

UNIVERSIDADE FEDERAL DO PARANÁ

CAMILE NUNES DOS ANJOS

EXTRAÇÃO DE CARACTERÍSTICAS RADIÔMICAS PARA CLASSIFICAÇÃO DE
NÓDULOS PULMONARES UTILIZANDO SVM E BRF

CURITIBA PR

2025

CAMILE NUNES DOS ANJOS

EXTRAÇÃO DE CARACTERÍSTICAS RADIÔMICAS PARA CLASSIFICAÇÃO DE
NÓDULOS PULMONARES UTILIZANDO SVM E BRP

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Informática Biomédica, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: Prof. Dr. Lucas Ferrari de Oliveira.

Coorientador: Prof. Dr. Fernando Roberto Pereira.

CURITIBA PR

2025

AGRADECIMENTOS

A conclusão deste trabalho só foi possível graças ao apoio, incentivo e colaboração de pessoas que estiveram ao meu lado durante toda a jornada.

Agradeço ao professor Lucas Ferrari de Oliveira, pelas orientações, sugestões valiosas e constante incentivo diante de cada desafio enfrentado. Ao Lucas José Ribeiro, que iniciou esta caminhada comigo e teve contribuição essencial na construção e implementação deste trabalho.

Sou grata também aos meus amigos e familiares, pelo suporte emocional, paciência e compreensão nos momentos mais desafiadores e, em especial, ao Eduardo Rosso Barbosa, por estar ao meu lado com apoio incondicional e palavras de incentivo fundamentais para a conclusão deste trabalho.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho, o meu sincero agradecimento.

RESUMO

O câncer de pulmão é uma das principais causas de mortalidade no mundo, sendo o diagnóstico precoce fundamental para aumentar as chances de sucesso no tratamento. Nesse contexto, a análise de imagens médicas, aliada a técnicas de inteligência artificial e à radiômica, têm se destacado como uma estratégia promissora no apoio ao diagnóstico. Este trabalho tem como objetivo avaliar o desempenho dos algoritmos Support Vector Machine (SVM) e Balanced Random Forest (BRF) na classificação de imagens de tomografia computadorizada (TC) para a identificação de nódulos pulmonares. O primeiro passo foi submeter as imagens a um processo de pré-processamento, que inclui limpeza, transformação e redução dos dados. A extração das características radiômicas foi realizada através da biblioteca PyRadiomics, envolvendo descritores estatísticos de primeira ordem, características de textura e métricas relacionadas à forma das regiões de interesse. Em seguida, os dados foram utilizados para treinar e validar os modelos de aprendizado de máquina, considerando técnicas para lidar com o desequilíbrio das classes "nódulo" e "não nódulo" da base de dados. Os resultados indicaram um desempenho limitado de ambos os modelos. O SVM apresentou maior sensibilidade (0,762) e F1-score (0,477), o que significa que conseguiu identificar uma maior parte dos casos positivos, porém com uma precisão muito baixa (0,349), resultando em muitos falsos positivos. O BRF obteve uma especificidade maior (0,871) e uma precisão razoavelmente maior (0,442), mas com recall baixo (0,492) e F1-score baixo (0,463), mostrando que o modelo falhou em detectar corretamente a maioria dos nódulos. De forma geral, ambos os modelos apresentaram desempenho insatisfatório, evidenciando a dificuldade em lidar com os desafios relacionados principalmente ao desbalanceamento da base de dados. Esses resultados ressaltam a necessidade de aprimoramento nas etapas de pré-processamento, seleção de atributos e na escolha de algoritmos mais robustos para esse tipo de tarefa.

Palavras-chave: Radiômica. Aprendizado de Máquinas. Classificação de Nódulos. Câncer de Pulmão.

