante mencionar que os atributos de intensidade e textura do conjunto N, referem-se aos atributos extraídos do nódulo, e os do conjunto P, aos extraídos do parênquima.

O uso do conjunto T+NB foi motivado por razões comparativas, já que se trata de um conjunto com atributos já extraídos por Ferreira Junior (JUNIOR; OLIVEIRA; AZEVEDO-MARQUES, 2016) e armazenado no BNP.

Em função da proposta de se usar a região do parênquima em torno do nódulo pulmonar para extração de atributos, algumas categorias de atributos foram extraídas de duas regiões diferentes (nódulo e parênquima), como pode ser observado na Seção 3.4. Por esse motivo foram criados os conjuntos N e P.

Por fim, um conjunto com a integração de todas as categorias de atributos justifica-se pelo conceito de *radiomics*, que trata justamente da análise em grande quantidade de atributos quantitativos. É importante ressaltar que neste trabalho a abordagem radiomics ficou restrita ao contexto de atributos de imagens, não foram utilizadas informações como dados genéticos ou clínicos dos pacientes.

Para o k-NN, k variou no intervalo ímpar [3,15]. Para cada valor de k, a métrica de similaridade usada foi a distância euclidiana. Com o MLP, os parâmetros do algoritmo usados foram: 0.3 de taxa de aprendizagem e 0.2 de *momentum* com uma camada oculta. Foi escolhido um valor baixo para taxa de aprendizagem para que ocorressem mudanças pequenas no ajuste de peso em cada etapa, a fim de evitar problemas com a convergência do algoritmo. Cinco testes foram realizados com estes valores de parâmetros, variando a quantidade de ciclos de treinamento do algoritmo para os respectivos valores: 100, 200, 300, 400 e 500 ciclos. O limite de ciclos definido até 500 foi devido ao elevado custo computacional que o algoritmo MLP possui, tornando inviável a utilização de valores de ciclos mais elevados. Por fim, para o RF foram realizados testes com a geração de 500, 750 e 1000 árvores, o critério de seleção adotado foi o ganho de informação, com profundidade máxima de 60. O valor de profundidade máxima foi definido em função dos resultados obtidos com o AGE nos algoritmos k-NN e MLP, onde os melhores resultados com estes algoritmos usaram respectivamente 55 e 65 atributos selecionados de um total de 121 atributos, o valor médio desse quantitativo de atributos selecionados foi definido como a profundidade máxima das árvores.

No algoritmo k-NN, o processo de classificação foi realizado sem e com a seleção de atributos, exatamente com os mesmos parâmetros. Com o algoritmo MLP, o processo de classificação sem seleção de atributos foi realizado usando todos os parâmetros definidos anteriormente. Já para etapa com seleção de atributos, o processo foi repetido apenas para o conjunto de parâmetros do algoritmo que alcançou o melhor desempenho de classificação sem seleção em cada conjunto de atributos. A seleção de atributos não foi aplicada ao algoritmo RF, dado que o mesmo por natureza já se encarrega de ponderar os atributos.

Capítulo 4

Resultados e Discussão

4.1 Resultados da Segmentação

Segundo a avaliação do radiologista especialista em câncer de pulmão, o método usado para segmentação se mostrou eficaz para o caso dos nódulos sólidos. A Figura 4.1 apresenta a avaliação do especialista que validou a segmentação, que classificou os resultados como ótimo/bom em 87,5%, moderado 12,5% e nenhum caso foi classificado como ruim/péssimo. Logo, os resultados foram considerados pelo especialista como satisfatórios.

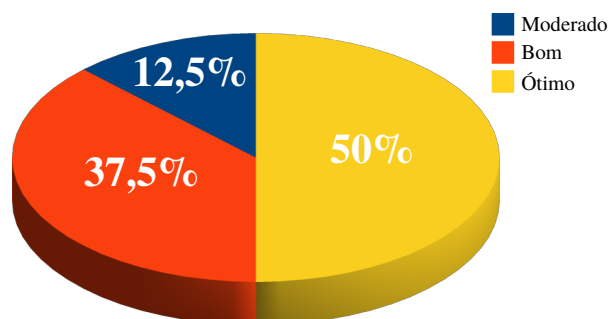


Figura 4.1: Avaliação do resultado de segmentação pelo especialista. Fonte: elaborado pelo autor.

As Figuras 4.2 a 4.5 exibem alguns dos resultados da segmentação da região do parênquima em torno do nódulo. Em cada exemplo é exibido a imagem de TC original com o nódulo identificado (seta vermelha), e abaixo dela, três imagens em fundo azul representando, da esquerda para à direita: a ROI extraída pelo método, a máscara gerada e a região de parênquima segmentada. Algumas imagens de resultado foram ampliadas para uma melhor visualização.

É possível observar que em alguns casos a segmentação não conseguiu remover totalmente os vasos do parênquima, exemplos 1 e 7 das Figuras 4.2 e 4.5, respectivamente.

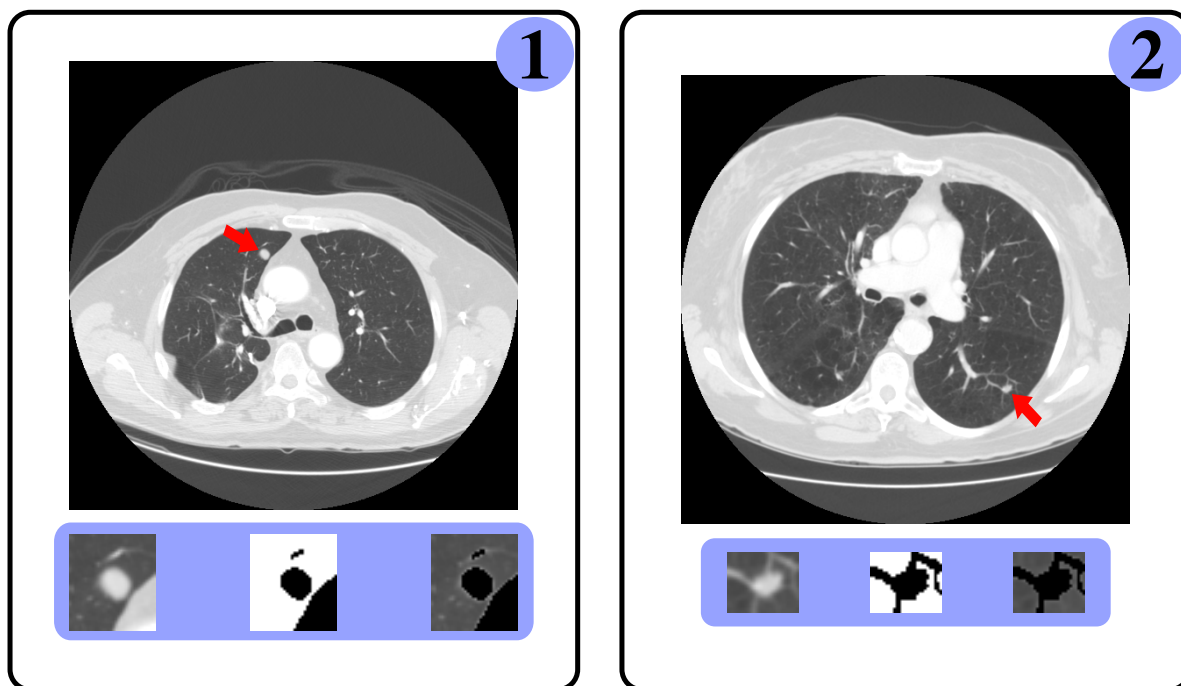


Figura 4.2: Resultados obtidos pela segmentação da região do parênquima em torno do nódulo.
Fonte: elaborado pelo autor.

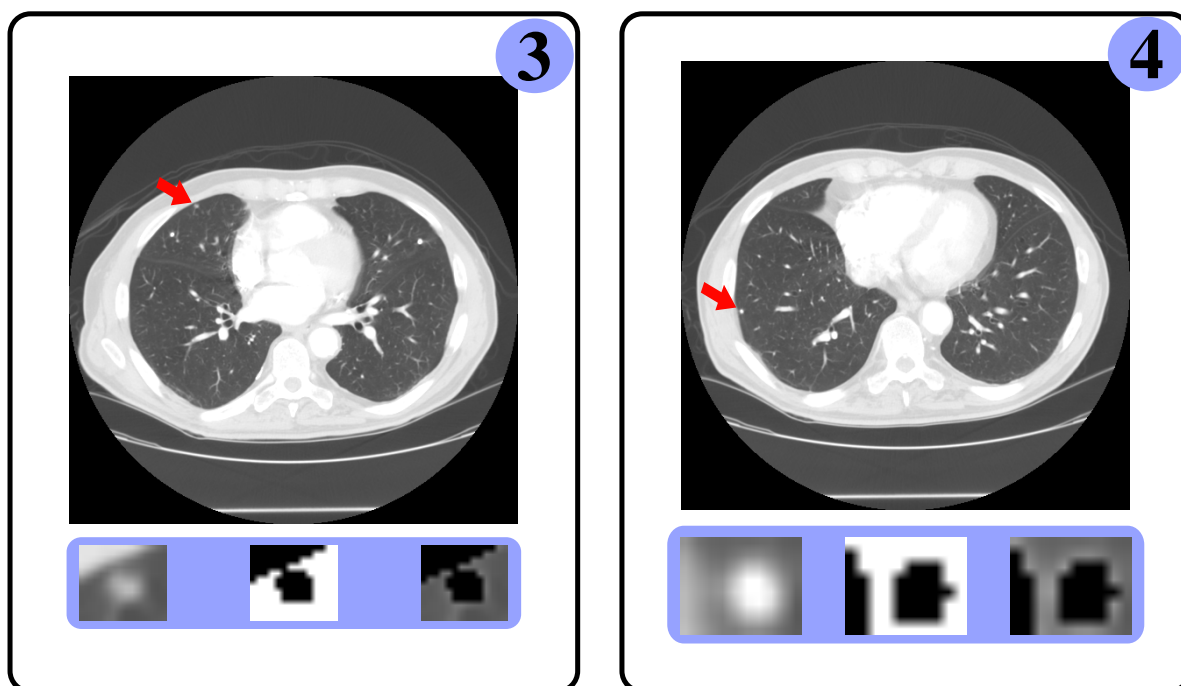


Figura 4.3: Resultados obtidos pela segmentação da região do parênquima em torno do nódulo.
Fonte: elaborado pelo autor.

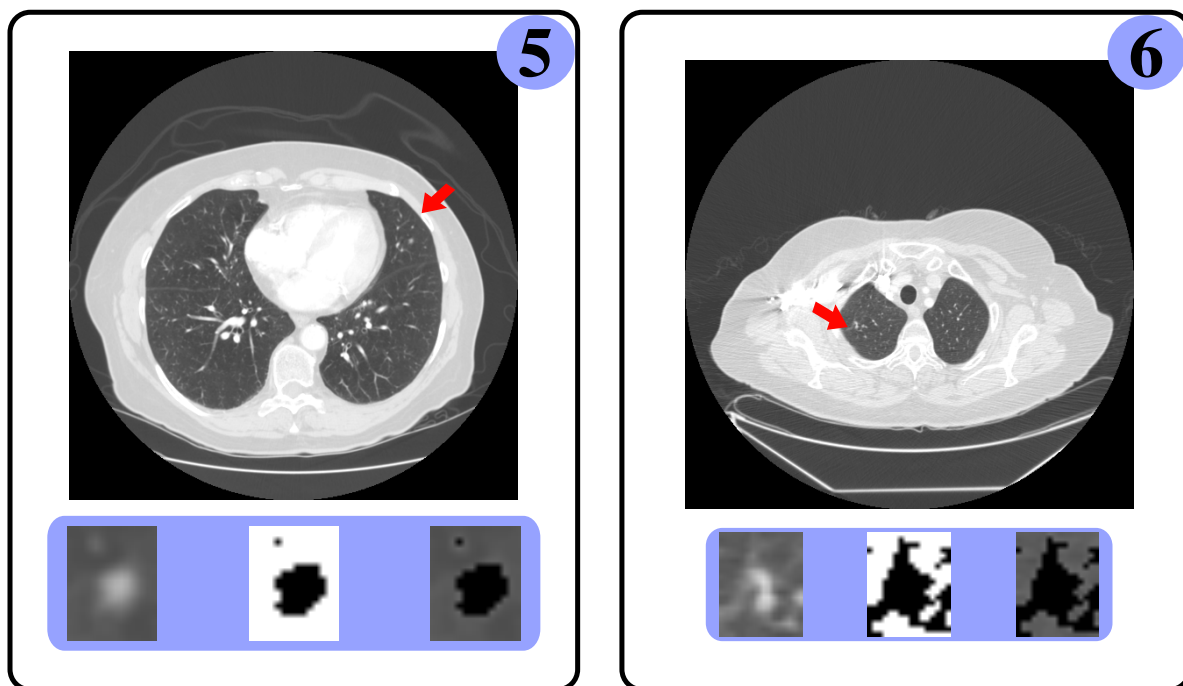


Figura 4.4: Resultados obtidos pela segmentação da região do parênquima em torno do nódulo.
Fonte: elaborado pelo autor.

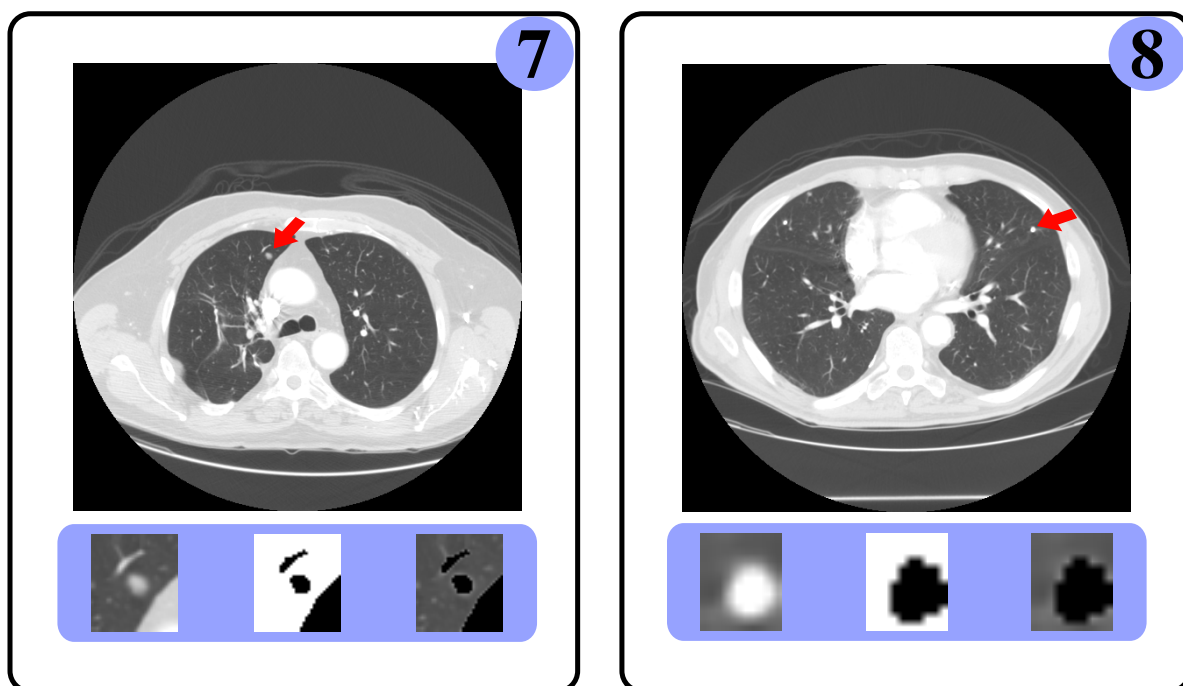


Figura 4.5: Resultados obtidos pela segmentação da região do parênquima em torno do nódulo.
Fonte: elaborado pelo autor.