ABSTRACT

Lung cancer is one of the leading causes of mortality worldwide, with early diagnosis being essential to improve treatment success rates. In this context, medical image analysis combined with artificial intelligence techniques and radiomics has emerged as a promising strategy to improve diagnosis. This study aims to evaluate the performance of the Support Vector Machine (SVM) and Balanced Random Forest (BRF) algorithms in classifying computed tomography (CT) images for the identification of pulmonary nodules. The first step involved a preprocessing pipeline, including data cleaning, transformation, and reduction. Radiomic feature extraction was performed using the PyRadiomics library, covering first-order statistical descriptors, texture features, and shape-related metrics from the regions of interest. The extracted features were then used to train and validate machine learning models, applying techniques to handle the class imbalance between “nodule” and “non-nodule” samples in the dataset. The results indicated a limited performance for both models. The SVM showed higher sensitivity (0.762) and F1-score (0.477), meaning it was able to identify a greater portion of positive cases; however, it achieved a very low precision (0.349), resulting in a high number of false positives. The BRF obtained higher specificity (0.871) and a slightly better precision (0.442), but with low recall (0.492) and a low F1-score (0.463), indicating that the model failed to correctly detect most nodules. Overall, both models demonstrated unsatisfactory performance, highlighting the challenges mainly related to the class imbalance in the dataset. These results emphasize the need for improvements in preprocessing steps, feature selection, and the adoption of more robust algorithms for this type of task.

Keywords: Radiomics. Machine Learning. Nodule Classification. Lung Cancer.

LISTA DE FIGURAS

2.1	Fluxo de processamento de uma imagem, desde a aquisição até o reconhecimento e interpretação. Fonte: (Marques e Oliveira, sd)..	9
2.2	Exemplo de um pipeline radiômico. Fonte: (Wagner et al., 2021).	14
2.3	Matriz de Confusão. TP, TN, FP e FN são, respectivamente, verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Fonte: A autora (2025)..	18
3.1	Visualização de um corte bidimensional de uma imagem normalizada. Fonte: A autora (2025)..	21
3.2	Imagem 3.1 com Gaussian Blur. Fonte: A autora (2025)..	21
3.3	Imagem 3.1 após aplicação da limiarização de Otsu, com fundo em vermelho e objeto em preto. Fonte: A autora (2025)..	22
3.4	Visualização dos 12 cortes axiais ao longo do eixo z de uma imagem, comparando a imagem original, suavizada com Gaussian Blur, e segmentada por Otsu onde a parte vermelha é o fundo. Fonte: A autora (2025)..	22
3.5	Projeção t-SNE das classes positivas e negativas em duas dimensões. Fonte: A autora (2025)..	24
3.6	Gráfico da Variância Acumulada Explicada para escolha do número de componentes ideal para aplicação da técnica de PCA. Fonte: A autora (2025). 25	
3.7	Mapa de calor com os pesos atribuídos a cada variável de cada componente principal do PCA. Fonte: A autora (2025)..	25
3.8	Resultado das métricas de desempenho do classificador SVM em cada iteração da validação cruzada. Fonte: A autora (2025)..	26
3.9	Resultado das métricas de desempenho do classificador BRF em cada iteração da validação cruzada com seu respectivo desvio padrão. Fonte: A autora (2025). . .	26
3.10	Média das métricas de desempenho obtidas pelos classificadores SVM e BRF. Fonte: A autora (2025)..	27
3.11	Resultado do cálculo do coeficiente de variação para os resultados das médias das métricas de desempenho com desvio padrão. Fonte: A autora (2025).	27

SUMÁRIO

1	INTRODUÇÃO	7
1.1	CONTEXTO E PROBLEMA.	7
1.2	OBJETIVOS	7
1.2.1	Objetivo Geral.	7
1.2.2	Objetivos Específicos	7
1.3	JUSTIFICATIVA	8
2	FUNDAMENTOS E APLICAÇÕES DA RADIÔMICA EM IMAGENS MÉDICAS	9
2.1	PROCESSAMENTO DE IMAGENS MÉDICAS	9
2.1.1	Pré-processamento	10
2.1.2	Tratamento dos Dados	10
2.1.3	Redução de Dimensionalidade	11
2.1.3.1	<i>Principal Component Analysis (PCA)</i>	11
2.1.3.2	<i>t-Distributed Stochastic Neighbor Embedding (t-SNE)</i> .	12
2.1.4	Técnicas de Segmentação e Filtragem	12
2.1.4.1	<i>Thresholding</i>	12
2.1.4.2	<i>Limiarização de Otsu</i>	13
2.1.4.3	<i>Gaussian Blur</i> .	13
2.1.4.4	<i>SimpleITK (Sitk)</i>	13
2.2	RADIÔMICA.	13
2.3	APRENDIZADO DE MÁQUINA EM RADIÔMICA.	15
2.3.1	Support Vector Machines (SVM).	16
2.3.2	Balanced Random Forest (BRF)	16
2.3.3	Estratégias de Validação	17
2.3.4	Métricas de Desempenho	17
3	METODOLOGIA	20
3.1	BASE DE DADOS	20
3.2	CANDIDATOS A NÓDULO.	20
3.3	PRÉ-PROCESSAMENTO	20
3.4	EXTRAÇÃO DE CARACTERÍSTICAS RADIÔMICAS	22
3.5	CLASSIFICAÇÃO	23
3.6	RESULTADOS	26
4	CONCLUSÃO	28
	REFERÊNCIAS	29