4.2 Desempenho do Modelo para Classificação

A área sob a curva (AUC) foi usada para avaliar o desempenho dos classificadores em cada conjunto de atributos. As Tabelas 4.1 a 4.5 apresentam os resultados da classificação (média \pm desvio padrão) com uma validação cruzada 10-*fold* de cada algoritmo de aprendizado de máquina. Em cada uma destas tabelas é possível ver o resultado da AUC alcançado em cada conjunto de atributos variando os parâmetros dos respectivos algoritmos de classificação na forma definida na metodologia deste trabalho.

O modelo de classificação usando o algoritmo k-NN alcançou maior AUC média de 0.773 ($\sigma = 0.097$) com o conjunto de atributos I (Tabela 4.1). Aplicando o algoritmo genético para seleção de atributos, o k-NN obteve novamente a melhor performance de AUC com o conjunto de atributos I, com média de 0.847 ($\sigma = 0.088$) (Tabela 4.2). Neste último resultado, o AGE selecionou um total de 55 atributos dos 121 que fazem parte do conjunto I (Tabela 4.6).

AUC :: K-NN				
k	$T+NB$	P	N	I
3	0.636 \pm 0.096	0.633 \pm 0.121	0.693 \pm 0.115	0.683 \pm 0.074
5	0.678 \pm 0.070	0.645 \pm 0.074	0.703 \pm 0.106	0.717 \pm 0.111
7	0.655 \pm 0.061	0.637 \pm 0.087	0.735 \pm 0.120	0.739 \pm 0.094
9	0.666 \pm 0.065	0.641 \pm 0.079	0.751 \pm 0.121	0.773 \pm 0.097
11	0.657 \pm 0.096	0.637 \pm 0.098	0.753 \pm 0.116	0.754 \pm 0.115
13	0.675 \pm 0.083	0.614 \pm 0.116	0.770 \pm 0.106	0.764 \pm 0.110
15	0.671 \pm 0.092	0.605 \pm 0.125	0.764 \pm 0.113	0.766 \pm 0.108

Tabela 4.1: Classificação de nódulos pequenos usando o k-NN.

AUC :: K-NN COM SELEÇÃO DE ATRIBUTOS				
k	$T+NB$	P	N	I
3	0.727 \pm 0.103	0.696 \pm 0.104	0.771 \pm 0.103	0.763 \pm 0.111
5	0.740 \pm 0.101	0.777 \pm 0.094	0.806 \pm 0.088	0.827 \pm 0.079
7	0.766 \pm 0.112	0.746 \pm 0.095	0.820 \pm 0.073	0.826 \pm 0.109
9	0.747 \pm 0.090	0.720 \pm 0.106	0.821 \pm 0.068	0.847 \pm 0.088
11	0.724 \pm 0.094	0.676 \pm 0.084	0.812 \pm 0.094	0.822 \pm 0.087
13	0.727 \pm 0.121	0.697 \pm 0.093	0.829 \pm 0.087	0.839 \pm 0.079
15	0.735 \pm 0.064	0.677 \pm 0.118	0.828 \pm 0.118	0.806 \pm 0.072

Tabela 4.2: Classificação de nódulos pequenos usando o k-NN com atributos selecionados pelo algoritmo genético.

A Figura 4.6 apresenta graficamente o desempenho do algoritmo k-NN com cada conjunto de atributos. O conjunto P possui o pior desempenho, com uma maior AUC média de 0.645 ($\sigma = 0.074$) em $k = 5$, ficando abaixo da curva de desempenho do k-NN com o conjunto T+NB que obteve maior AUC média de 0.678 ($\sigma = 0.070$) em $k = 5$. Apesar do algoritmo ter alcançado maior AUC média de 0.773 ($\sigma = 0.097$) com o conjunto I em $k = 9$, seu desempenho ficou muito próximo com o conjunto N, que obteve maior AUC média de 0.770 ($\sigma = 0.106$) em $k = 13$.

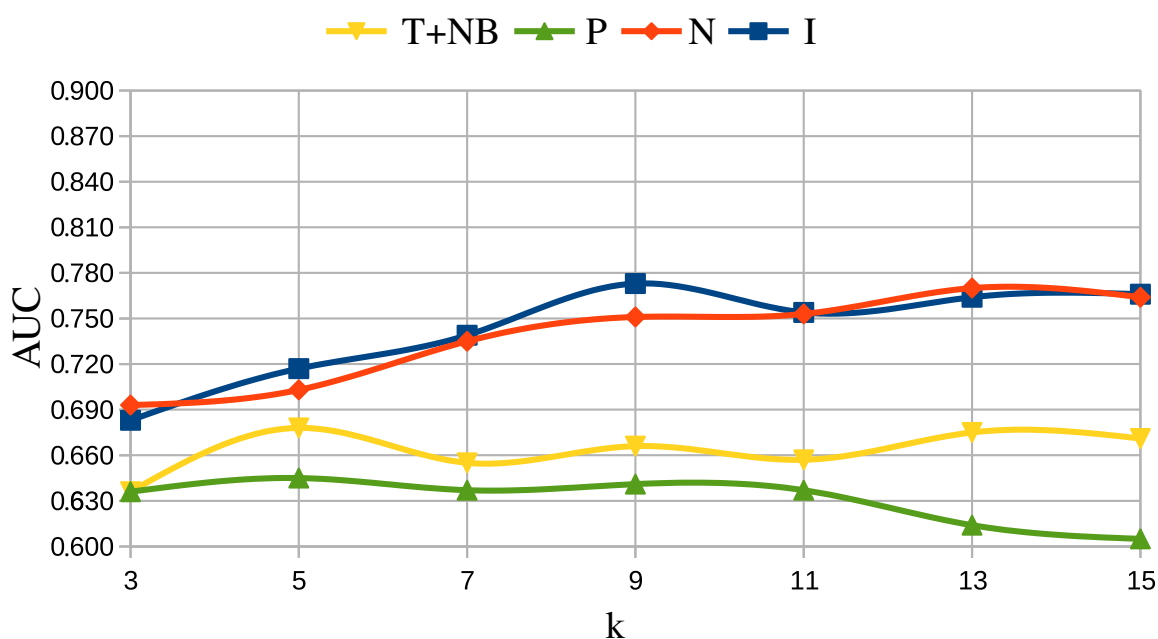


Figura 4.6: Classificação de nódulos pequenos usando o k-NN com variação de k entre 3 a 15. Fonte: elaborado pelo autor.

A Figura 4.7 apresenta graficamente o desempenho do algoritmo k-NN com cada conjunto de atributos fazendo uso do algoritmo de seleção de atributos. Com o conjunto P, o k-NN alcançou uma maior AUC média de 0.777 ($\sigma = 0.094$) em $k = 5$, um valor maior que o melhor resultado com T+NB, que alcançou uma maior AUC média de 0.766 ($\sigma = 0.112$) em $k = 7$, contudo, o algoritmo obteve a menor AUC média de 0.677 ($\sigma = 0.118$) com o conjunto P em $k = 15$. O gráfico confirma a superioridade de desempenho do algoritmo k-NN com o conjunto I em relação ao desempenho com os outros conjuntos, que alcançou uma AUC média de 0.847 ($\sigma = 0.088$) em $k = 9$, ficando a uma diferença positiva de AUC média de 0.018 em relação ao melhor desempenho com o conjunto N, que obteve uma AUC média de 0.829 ($\sigma = 0.087$) em $k = 13$.

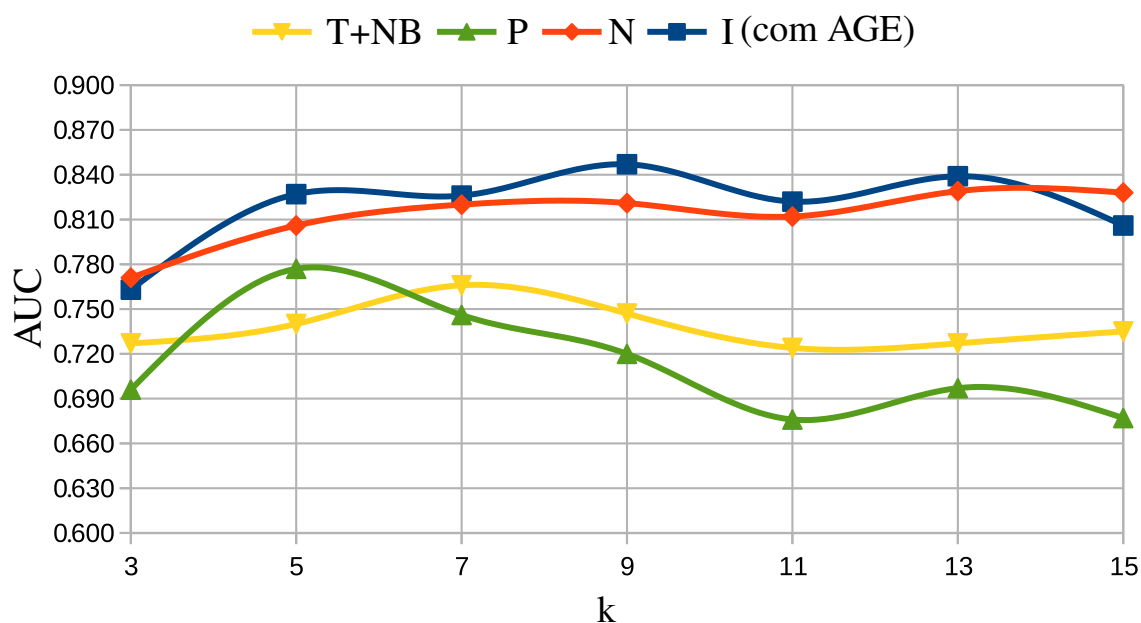


Figura 4.7: Classificação de nódulos pequenos usando o k-NN com atributos selecionados pelo algoritmo genético e variação de k entre 3 e 15. Fonte: elaborado pelo autor.

O modelo de classificação usando o algoritmo RF alcançou maior AUC média de 0.813 ($\sigma = 0.125$) com o conjunto de atributos I (Tabela 4.3).

AUC :: RANDOM FOREST				
Número de Árvores	T+NB	P	N	I
500	0.755 ± 0.108	0.750 ± 0.101	0.785 ± 0.131	0.813 ± 0.125
750	0.756 ± 0.112	0.763 ± 0.092	0.788 ± 0.125	0.812 ± 0.129
1000	0.766 ± 0.105	0.762 ± 0.097	0.778 ± 0.127	0.805 ± 0.122

Tabela 4.3: Classificação de nódulos pequenos usando o RF.

A Figura 4.8 apresenta graficamente o desempenho do algoritmo RF com cada conjunto de atributos. O desempenho do algoritmo RF com os conjuntos T+NB e P foi muito próximo, dado que a diferença absoluta entre seus respectivos maiores valores médios de AUC foi de 0.003, obtidos com um número de árvores igual a 1000 e 750, respectivamente. Com o conjunto N, o algoritmo obteve uma maior AUC média de 0.788 ($\sigma = 0.125$) com 750 árvores, uma diferença negativa de AUC média de -0.025 em relação ao seu desempenho com o conjunto I, que obteve uma maior AUC média de 0.813 ($\sigma = 0.125$) com 500 árvores. Portanto, o gráfico só sustenta a superioridade do algoritmo RF com o conjunto I.

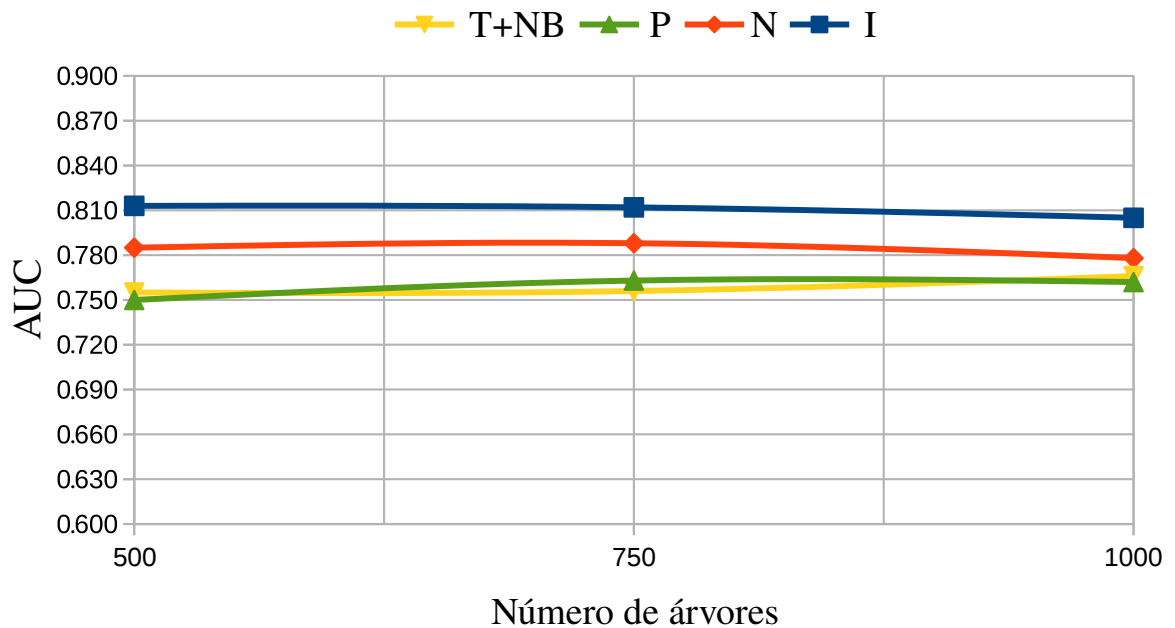


Figura 4.8: Classificação de nódulos pequenos usando o RF. Fonte: elaborado pelo autor.

Com o algoritmo MLP, o modelo alcançou maior AUC média de 0.820 ($\sigma = 0.092$) com o conjunto de atributos I (Tabela 4.4). Aplicando o AGE ao MLP em cada conjunto com uma quantidade de ciclos de treinamento que obtiveram as melhores performances de AUCs, respectivamente, o algoritmo obteve novamente o melhor resultado de AUC com o conjunto de atributos I, com média de 0.875 ($\sigma = 0.048$) (Tabela 4.5), o melhor resultado alcançado neste trabalho. Neste último resultado, o MLP usou um total de 65 atributos dos 121 que compõem o conjunto I (Tabela 4.6).

AUC :: PERCEPTRON DE MÚLTIPLAS CAMADAS				
<i>Ciclos</i>	<i>T+NB</i>	<i>P</i>	<i>N</i>	<i>I</i>
100	0.725 ± 0.119	0.639 ± 0.148	0.817 ± 0.099	0.810 ± 0.081
200	0.726 ± 0.100	0.671 ± 0.180	0.807 ± 0.093	0.817 ± 0.095
300	0.703 ± 0.121	0.679 ± 0.175	0.798 ± 0.098	0.817 ± 0.095
400	0.700 ± 0.110	0.665 ± 0.165	0.775 ± 0.125	0.820 ± 0.092
500	0.713 ± 0.104	0.640 ± 0.152	0.772 ± 0.124	0.810 ± 0.106

Tabela 4.4: Classificação de nódulos pequenos usando o MLP.

AUC :: PERCEPTRON DE MÚLTIPLAS CAMADAS COM SELEÇÃO DE ATRIBUTOS			
<i>T+NB :: 200 ciclos</i>	<i>P :: 300 ciclos</i>	<i>N :: 100 ciclos</i>	<i>I :: 400 ciclos</i>
0.774 ± 0.083	0.721 ± 0.103	0.857 ± 0.083	0.875 ± 0.048

Tabela 4.5: Classificação de nódulos pequenos usando o MLP com atributos selecionados pelo algoritmo genético.

A Figura 4.9 apresenta graficamente o desempenho do algoritmo MLP em cada conjunto de atributos. Usando os atributos do conjunto P, o MLP obteve uma maior AUC média de 0.679 ($\sigma = 0.175$) com 300 ciclos de treino, ficando abaixo do menor valor médio de AUC obtido pelo algoritmo MLP com T+NB, que foi 0.700 ($\sigma = 0.110$) com 400 ciclos. Desta forma, o gráfico sustenta que o MLP obteve o pior desempenho com o conjunto P. Apesar do algoritmo ter alcançado maior AUC média de 0.820 ($\sigma = 0.092$) com o conjunto I com 400 ciclos de treinamento, seu desempenho ficou muito próximo com o conjunto N, que obteve maior AUC média de 0.817 ($\sigma = 0.099$) com 100 ciclos. Porém, com a aplicação do AGE, o MLP obteve a mais alta AUC média deste trabalho com o conjunto I com 400 ciclos, com uma diferença positiva de AUC média de 0,018 em relação ao melhor desempenho com o conjunto N, que obteve uma AUC média de 0.857 ($\sigma = 0.083$) com 100 ciclos de treinamento.

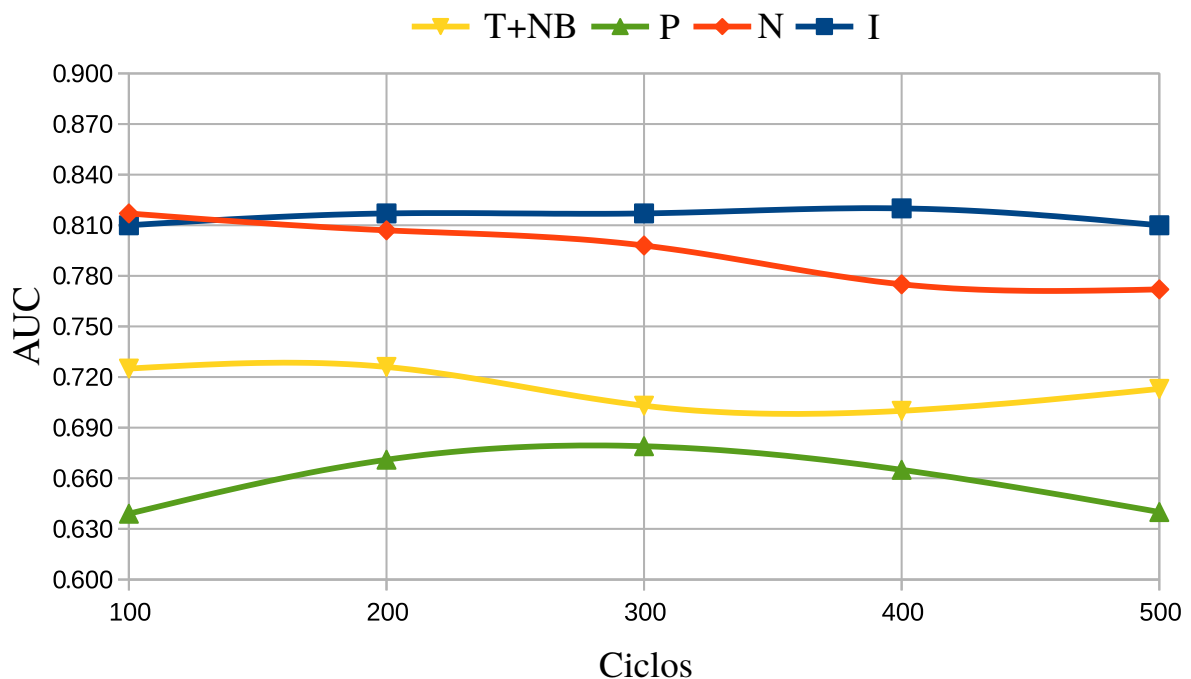


Figura 4.9: Classificação de nódulos pequenos usando o MLP. Fonte: elaborado pelo autor.

A curva ROC exibida na Figura 4.10 apresenta a comparação entre os melhores resultados obtidos pelo modelo em função de cada conjunto de atributo abordado neste trabalho, confirmando a superioridade do modelo de classificação com o conjunto de atributos I em relação

aos outros conjuntos. Outro ponto importante a ser observado é o desempenho semelhante do modelo com o uso dos conjuntos T+NB e P.

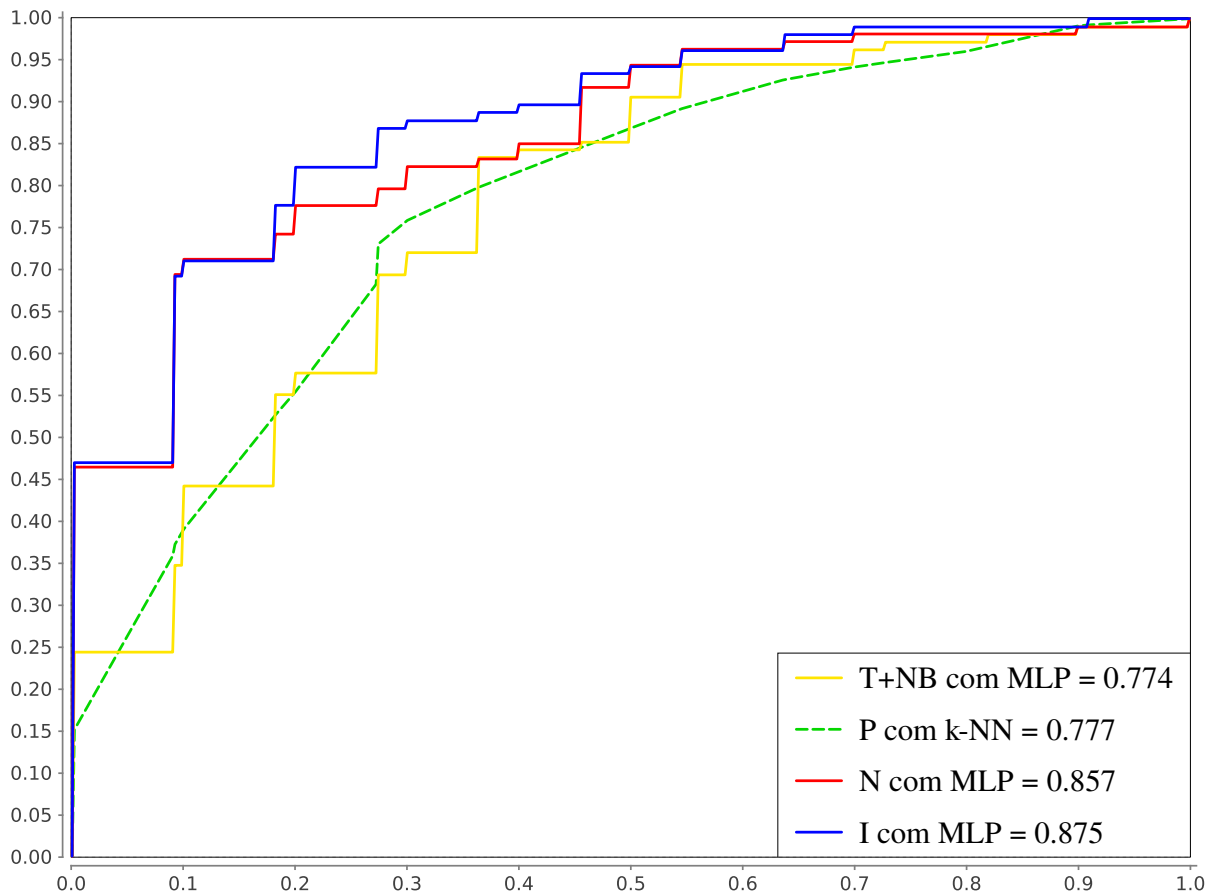


Figura 4.10: Comparação da curva ROC entre os melhores resultados do modelo de classificação em função dos conjuntos de atributos T+NB, P, N e I. Fonte: elaborado pelo autor.

A Tabela 4.6 exibe os atributos selecionados pelo AGE na obtenção dos melhores resultados com os algoritmos k-NN e MLP, resultados estes que foram obtidos com o conjunto de atributos I (Tabelas 4.2 e 4.5). Na tabela é possível observar que todas as categorias de atributos foram selecionadas pelo AGE (intensidade, forma, textura e nitidez de borda), bem como é importante observar que os atributos extraídos da região do parênquima (intensidade e textura) também foram selecionados pelo AGE, e tiveram sua parcela de contribuição na obtenção dos resultados.

RESULTADO DO AGE COM O CONJUNTO I		
ATRIBUTO	ALGORITMO K-NN (AUC = 0.847)	ALGORITMO MLP (AUC = 0.875)
INTENSIDADE NÓDULO	Energia Entropia Intensidades média, mediana e mínima Skewness Uniformidade Variância	Desvio médio absoluto Intensidades mínima, mediana e máxima Range Raiz quadrada média Skewness Desvio padrão
INTENSIDADE PARÊNQUIMA	Desvio médio absoluto Intensidades mediana e mínima Skewness Uniformidade	Energia Entropia Desvio médio absoluto Intensidades mediana e mínima Skewness Uniformidade
FORMA	Compacidade 1 Compacidade 2 Desproporção esférica Esfericidade	Compacidade 1 Desproporção esférica Esfericidade Área Relação superfície-volume Diâmetro
TEXTURA NÓDULO	Energia em 45° e 135° Entropia em 90° MDI em 0° e 135° Contraste em 0°, 45°, 90° e 135° Variância em 90° e 135° Matiz em 0°, 45°, 90° e 135° Proeminência em 0° e 45° Correlação em 90° Homogeneidade em 45° e 135°	Energia em 0° e 135° Entropia em 0° e 45° MDI em 45°, 90° e 135° Contraste em 45°, 90° e 135° Variância em 0°, 45°, 90° e 135° Matiz em 0° e 90° Correlação em 90° e 135° Homogeneidade em 45°, 90° e 135°
TEXTURA PARÊNQUIMA	Energia em 90° Entropia em 45°, 90° e 135° MDI em 45° e 135° Matiz em 0° Proeminência em 45°, 90° Homogeneidade em 0° e 90°	Energia em 0° Entropia em 0° e 45° MDI em 90° Contraste em 0°, 90° e 135° Variância em 45° e 135° Matiz em 45° e 90° Proeminência em 45° e 90° Correlação em 45° e 90° Homogeneidade em 0° e 135°
NITIDEZ DE BORDA	Diferença entre os extremos Soma dos quadrados Soma dos logs Variância da população Variância da amostra Desvio padrão Medida de kurtosis	Diferença entre os extremos Soma dos quadrados Soma dos logs Variação da população Variação da amostra Medida de kurtosis

Tabela 4.6: Atributos selecionados do conjunto I com o AGE nos melhores resultados de classificação obtidos com k-NN e MLP.

4.3 Discussão

Diante dos resultados obtidos neste trabalho com os conjuntos de atributos T+NB e N, o conjunto N foi aquele que obteve melhor desempenho em todos os algoritmos de classificação se comparados aos desempenhos obtidos com T+NB. Este resultado pode ser justificado pela maior representatividade/caracterização do nódulo no conjunto N em relação ao T+NB, já que o primeiro possui justamente mais atributos de categorias diferentes se comparado ao conjunto T+NB, devido a inclusão dos atributos de forma e intensidade.

O modelo para classificação de nódulos pulmonares pequenos desenvolvido neste trabalho obteve seu melhor desempenho com o algoritmo MLP usando os atributos do conjunto I, alcançando uma AUC média de 0.875 ± 0.048 , além disso, todos os outros algoritmos de classificação usados, k-NN e RF, obtiveram os melhores resultados também com este conjunto. O conjunto I é formado justamente pela integração de todas as categorias de atributos extraídos neste trabalho.

Portanto, utilizar descritores *radiomics* de imagem com uma maior variedade de categorias de atributos para caracterizar nódulos pulmonares pequenos, mostrou ser uma estratégia promissora para classificação destes nódulos.

Os algoritmos de classificação obtiveram resultados mais baixos com o conjunto de atributos P, que possui apenas atributos extraídos da região do parênquima no entorno do nódulo. Isto pode ser justificado pelo método de segmentação da região do parênquima adotado neste trabalho, pois nos resultados obtidos é possível observar alguns problemas, como a não eliminação total da pleura, dos vasos e até mesmo do nódulo, estruturas que não são do nosso interesse na segmentação em questão, bem como a inclusão de informação excessiva de parênquima que alguns resultados mostraram devido ao tamanho de algumas ROIs.

A integração dos atributos dos conjuntos P e N, e a consequente formação do conjunto de atributos I, levou o modelo a alcançar os melhores desempenhos nas classificações dos nódulos pulmonares pequenos, como já foi discutido anteriormente. Diante deste argumento, podemos afirmar que a utilização da região do parênquima na caracterização dos nódulos pulmonares contribuiu para uma melhora de desempenho na classificação dos nódulos pulmonares analisados neste trabalho.

Os conjuntos de atributos definidos neste trabalho possuem vetores da ordem de 48 dimensões (conjunto T+NB), chegando até a vetores com dimensão 121 (conjunto I). Com um espaço dimensional desta proporção, a probabilidade de existência de atributos irrelevantes (ruídos) é alta. E diante dos resultados apresentados, é possível observar como este argumento pode ser sustentado, bem como fica evidente a importância da etapa de seleção de atributos em um modelo para classificação, já que com a aplicação do AGE, todos os resultados individualmente obtiveram melhora de desempenho, sem exceção.

Os resultados alcançados com o modelo desenvolvido neste trabalho podem ser comparados com os resultados encontrados na literatura para classificação de nódulos pulmonares. Apesar

disto, é importante ressaltar que os trabalhos encontrados usaram alguns parâmetros diferentes em relação ao modelo apresentado neste trabalho.

Reeves (REEVES; XIE; JIRAPATNAKUL, 2015) usou os algoritmos k-Nearest-Neighbors com distância ponderada (dwNN) e Máquina de Vetores de Suporte (SVM) combinados com um conjunto de 46 atributos de imagem como: atributos geométricos 3D, atributos de distribuição de densidade 3D, atributos de margem e curvatura da superfície, para determinar a malignância de nódulos pulmonares avaliados em diferentes subconjuntos a fim de verificar o impacto da distribuição do tamanho do nódulo não-balanceada¹ na classificação. Com um conjunto de 326 nódulos balanceado por tamanho e limiar de diâmetro 5-14mm, uma AUC média de $0,708 \pm 0,062$ foi alcançada com uma validação cruzada 5-fold. Para o conjunto não-balanceado, com 736 nódulos com diâmetros entre 3-29mm, a classificação obteve uma AUC média de $0,772 \pm 0,031$. Nós obtivemos uma diferença positiva de AUC média de 0.167 em relação ao resultado apresentado por Reeves com a base balanceada, que usou um diâmetro de nódulo próximo do utilizado em nosso trabalho. Vale ressaltar aqui também a diferença entre as quantidades de nódulos, Reeves usou uma base de nódulos diferente, com um quantitativo maior que o nosso, bem como a diferença entre os atributos e algoritmos de classificação.

O trabalho apresentado por Tartar (TARTAR; KILIC; AKAN, 2013), utilizou os algoritmos RNA e RF para classificar 170 lesões como nódulos ou não nódulos com diâmetros entre 2-20mm (média $6,42 \pm 3,00$ mm) com vários métodos de extração e seleção de atributos, tais como Análise de Componentes Principais 2D (do inglês *Principal Component Analysis- PCA*), atributos estatísticos de PCA 2D, processamento de imagem morfológica baseado em atributos geométricos, e o método de seleção de atributos mínima Redundância e máxima Relevância (mRMR). A abordagem de classificação proposta alcançou uma AUC de 0,940 com uma validação cruzada 5-fold. Comparando nosso resultado com o de Tartar, obtivemos uma diferença negativa de AUC média de -0.065. É importante salientar nesta comparação a diferença de tamanho dos nódulos em relação ao nosso trabalho, Tartar utilizou nódulos com até duas vezes o tamanho dos nódulos que usamos em nosso trabalho. Existe também uma diferença entre as categorias de atributos usadas em ambos os trabalhos. Além disso, Tartar não tratou do problema de classificação da lesão quanto a sua malignidade.