1 INTRODUÇÃO

1.1 CONTEXTO E PROBLEMA

O câncer de pulmão é o tipo mais frequente e letal de câncer no mundo, responsável por uma parcela significativa da incidência e mortalidade global, conforme os dados de 2022 da Agência Internacional de Pesquisa sobre o Câncer (IARC) da Organização Mundial da Saúde (OMS) (Ferlay J, 2024). Diante desse cenário desafiador, o diagnóstico precoce de nódulos pulmonares é fundamental para o sucesso no tratamento e aumento da taxa de sobrevivência dos pacientes (JATOBÁ, 2021).

Atualmente, a tomografia computadorizada de baixa dose (LDCT) é considerada o padrão ouro para o rastreamento da doença, permitindo a detecção precoce. No entanto, a análise manual das imagens é uma tarefa complexa e suscetível à subjetividade, além de estar associada a elevadas taxas de falsos positivos, que geram estresse nos pacientes e demandam procedimentos desnecessários, onerando o sistema de saúde (Guerreiro et al., 2024; JATOBÁ, 2021).

Nesse contexto, técnicas de inteligência artificial (IA), como o aprendizado de máquina aliado à radiômica, vêm sendo destacadas como abordagens promissoras. A radiômica permite a extração de características quantitativas das imagens médicas, convertendo informações visuais em dados mensuráveis. Esses dados podem ser utilizados por algoritmos de aprendizado de máquina para construir modelos preditivos capazes de auxiliar na detecção e caracterização de nódulos pulmonares, mitigando limitações da avaliação subjetiva (van Timmeren et al., 2020).

Entretanto, a seleção adequada dos algoritmos e a análise de seu desempenho continuam sendo desafios relevantes, especialmente considerando o desequilíbrio presente em bases de dados médicas, onde há uma quantidade muito maior de amostras de tecido saudável em comparação com amostras contendo nódulos. Isso impacta diretamente na eficácia dos modelos e na sua capacidade de generalização (Salmi et al., 2024).

Nesse cenário, torna-se fundamental investigar e avaliar o desempenho de diferentes algoritmos de aprendizado de máquina aplicados na detecção e caracterização de nódulos pulmonares, levando em consideração seus desafios recorrentes, bem como validar o potencial da radiômica como ferramenta complementar na análise quantitativa de imagens médicas.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Este trabalho tem como objetivo avaliar o desempenho dos algoritmos Support Vector Machine (SVM) e Balanced Random Forest (BRF) na classificação de imagens médicas de tomografia computadorizada utilizando a radiômica como gerador de características, visando identificar candidatos a nódulos pulmonares. A proposta busca contribuir com o desenvolvimento de sistemas de suporte à decisão, capazes de auxiliar no diagnóstico precoce de câncer de pulmão.

1.2.2 Objetivos Específicos

- Validar o uso da radiômica em conjunto com ferramentas de aprendizado de máquinas para apoio ao diagnóstico de doenças;
- Fornecer uma base para pesquisas futuras na área de processamento de imagens médicas e inteligência artificial aplicada à saúde.