Dilger (DILGER et al., 2015) usou uma RNA para classificar entre maligno ou benigno, 50 nódulos com diâmetros entre 4-30mm e um conjunto de 47 atributos selecionados (intensidade, forma, tamanho, textura), extraídos tanto dos nódulos pulmonares como do tecido da região de parênquima do pulmão. O modelo alcançou uma AUC média de $0,935 \pm 0,0096$ com uma validação cruzada *leave-on-out*. Em comparação ao resultado do nosso trabalho, obtivemos uma diferença negativa de AUC média de -0.060. É importante reparar que Dilger também utilizou atributos da região do parênquima, porém, não levou em consideração nódulos pequenos, pois foram usados nódulos com até 30mm em seus testes, o que eliminou alguns desafios inerentes

¹No contexto do trabalho do autor, o termo "distribuição do tamanho do nódulo não-balanceada", refere-se a um conjunto de nódulos nas duas classes, malignos e benignos, com limiares de tamanho de nódulos diferentes para as duas classes.

à utilização de nódulos pequenos.

Nos resultados preliminares deste trabalho (FELIX et al., 2016), foram usados os algoritmos k-NN, MLP e RF para classificar 274 nódulos pulmonares pequenos, com diâmetros entre 3-10mm usando o conjunto de atributos T+NB. O modelo desenvolvido obteve uma AUC média de 0.820 ± 0.053 com uma validação cruzada 10-*fold*. Comparando-o ao resultado atual do nosso trabalho, nós obtivemos uma diferença positiva de AUC média de 0.055, sustentando a importância da utilização de atributos do parênquima na classificação de nódulos pulmonares pequenos. Destacamos também aqui que, em ambos os trabalhos foram usadas a mesma base de nódulos.

Capítulo 5

Conclusão

Este trabalho apresentou o desenvolvimento de um modelo para classificação de nódulos pulmonares pequenos usando atributos *radiomics* extraídos da região de microambiente do nódulo. Com base nos resultados alcançados, demonstramos que a utilização de atributos *radiomics* de imagem é promissora para classificação de nódulos pulmonares pequenos.

Em função dos resultados obtidos neste trabalho, com os algoritmos de aprendizagem e atributos usados, podemos afirmar que o melhor modelo para classificação de nódulos pulmonares pequenos deve usar o algoritmo MLP com um subconjunto de atributos selecionados do conjunto I, já que os melhores resultados de classificação foram alcançados usando este conjunto, com uma AUC média de 0.875 ($\sigma = 0.048$). O subconjunto de atributos com o melhor resultado foi formado pelos atributos:

- Intensidade do nódulo: desvio médio absoluto, intensidades mínima, mediana e máxima, range, raiz quadrada média, skewness e desvio padrão;
- Intensidade do parênquima: energia, entropia, desvio médio absoluto, intensidades mediana e mínima, skewness e uniformidade;
- Forma: compacidade 1, desproporção esférica, esfericidade, área, relação superfície-volume e diâmetro;
- Textura do nódulo: energia em 0° e 135° , entropia em 0° e 45° , MDI em 45° , 90° e 135° , contraste em 45° , 90° e 135° , variância em 0° , 45° , 90° e 135° , matiz em 0° e 90° , correlação em 90° e 135° e homogeneidade em 45° , 90° e 135° ;
- Textura do parênquima: energia em 0° , entropia em 0° e 45° , MDI em 90° , contraste em 0° , 90° e 135° , variância em 45° e 135° , matiz em 45° e 90° , proeminência em 45° e 90° , correlação em 45° e 90° e homogeneidade em 0° e 135° ;
- Nitidez de borda: diferença entre os extremos, soma dos quadrados, soma dos logs, variação da população, variação da amostra e medida de kurtosis.

O conjunto I foi composto justamente por atributos extraídos da região do nódulo (conjunto N), bem como da região de parênquima do pulmão (conjunto P). Isto comprova o teste de hipótese de que considerar o parênquima do pulmão implica em uma melhoria de desempenho na classificação de nódulos pulmonares pequenos, já que o melhor desempenho do modelo com conjunto de atributos N obteve uma AUC média de 0.857 ($\sigma = 0.083$), apresentando uma diferença negativa de AUC média de -0.018 em relação ao melhor resultado com o conjunto I.

Na literatura foi possível encontrar diversos trabalhos que tratam da classificação de nódulos pulmonares. Neles, foram abordados diversos parâmetros que diferem da metodologia adotada neste trabalho, como: diâmetro do nódulo, classificadores e até mesmo descritores diferentes, o que tornou a comparação complexa. Desprezando esta discordância, podemos dizer que o modelo desenvolvido aqui, apesar de ter alcançado resultados melhores que alguns outros trabalhos, ainda obteve um desempenho inferior em relação a literatura. Porém, o trabalho desenvolvido utilizou um cenário muito específico de nódulos pequenos, utilizando um tamanho/diâmetro não comumente trabalhado na literatura no auxílio ao diagnóstico e de extrema importância para a avaliação precoce do câncer de pulmão, deparando-se com a complexidade do diagnóstico de nódulos pequenos e mostrando o potencial que os algoritmos de classificação e os atributos *radiomics* possuem em relação ao difícil problema da classificação de nódulos pulmonares pequenos. Avanços nesta área são importantes, visto que a classificação de nódulos pulmonares pequenos é um grande desafio para o especialista e crítica à sobrevida do paciente.

5.1 Trabalhos Futuros

Com o objetivo de melhorar o modelo de classificação para nódulo pulmonares pequenos desenvolvido neste trabalho, algumas sugestões são propostas como trabalhos futuros:

- Utilizar mais categorias de atributos, para uma melhor caracterização dos nódulos pulmonares;
- Utilizar mais algoritmos de aprendizagem de máquina para classificação;
- Melhorar o algoritmo de segmentação da região do parênquima do pulmão que circunda o nódulo para uma melhor eliminação de ruídos, bem como uma melhoria na quantidade de informação de parênquima a ser adicionada na ROI;

5.2 Contribuições Científicas

Este trabalho foi desenvolvido no Laboratório de Telemedicina e Informática Médica (LaTIM) que está vinculado à Universidade Federal de Alagoas (UFAL). Durante 1 ano e 6 meses o projeto esteve amparado pelo Programa de Pós-Graduação *stricto sensu* da UFAL.

Os resultados parciais deste trabalho foram aceitos em eventos de relevância internacional e nacional com Qualis CAPES durante o período de desenvolvimento da proposta. Os trabalhos científicos aceitos para apresentação foram:

- FELIX, A., OLIVEIRA, M., & RANIERY, J. 2017. Machine learning models for small lung nodules classification using image features. *2017 VI ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing (VipIMAGE)*;
- OLIVEIRA, M., LUCENA, D. & FELIX, A. 2017. Recuperação de Nódulos Pulmonares por Conteúdo: uma abordagem *Radiomics* em Pesquisa Reprodutível. *2017 XVII Workshop de Informática Médica (WIM)*.
- FELIX, A., OLIVEIRA, M., MACHADO, A., & RANIERY, J. 2016 (Oct). Using 3D Texture and Margin Sharpness Features on Classification of Small Pulmonary Nodules. *Pages 394–400 of: 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*;

Referências Bibliográficas

- AERTS, H. J. W. L. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. In: *Nature communications*. [S.l.: s.n.], 2014. v. 5, n. 4006, p. 1–8.
- AGGARWAL, P.; SARDANA, H.; VIG, R. Content based image retrieval approach in creating an effective feature index for lung nodule detection with the inclusion of expert knowledge and proven pathology. In: REVIEWS, C. M. I. (Ed.). [S.l.], 2014. v. 10, p. 178–204.
- AKGÜL, C. B. et al. Content-based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging*, v. 24, n. 2, p. 208–222, 2010. ISSN 1618-727X.
- AL-ABSI, H.; SAMIR, B.; SULAIMAN, S. A computer aided diagnosis system for lung cancer based on statistical and machine learning techniques. *Journal of Computers*, v. 9, n. 2, 2014.
- ALILOU, M. et al. A comprehensive framework for automatic detection of pulmonary nodules in lung ct images. *Image Analysis & Stereology*, v. 33, n. 1, p. 13–27, 2014. ISSN 1854-5165.
- ARMATO, S. G. et al. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, v. 38, n. 2, p. 915–931, 2011.
- AZEVEDO-MARQUES, P. M. *Diagnóstico auxiliado por computador na radiologia*. [S.l.]: Radiologia Brasileira, 2001. v. 34. (285-293, 5).
- BARTHOLMAI, B. J. et al. Pulmonary nodule characterization, including computer analysis and quantitative features. *Journal of Thoracic Imaging*, v. 30, n. 2, p. 139–156, March 2015. ISSN 1076-6332.
- BEDO, M. V. N. et al. Endowing a content-based medical image retrieval system with perceptual similarity using ensemble strategy. *Journal of Digital Imaging*, v. 29, n. 1, p. 22–37, Feb 2016. ISSN 1618-727X.
- BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2007.
- BOLT, A.; LEONI, M. de; AALST, W. M. P. van der. Scientific workflows for process mining: building blocks, scenarios, and implementation. *International Journal on Software Tools for Technology Transfer*, v. 18, n. 6, p. 607–628, Nov 2016. ISSN 1433-2787.
- BRAGA, A. de P.; CARVALHO, A. P. de Leon F. de; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: Rio de Janeiro, RJ: Livros Técnicos e Científicos, 2007.

- BRANDMAN SCOTT MD; KO, J. P. M. Pulmonary nodule detection, characterization, and management with multidetector computed tomography. *Journal of Thoracic Imaging*, v. 26, n. 2, p. 90–105, May 2011.
- BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. ISSN 1573-0565.
- BUSHBERG, J. et al. *The Essential Physics of Medical Imaging*. 3. ed. [S.l.]: Wolters Kluwer Health, 2011. ISBN 9781451153941.
- CUI, J.-W. et al. Screening for lung cancer using low-dose computed tomography: concerns about the application in low-risk individuals. *Translational Lung Cancer Research*, v. 4, n. 3, 2015. ISSN 2226-4477.
- DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM, 2006. (ICML '06), p. 233–240. ISBN 1-59593-383-2.
- DICIOTTI, S. et al. The *LoG* characteristic scale: A consistent measurement of lung nodule size in ct imaging. *IEEE Transactions on Medical Imaging*, v. 29, n. 2, p. 397–409, Feb 2010. ISSN 0278-0062.
- DILGER, S. K. et al. Improved pulmonary nodule classification utilizing lung parenchyma texture features. *Proc. SPIE*, v. 9414, p. 94142T–10, 2015.
- DILGER, S. K.; JUDISCH, A.; UTHOFF, J. e. a. Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. *Journal of Medical Imaging*, v. 2, n. 4, 2015. ISSN 1861-6429.
- DOI, K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, v. 31, n. 4-5, p. 198–211, 2007.
- ERASMUS, J. J. et al. Solitary pulmonary nodules: Part i. morphologic evaluation for differentiation of benign and malignant lesions. *RadioGraphics*, v. 20, n. 1, p. 43–58, 2000. PMID: 10682770.
- ERICKSON, B. J. et al. Machine learning for medical imaging. *RadioGraphics*, v. 37, n. 2, p. 505–515, 2017. PMID: 28212054.
- FACELI, K. et al. *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. [S.l.]: Livros Técnicos e Científicos Editora - LTC, 2011.
- FAWCETT, T. Roc graphs: Notes and practical considerations for researchers. *ReCALL*, Citeseer, v. 31, n. HPL-2003-4, p. 1–38, 2004.
- FELIX, A. et al. Using 3d texture and margin sharpness features on classification of small pulmonary nodules. In: *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2016. p. 394–400.
- GIGER, M. L. Update on the potential of computer-aided diagnosis for breast cancer. *Future Oncology*, v. 6, n. 1, p. 1–4, 2010. PMID: 20021201.
- GILLIES, R. J.; KINAHAN, P. E.; HRICAK, H. Radiomics: Images are more than pictures, they are data. *Radiology*, v. 278, n. 2, p. 563–577, 2016. PMID: 26579733.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2008. ISBN 013168728X.

GONZALEZ, R. C.; WOODS, R. E. *Processamento Digital De Imagens*. 3. ed. [S.l.]: Pearson Education, 2010.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, n. 6, p. 610–621, Nov 1973. ISSN 0018-9472.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999. ISBN 0132733501.

HENSCHKE, C. I. et al. Early lung cancer action project: overall design and findings from baseline screening. *The Lancet*, v. 354, n. 9173, p. 99 – 105, 1999. ISSN 0140-6736.

HOLLAND, J. *Adaptation in natural and artificial systems*. 2. ed. [S.l.]: Ann Arbor, 1992. ISBN 0132733501.

HOWLADER, N. et al. *SEER Cancer Statistics Review, 1975-2013*. 2016. National Cancer Institute. Bethesda, MD. <https://seer.cancer.gov/csr/1975_2013/>. Acessado em 14-07-2017.

HU, Q.; YU, D.; XIE, Z. Neighborhood classifiers. *Expert Systems with Applications*, v. 34, n. 2, p. 866 – 876, 2008. ISSN 0957-4174.

HUA, K.-L. et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and Therapy*, 8:2015–2022, 2015.

HUANG, L.-K.; IUN, M. j; WANGT, J. Image thresholding by minimizing the measure of fuzziness. In: *Pattern Recognition, 28(1):41-51. 113 International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B2. Beijing 2008*. [S.l.: s.n.], 1995.

INCA. *Instituto Nacional de Câncer*, <<http://www.inca.gov.br/>>. 2016. Acessado em 22-09-2016.

INCA. *Tipos de Câncer*, <<http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao>>. 2017. Acessado em 10-07-2017.

J. R. Parker. *Algorithms for Image Processing and Computer Vision*. [S.l.]: Wiley Publishing, Inc, 1997.

JEMAL, A. et al. *The Cancer Atlas*. Second. [S.l.]: American Cancer Society, 2014.

JUNIOR, J. R. F.; AZEVEDO-MARQUES, P. M. de; OLIVEIRA, M. C. Selecting relevant 3d image features of margin sharpness and texture for lung nodule retrieval. *International Journal of Computer Assisted Radiology and Surgery*, p. 1–9, 2016. ISSN 1861-6429.

JUNIOR, J. R. F.; OLIVEIRA, M. C. Evaluating margin sharpness analysis on similar pulmonary nodule retrieval. In: *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. [S.l.: s.n.], 2015. p. 60–65. ISSN 1063-7125.