1.3 JUSTIFICATIVA

A digitalização na medicina, aliada ao avanço das técnicas de inteligência artificial (IA), tem possibilitado o desenvolvimento de ferramentas capazes de apoiar o diagnóstico e a tomada de decisão clínica. Técnicas como aprendizado de máquina, quando aplicadas a imagens médicas por meio da radiômica, permitem a extração de informações quantitativas e padrões complexos que são frequentemente imperceptíveis à análise visual humana. Essa integração de tecnologias tem demonstrado eficácia na identificação de marcadores associados a diversas patologias, incluindo o câncer de pulmão, proporcionando a base para diagnósticos mais precisos e tratamentos mais assertivos (van Timmeren et al., 2020).

Portanto, este trabalho se justifica pela necessidade de aprimorar o diagnóstico precoce do câncer de pulmão, utilizando o potencial transformador da IA. Ao investigar o desempenho dos algoritmos Support Vector Machine (SVM) e Balanced Random Forest (BRF) na classificação de imagens de tomografia computadorizada para identificar candidatos a nódulos pulmonares, busca-se não apenas validar a aplicabilidade dessas técnicas, mas também oferecer uma base para o desenvolvimento de sistemas de suporte à decisão. Espera-se que os resultados contribuam para a saúde dos pacientes, otimizando o processo diagnóstico e abrindo caminhos para futuras pesquisas na interseção entre processamento de imagens médicas e inteligência artificial aplicada à saúde.

2 FUNDAMENTOS E APLICAÇÕES DA RADIÔMICA EM IMAGENS MÉDICAS

Este capítulo apresenta os principais fundamentos teóricos para compreender a radiômica e suas aplicações em imagens médicas. Serão abordadas as etapas do processamento de imagens, técnicas de segmentação, filtragem, redução de dimensionalidade e normalização, além de uma visão geral da radiômica e dos classificadores de aprendizado de máquina utilizados neste trabalho. Por fim, são apresentadas também as métricas de desempenho e as estratégias de validação usadas na avaliação dos modelos.