- JUNIOR, J. R. F.; OLIVEIRA, M. C.; AZEVEDO-MARQUES, P. M. de. Cloud-based nosql open database of pulmonary nodules for computer-aided lung cancer diagnosis and reproducible research. *Journal of Digital Imaging*, p. 1–14, 2016. ISSN 1618-727X.
- KUMAR, D. et al. Discovery radiomics for computed tomography cancer detection. *CoRR*, abs/1509.00117, 2015.
- KURUVILLA, J.; GUNAVATHI, K. Lung cancer classification using neural networks for {CT} images. *Computer Methods and Programs in Biomedicine*, v. 113, n. 1, p. 202 – 209, 2014. ISSN 0169-2607.
- LEVMAN, J. E.; MARTEL, A. L. A margin sharpness measurement for the diagnosis of breast cancer from magnetic resonance imaging examinations. *Academic Radiology*, v. 18, n. 12, p. 1577 – 1581, 2011. ISSN 1076-6332.
- LIANG, M. et al. Low-dose ct screening for lung cancer: Computer-aided detection of missed lung cancers. *Radiology*, v. 281, n. 1, p. 279–288, 2016. PMID: 27019363.
- LIMA, L. L. de; JUNIOR, J. R. F.; OLIVEIRA, M. C. Performance comparison of medical image similarity measures in a heterogeneous architecture. *XIV Congresso Brasileiro de Informática em Saúde (CBIS)*, 2014. ISSN 1618-727X.
- LU, Y. et al. A modality synthesis framework: Using patch based intensity histogram and weber local descriptor features. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. [S.l.: s.n.], 2015. p. 1126–1129. ISSN 1945-7928.
- MAHMOUD-GHONEIM, D. et al. Three dimensional texture analysis in mri: a preliminary evaluation in gliomas. *Magnetic Resonance Imaging*, v. 21, n. 9, p. 983–987, 2003.
- MASON, J. E. et al. An n-gram based approach to multi-labeled web page genre classification. *2014 47th Hawaii International Conference on System Sciences*, IEEE Computer Society, Los Alamitos, CA, USA, v. 0, p. 1–10, 2010.
- MITCHELL, T. M. *Machine Learning*. 1. ed. [S.l.]: McGraw-Hill Education, 1997.
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos de aprendizado de máquina. In: _____. [S.l.]: Manole, 2003. p. 89–114. ISBN 9788520416839.
- NAMIN, S. T. et al. Automated detection and classification of pulmonary nodules in 3d thoracic ct images. In: *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*. [S.l.: s.n.], 2010. p. 3774–3779. ISSN 1062-922X.
- NITHILA, E. E.; KUMAR, S. Automatic detection of solitary pulmonary nodules using swarm intelligence optimized neural networks on {CT} images. *Engineering Science and Technology, an International Journal*, 2016. ISSN 2215-0986.
- NOVAES, F. T. et al. Câncer de pulmão: histologia, estágio, tratamento e sobrevida. *Jornal Brasileiro de Pneumologia*, scielo, v. 34, p. 595 – 600, 08 2008. ISSN 1806-3713.
- NOWIK, S. D. K. *The use of surrounding lung parenchyma for the automated classification of pulmonary nodules*. Tese (MS (Master of Science) thesis) — University of Iowa, 2013.

- NUZHAYYA, T. et al. Classification of texture patterns in ct lung imaging. *SPIE Medical Imaging*, v. 7963, 2011.
- OLIVEIRA, M. C. *Visualização de volumes em morfometria de estruturas cerebrais a partir de imagens de ressonância magnética nuclear*. Tese (Doutorado) — Universidade de São Paulo (USP), 2002.
- OLIVEIRA, M. C. *Grids Computacionais para recuperação de imagens médicas a partir de conteúdo: um estudo de viabilidade*. Tese (Tese de Doutorado) — Universidade de São Paulo (USP), 2006.
- OLIVEIRA, M. C.; CIRNE, W.; MARQUES, P. M. de A. Towards applying content-based image retrieval in the clinical routine. *Future Generation Computer Systems*, v. 23, n. 3, p. 466–474, 2007. ISSN 0167-739X.
- OLIVEIRA, M. C.; FERREIRA, J. R. A bag-of-tasks approach to speed up the lung nodules retrieval in the bigdata age. In: *e-Health Networking, Applications Services (Healthcom), 2013 IEEE 15th International Conference on*. [S.l.: s.n.], 2013. p. 632–636.
- PATZ, E. et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Internal Medicine*, v. 174, n. 2, p. 269–274, 2014.
- RAMLI, S. et al. Histogram of intensity feature extraction for automatic plastic bottle recycling system using machine vision. *American Journal of Environmental Sciences, Science Publications*, v. 4, n. 6, p. 583–588, 2008. ISSN 1553-345X.
- REEVES, A. P. et al. On measuring the change in size of pulmonary nodules. *IEEE Transactions on Medical Imaging*, v. 25, n. 4, p. 435–450, April 2006. ISSN 0278-0062.
- REEVES, A. P.; KOSTIS, W. J. Computer-aided diagnosis of small pulmonary nodules. *Seminars in Ultrasound, CT and MRI*, v. 21, n. 2, p. 116 – 128, 2000. ISSN 0887-2171.
- REEVES, A. P.; XIE, Y.; JIRAPATNAKUL, A. Automated pulmonary nodule ct image characterization in lung cancer screening. *International Journal of Computer Assisted Radiology and Surgery*, v. 11, n. 1, p. 73–88, 2015. ISSN 1861-6429.
- REZENDE, S. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Manole, 2003. ISBN 9788520416839.
- ROZSYPAL, A.; KUBAT, M. Selecting representative examples and attributes by a genetic algorithm. *Intell. Data Anal.*, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 7, n. 4, p. 291–304, set. 2003. ISSN 1088-467X.
- RUMELHART, J. L. M. D. E. *Parallel Distributed Processing*. [S.l.]: MIT Press, 1986. v. 1: Foundations. ISBN 1107057132, 9781107057135.
- SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press, 2014. ISBN 1107057132, 9781107057135.
- SHEWAYE, T. N.; MEKONNEN, A. A. Benign-malignant lung nodule classification with geometric and appearance histogram features. *CoRR*, abs/1605.08350, 2016.

- SIEGLE, R. L. et al. Rates of disagreement in imaging interpretation in a group of community hospitals. *Academic Radiology*, v. 5, n. 3, p. 148 – 154, 1998. ISSN 1076-6332.
- SILVA, A. C.; CARVALHO, P. C. P.; GATTASS, M. Diagnosis of lung nodule using semivariogram and geometric measures in computerized tomography images. *Computer Methods and Programs in Biomedicine*, v. 79, p. 31–38, 2005.
- SILVA, M. P. da. *Processing similarity queries in medical images to the perceptual recovery guided by the user*. Tese (Doutorado) — University of São Paulo (USP), 2009.
- STRAUCH, C. *NoSQL databases*. [S.l.]: Stuttgart Media University, 2011.
- SUZUKI, K. Machine learning in computer-aided diagnosis of the thorax and colon in ct: A survey. *IEICE Transactions*, v. 96-D, p. 772–783, 2013.
- TAKASHIMA, S. et al. Indeterminate solitary pulmonary nodules revealed at population-based ct screening of the lung: using first follow-up diagnostic ct to differentiate benign and malignant lesions. *American Journal of Roentgenology*, v. 180, n. 5, p. 1255–1263, 2003.
- TARTAR, A.; KİLİC, N.; AKAN, A. Classification of pulmonary nodules by using hybrid features. *Computational and Mathematical Methods in Medicine*, v. 2013, 2013.
- TARTAR, A.; KİLİC, N.; AKAN, A. A new method for pulmonary nodule detection using decision trees. In: *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. [S.l.: s.n.], 2013. p. 7355–7359. ISSN 1557-170X.
- TIWARI, S. *Professional NoSQL*. [S.l.]: John Wiley and Sons, Inc, 2011.
- W. J. Choi and T. S. Choi. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Computer Methods and Programs in Biomedicine*, v. 113, n. 1, p. 37–54, 2014. ISSN 1861-6429.
- WANG, S.; RONALD, M. *Machine Learning and Radiology*. 2012.
- WERNICK, M. N. et al. Machine learning in medical imaging. *IEEE Signal Processing Magazine*, v. 27, n. 4, p. 25–38, 2010.
- WU, H.; HE, L. Combining visual and textual features for medical image modality classification with ℓ_p -norm multiple kernel learning. *Neurocomputing*, v. 147, p. 387 – 394, 2015. ISSN 0925-2312. Advances in Self-Organizing Maps Subtitle of the special issue: Selected Papers from the Workshop on Self-Organizing Maps 2012 (WSOM 2012).
- WU, H. et al. Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography. *Journal of Digital Imaging*, v. 26, n. 4, p. 797–802, Aug 2013. ISSN 1618-727X.
- WU, J. et al. Early-stage non-small cell lung cancer: Quantitative imaging characteristics of 18f fluorodeoxyglucose pet/ct allow prediction of distant metastasis. *Radiology*, v. 281, n. 1, p. 270–278, 2016. PMID: 27046074.
- XU, J. et al. Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval. *Medical Physics*, v. 39, n. 9, p. 5405–5418, 2012.

Y.-X. J. Wang, J.-S. Gong, K. Suzuki, and S. K. Morcos. Evidence based imaging strategies for solitary pulmonary nodule. *Journal of Thoracic Disease*, v. 6, n. 7, p. 872, 2014.

YANKELEVITZ, D. F. et al. Small pulmonary nodules: Evaluation with repeat ct—preliminary experience. *Radiology*, v. 212, n. 2, p. 561–566, 1999. PMID: 10429718.

YIP, S. S. F.; AERTS, H. J. W. L. Applications and limitations of radiomics. *Physics in Medicine and Biology*, v. 61, n. 13, p. R150, 2016.

ZHANG, L. et al. ibex: An open infrastructure software platform to facilitate collaborative work in radiomics. *Medical Physics*, v. 42, n. 3, p. 1341–1353, 2015.

Apêndice A

Documento Modelo para Validação da Segmentação

O modelo do documento elaborado com alguns dos resultados da etapa de segmentação da região do parênquima enviado ao médico radiologista, encontra-se nas páginas seguintes.

1. Banco de Nódulos Pulmonares

Os nódulos analisados foram extraídos de um repositório de imagens médicas denominado Lung Image Database Consortium (LIDC), o qual consiste de imagens de Tomografia Computadorizada (TC) do pulmão com lesões identificadas e classificadas por quatro radiologistas, em um processo de interpretação de imagem o qual os especialistas leram as imagens de TCs e marcaram as lesões usando uma interface gráfica.

Os nódulos identificados foram classificados pelos radiologistas de acordo com características subjetivas, dentre elas, vale destacar as duas de maior interesse para nós: malignância e textura (*malignancy* e *texture*, originalmente no banco).

À malignância, foram atribuídos valores inteiros entre 1 e 5, seguindo as seguintes condições:

- 1 – probabilidade alta para ser benigno;
- 2 – probabilidade moderada para ser benigno;
- 3 – probabilidade indeterminada;
- 4 – probabilidade moderada para ser maligno;
- 5 – probabilidade alta para ser maligno.

Quanto a textura, foram atribuídos 3 termos, seguindo as seguintes condições:

- 1 – nódulo não sólido/vidro fosco;
- 3 – nódulo parcialmente sólido;
- 5 – nódulo sólido;

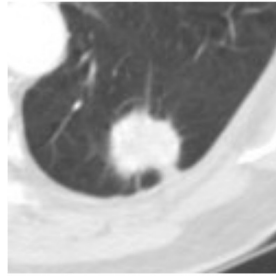
2. Segmentação Parênquima

A necessidade da extração da região do parênquima em torno dos nódulos vem de que, em meu trabalho, extrairei descritores quantitativos dessa região, a fim de usá-los em um modelo de classificação benigno e maligno. Para isso, é necessário a geração de uma Região de Interesse (Region of Interest - ROI) contendo apenas a região do parênquima e excluir qualquer outra informação (ruído) como: nódulo, pleuro e vasos respiratórios. **É importante mencionar que a exclusão total dos ruídos não é o que pretendemos alcançar (esse seria um caso ideal), mas sim uma exclusão subjetivamente significativa, que já seria suficiente (como a exibida na Figura 1d).**

O objetivo final do modelo é entregar ao usuário (médico ou radiologista) uma classificação quanto a malignidade do nódulo com uma probabilidade de acerto, para que o mesmo tenha um suporte na sua tomada de decisão quanto ao diagnóstico final do paciente, principalmente nos casos em que o grau de incerteza quanto ao diagnóstico for alto.



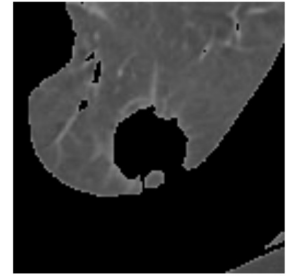
(a) Imagem de TC com um nódulo marcado pelo radiologista.



(b) ROI contendo nódulo, parênquima e outras estruturas.



(c) Máscara do parênquima.



(d) Imagem contendo apenas o parênquima.

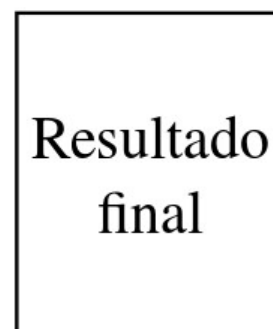
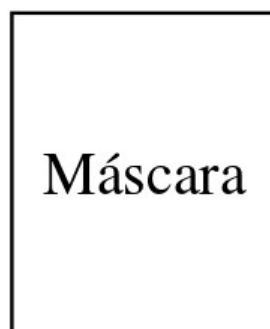
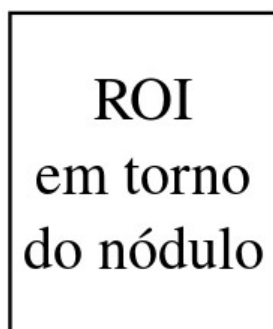
Figura 1: Ilustração do processo de segmentação do parênquima pulmonar.

→ Metodologia de Segmentação

O processo de segmentação foi realizado de forma automática (**Figura 1**), partindo do princípio de que a informação de marcação dos nódulos nas imagens de TC já foi recuperada (**Figura 1a**). Primeiro, uma Região de Interesse (Region of Interest - ROI) foi criada em torno do nódulo usando as coordenadas das marcações feitas pelos radiologistas nas imagens de TC (**Figura 1b**). Para isto, foi recuperada a informação de máximos e mínimos nas respectivas coordenadas x e y das marcações e a quantidade de parênquima incluída em cada ROI foi proporcional ao tamanho do nódulo (duas vezes o tamanho do nódulo). Em seguida, foi criada uma máscara contendo a informação de interesse (**Figura 1c**) usando um método de limiarização de imagem para eliminar nódulo, pleura e vasos respiratórios. Por fim, os resultados (b) e (c) foram combinados para se ter como resultado final uma imagem contendo apenas o tecido do parênquima em torno do nódulo (**Figura 1d**).

Avaliação

Os resultados da segmentação da região do parênquima mostrados a seguir foram gerados a partir de alguns dos nódulos sólidos, ou seja, nódulos com atributo textura no valor 5. A figura abaixo mostra o esquema que as imagens para avaliação possuirão, bem como, o ranking de avaliação para o resultado final.



Péssimo

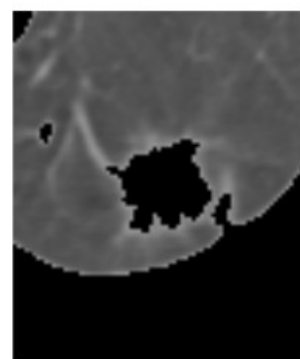
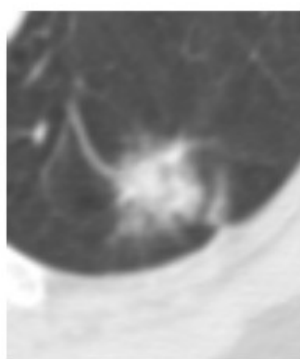
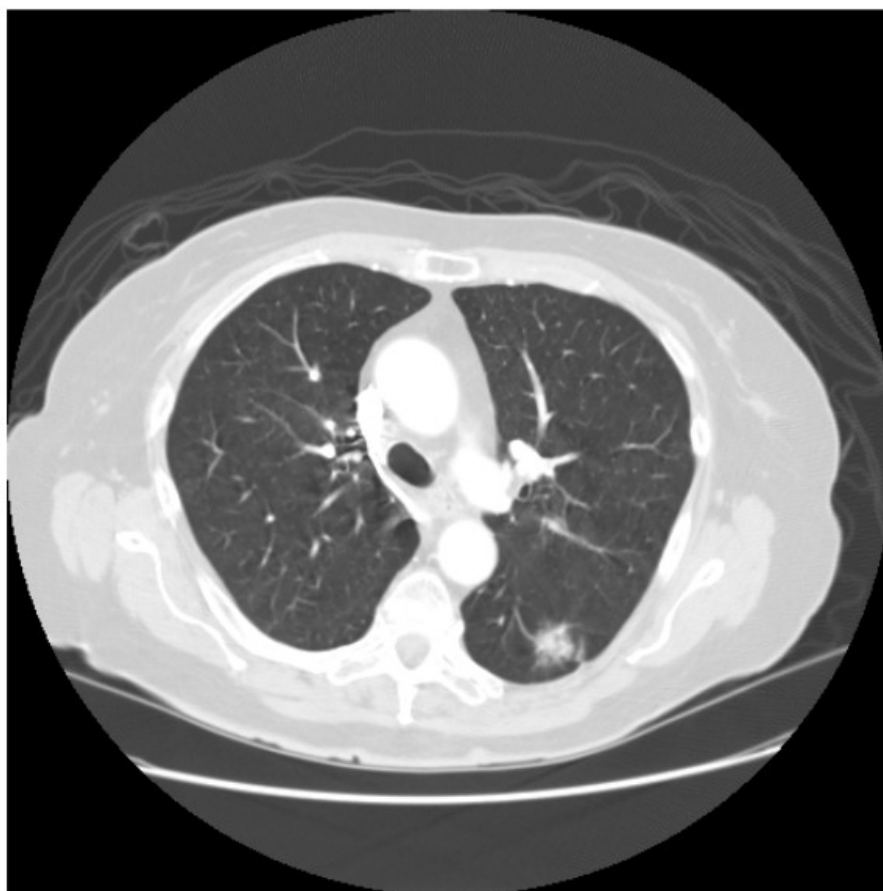
Ruim

Moderado

Bom

Ótimo

Resultado 1



Péssimo

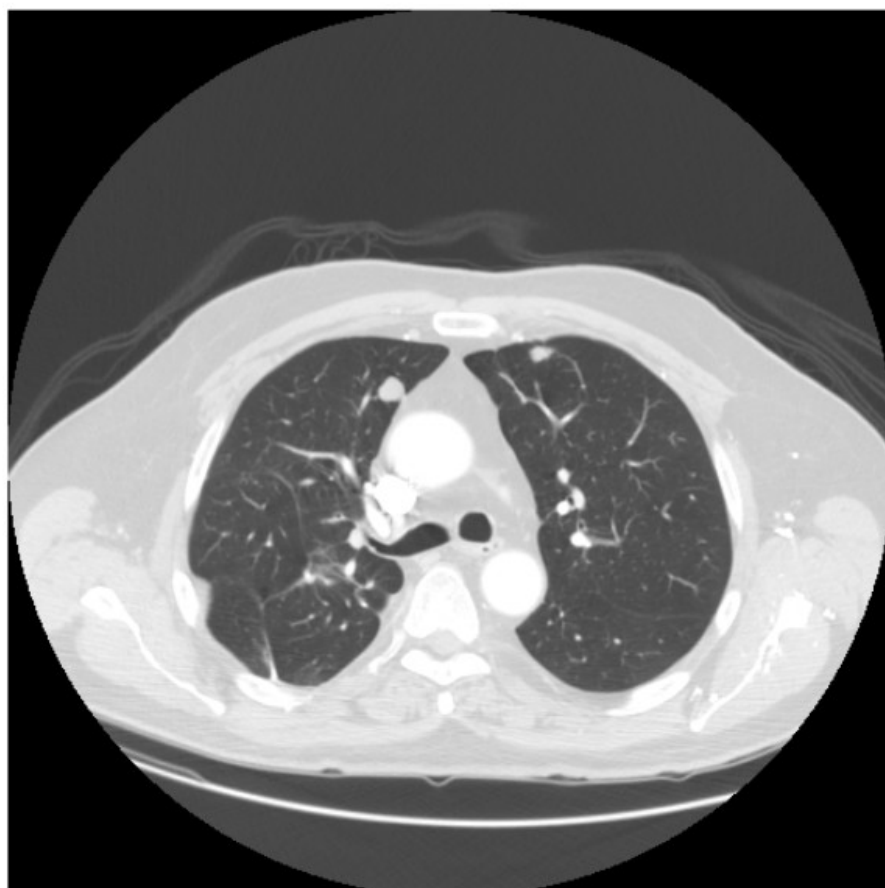
Ruim

Moderado

Bom

Ótimo

Resultado 2



Péssimo

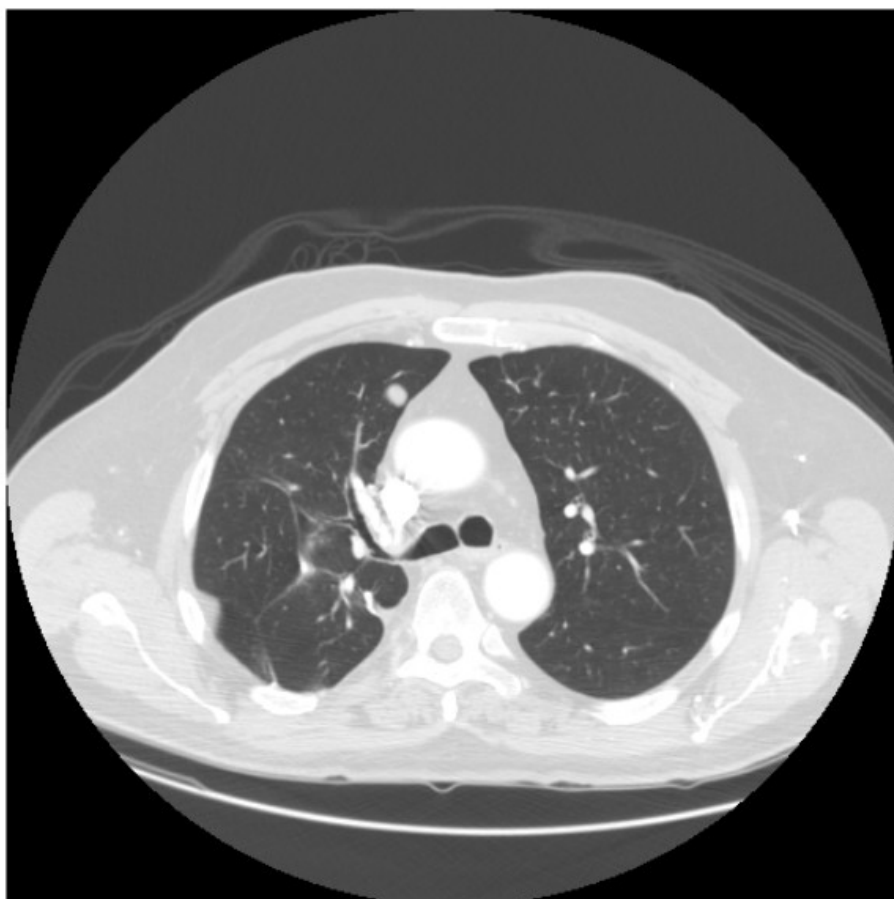
Ruim

Moderado

Bom

Ótimo

Resultado 3



Péssimo

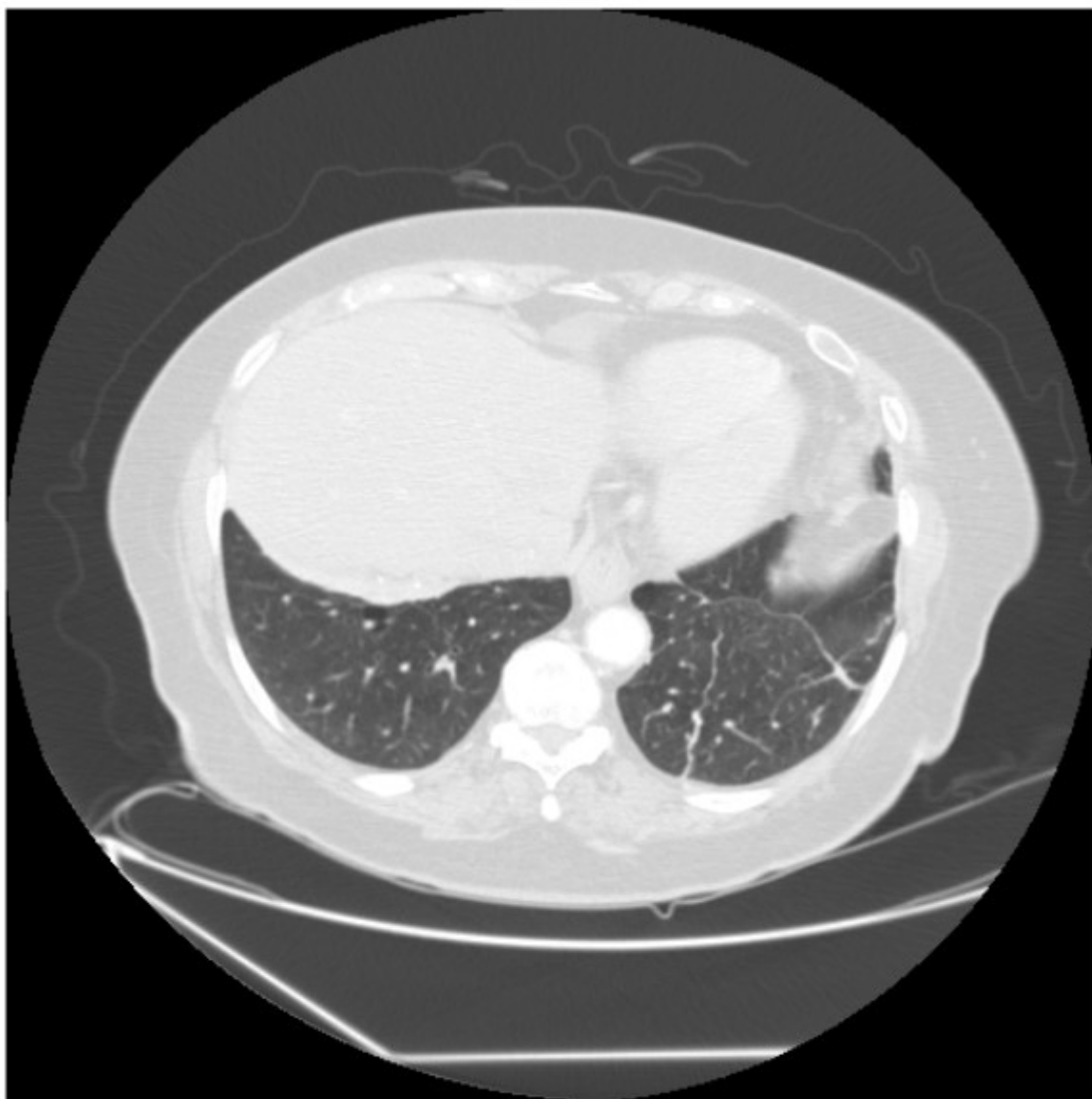
Ruim

Moderado

Bom

Ótimo

Resultado 4



não tenho certeza que há nódulo nesta imagem ... pode ser vaso

Péssimo

Ruim

Moderado

Bom

Ótimo

Resultado 5



Péssimo

Ruim

Moderado

Bom

Ótimo

Resultado 6



este é um linfonodo intrapulmonar, não é nódulo

Péssimo

Ruim

Moderado

Bom

Ótimo

Resultado 7



mesma situação do caso anterior ... os linfonodos são comuns e certamente benignos

Péssimo

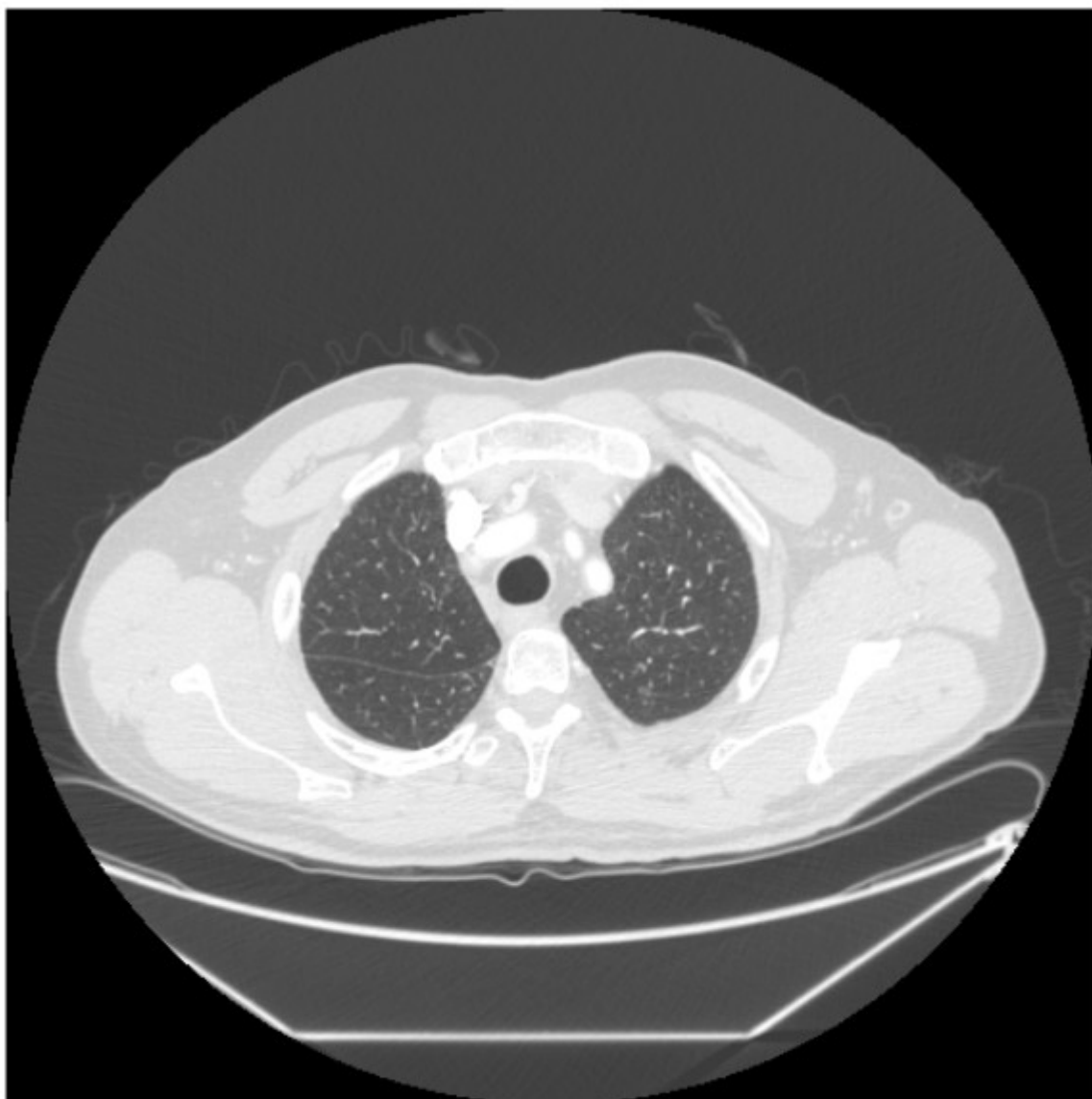
Ruim

Moderado

Bom

Ótimo

Resultado 8



Péssimo

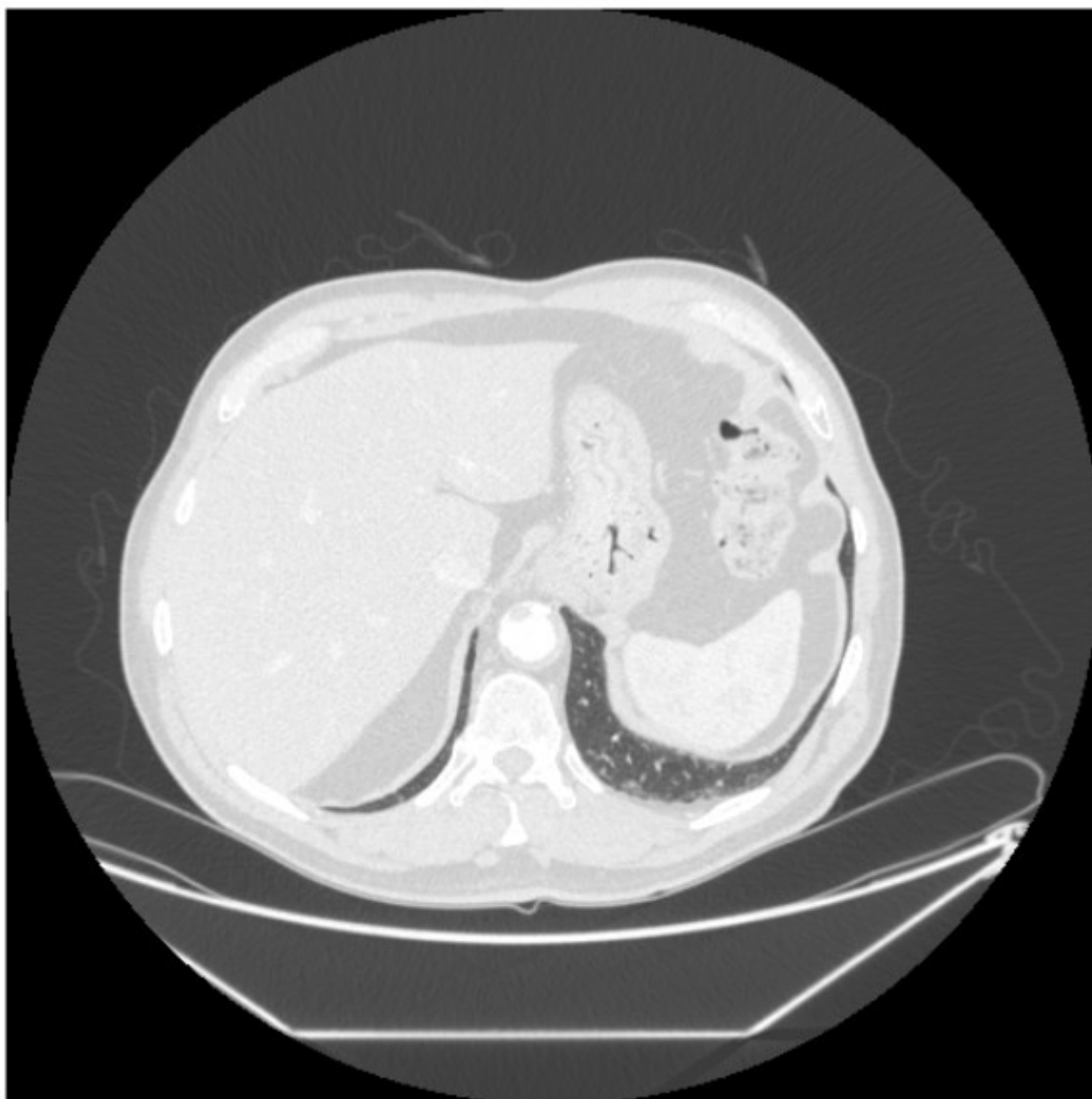
Ruim

Moderado

Bom

Ótimo

Resultado 9



não vejo nódulo ... estas lesões muito pequenas (< 5 mm) eu deixaria de fora

Péssimo

Ruim

Moderado

Bom

Ótimo

Resultado 10



Péssimo

Ruim

Moderado

Bom

Ótimo

Resultado 11



Pésimo

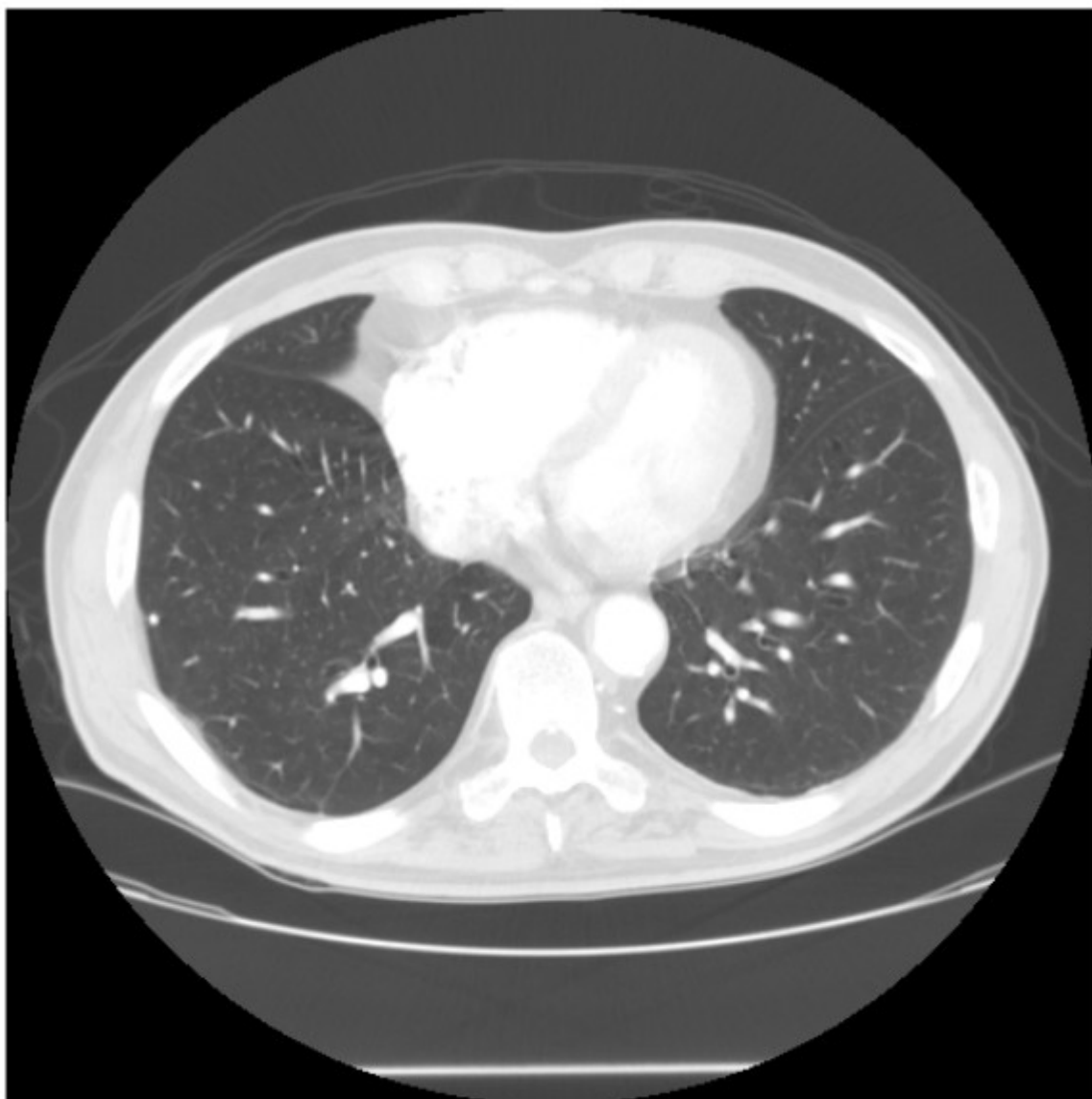
Ruim

Moderado

Bom

Ótimo

Resultado 12



Péssimo

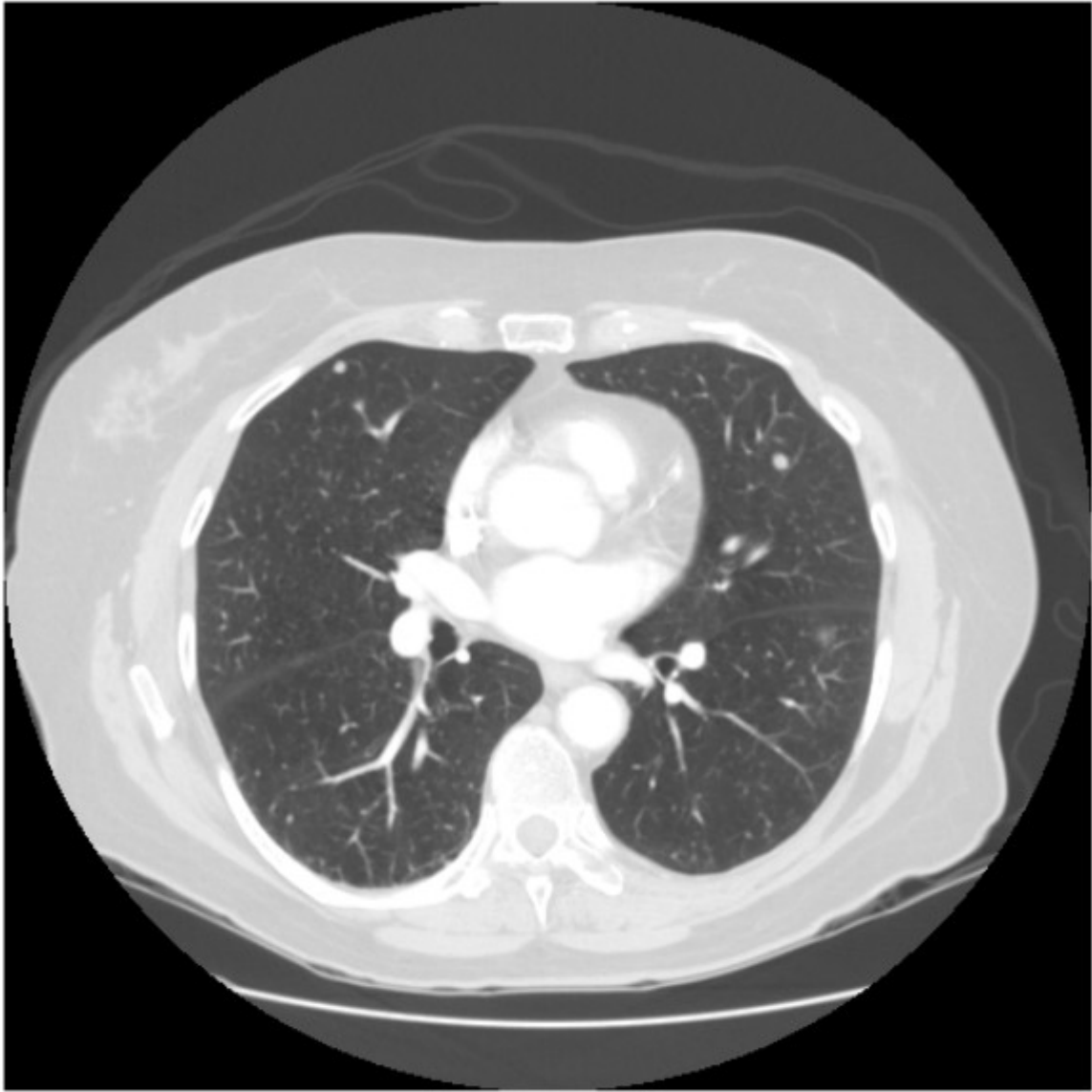
Ruim

Moderado

Bom

Ótimo

Resultado 13



Péssimo

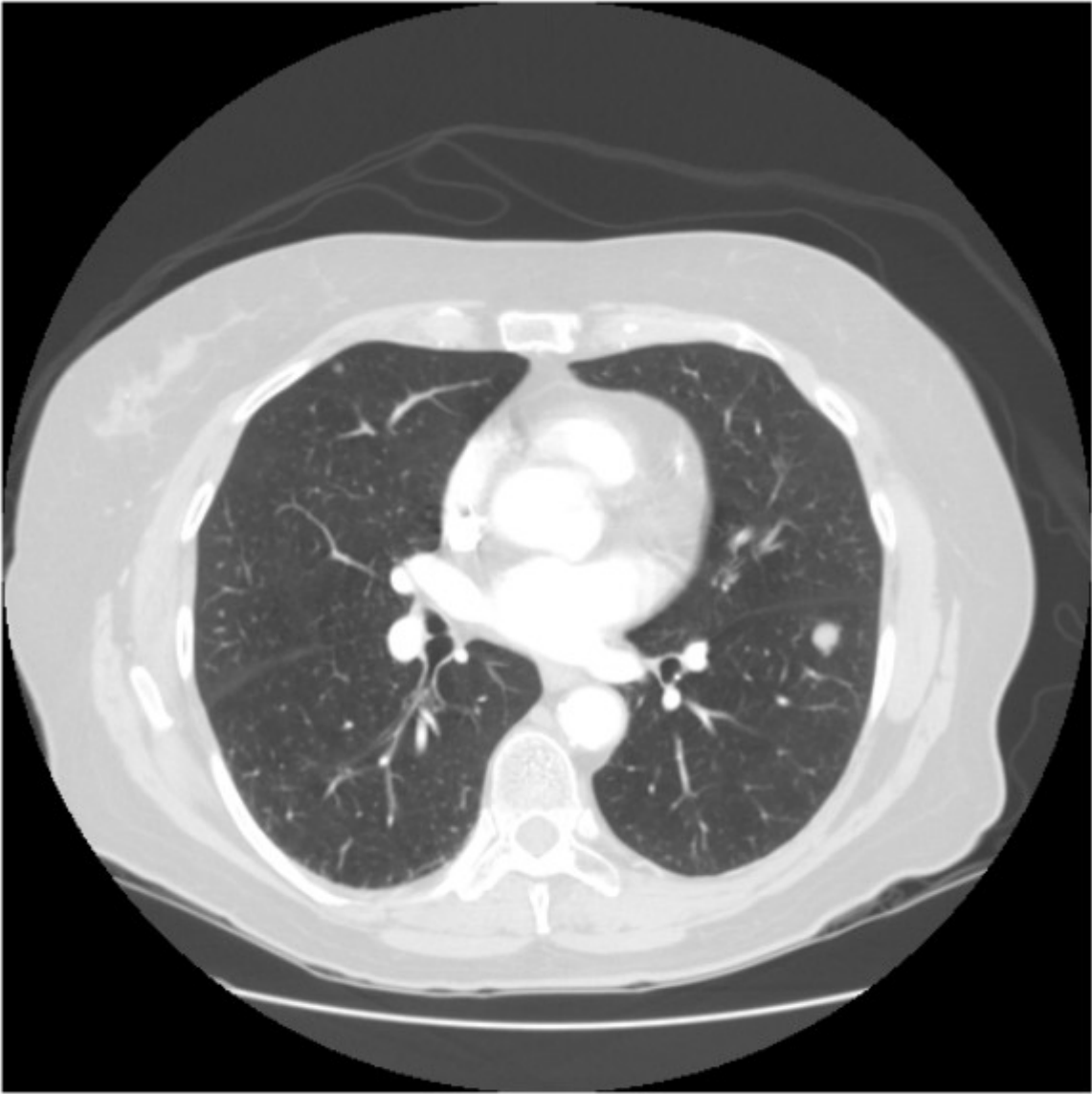
Ruim

Moderado

Bom

Ótimo

Resultado 14



Pésimo

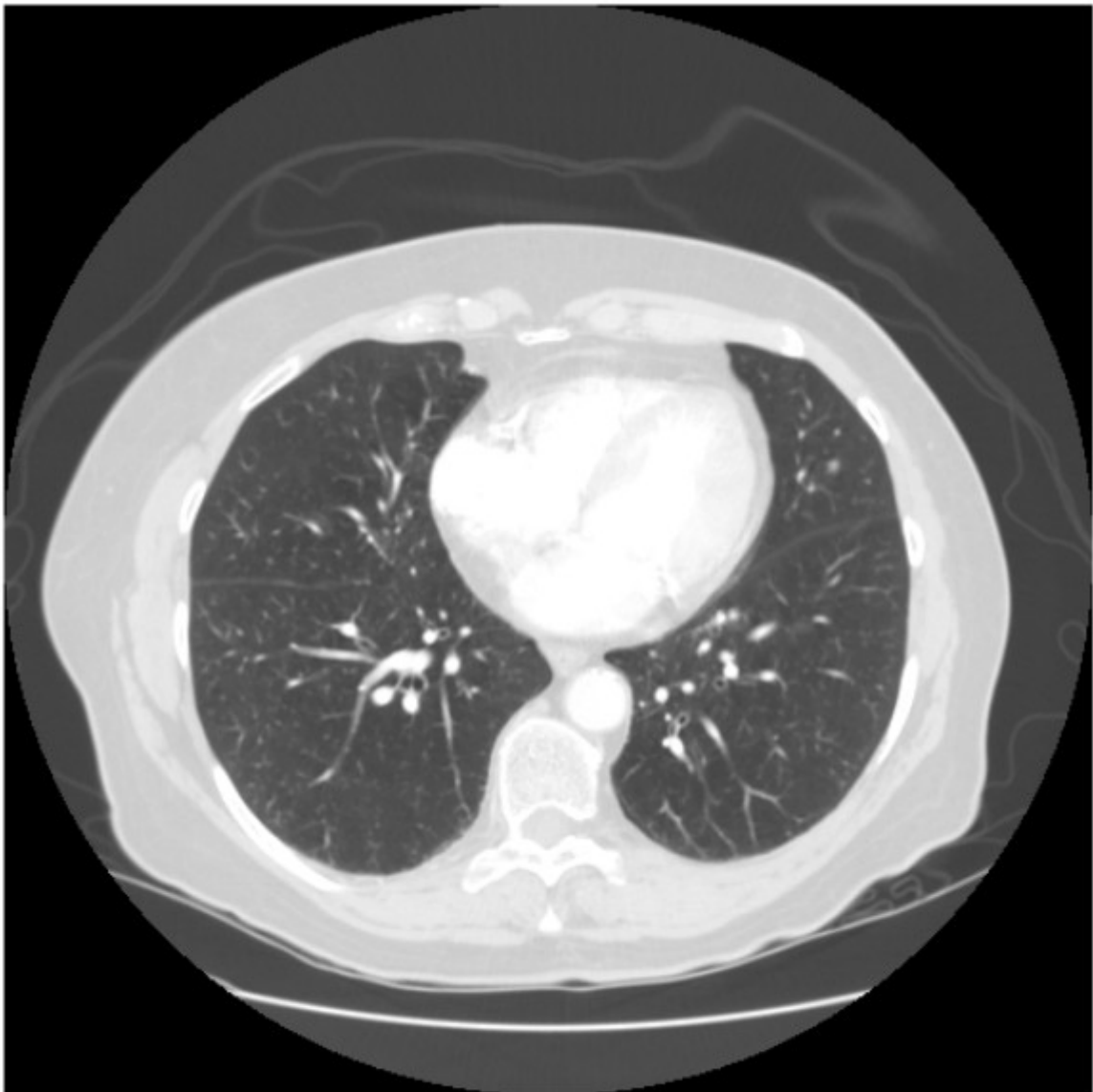
Ruim

Moderado

Bom

Ótimo

Resultado 15



Pésimo

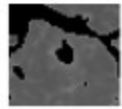
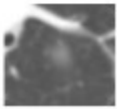
Ruim

Moderado

Bom

Ótimo

Resultado 16



Péssimo

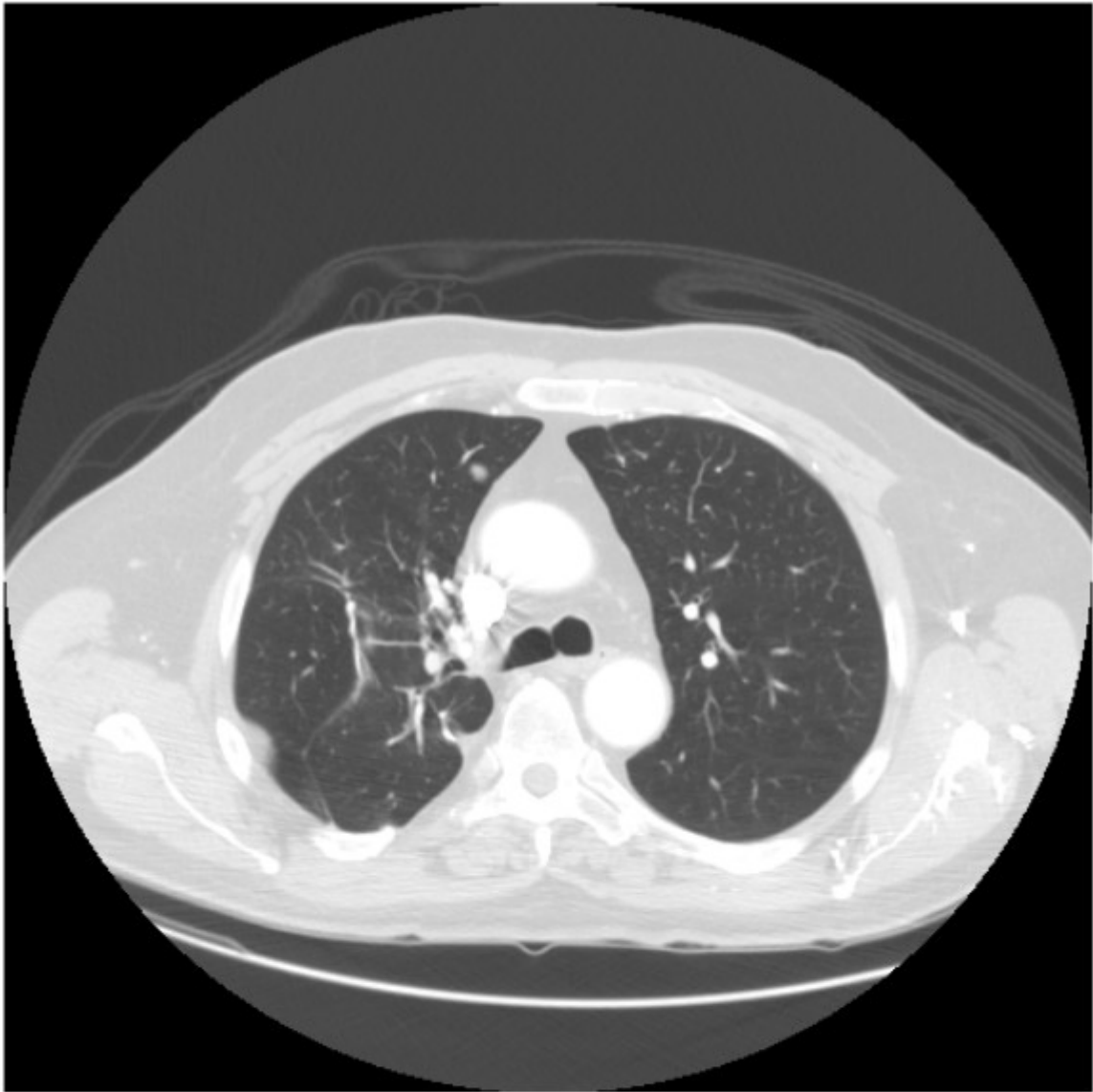
Ruim

Moderado

Bom

Ótimo

Resultado 17



Péssimo

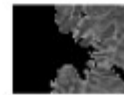
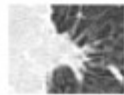
Ruim

Moderado

Bom

Ótimo

Resultado 18



Péssimo

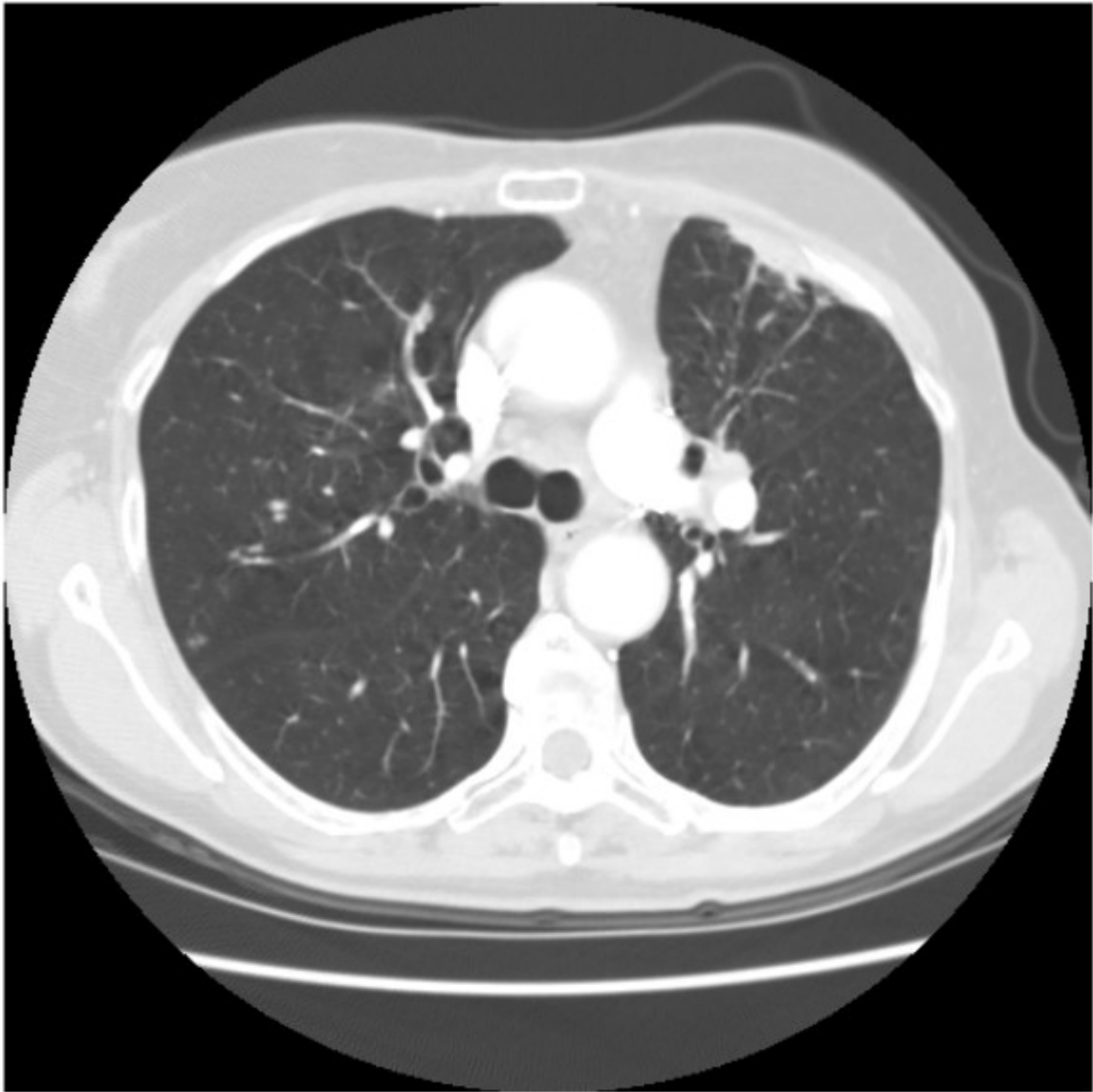
Ruim

Moderado

Bom

Ótimo

Resultado 19



Péssimo

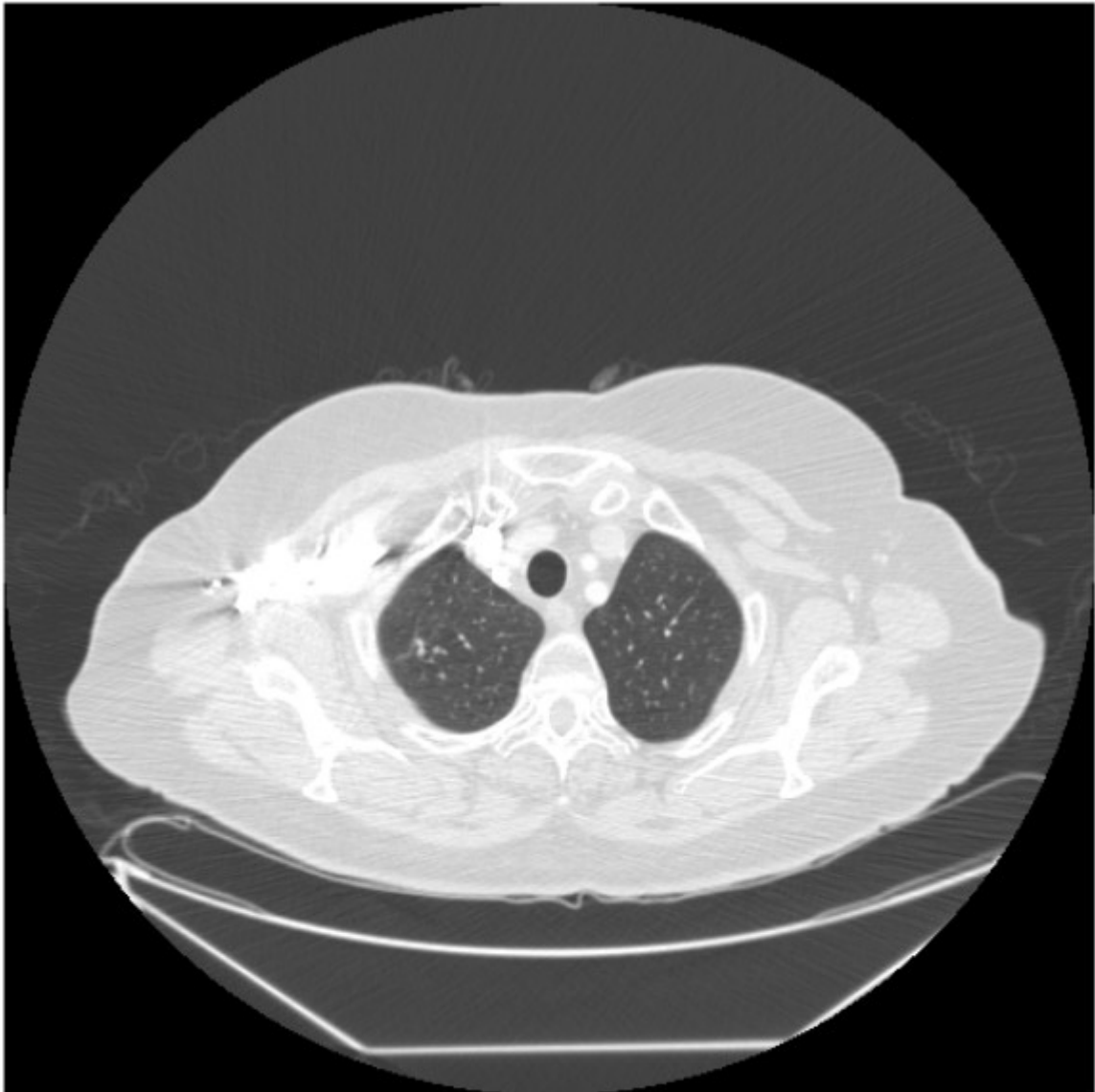
Ruim

Moderado

Bom

Ótimo

Resultado 20



Péssimo

Ruim

Moderado

Bom

Ótimo