2.1 PROCESSAMENTO DE IMAGENS MÉDICAS

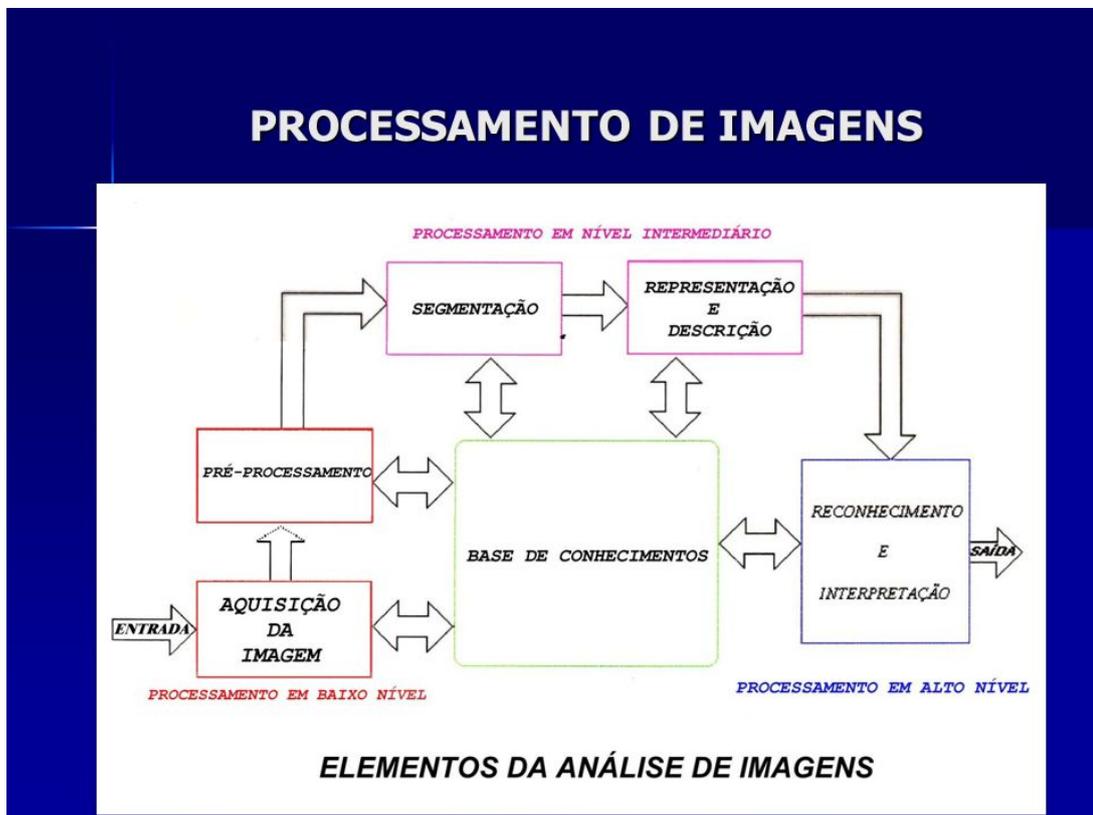


Figura 2.1: Fluxo de processamento de uma imagem, desde a aquisição até o reconhecimento e interpretação. Fonte: (Marques e Oliveira, sd).

Imagens médicas são representações digitais obtidas a partir da captação de radiações que interagem com estruturas do corpo humano, permitindo a análise de diferentes características anatômicas e funcionais. Na prática, uma imagem digital é como uma matriz de pixels, onde cada pixel representa a intensidade do sinal em um ponto específico da imagem, em tons de cinza. O pixel é a menor unidade da imagem e sua dimensão depende da resolução espacial do equipamento utilizado. (Silva, 2009).

Essa organização em forma de matriz permite que a imagem seja manipulada, analisada e processada, extraindo informações relevantes para diagnóstico e tomada de decisão. (Silva, 2009).

A descoberta dos Raios-X em 1895 possibilitou a visualização de estruturas internas do corpo, e desde então surgiram diversas tecnologias que permitem analisar diferentes características anatômicas e funcionais do organismo (Silva, 2009).

Segundo Silva (2009), podemos dividir as modalidades de imagens médicas em dois grandes grupos:

- Anatômicas: focam na visualização da morfologia, como Raio-X, Tomografia Computadorizada (TC), Ressonância Magnética (RM) e Ultrassonografia (US).
- Funcionais: fornecem informações sobre o metabolismo e funcionamento dos tecidos, como Cintilografia, SPECT, PET e Ressonância Magnética Funcional.

Cada modalidade possui características específicas e desafios próprios, que podem ser minimizados com o uso de técnicas de processamento de imagens (Silva, 2009).

2.1.1 Pré-processamento

O pré-processamento dos dados de entrada, segundo Mahmud Sujon et al., é essencial para a construção de modelos preditivos. Qualquer processamento aplicado a dados brutos com o objetivo de prepará-los para análises ou outros processamentos, como por exemplo o treinamento de modelos de aprendizado de máquinas, pode ser entendido como pré-processamento (Datacamp, 2025).

Existem 3 principais etapas envolvidas neste processo, são elas: limpeza de dados, transformação de dados e redução de dados (Gomes, 2019). A limpeza envolve atividades como preenchimento de dados ausentes, redução de ruídos, remoção de dados duplicados e correção de formatos inconsistentes. Já a transformação diz respeito ao dimensionamento de dados, normalização, seleção de atributos, entre outros (Gomes, 2019; Datacamp, 2025).

A redução, por sua vez, é usada para lidar com um grande volume de dados, reduzindo o número de recursos e preservando informações essenciais. São exemplos de aplicação a análise de componentes principais (PCA), seleção de recursos, métodos de amostragem, etc (Gomes, 2019; Datacamp, 2025).

Imagens médicas digitais geralmente necessitam de algum tipo de pré-processamento para que sejam corrigidas ou realçadas adequadamente, seja para melhorar o contraste, corrigir pixels ou reduzir o ruído. Isso permite que técnicas avançadas de processamento sejam mais eficientes (Silva et al., 2019).

2.1.2 Tratamento dos Dados

A normalização e a padronização são técnicas de escalonamento importantes no tratamento de dados, pois elas compactam uma quantidade grande de características em formatos mais fáceis de interpretar e gerenciar, diminuindo também a complexidade computacional dos modelos (Mahmud Sujon et al., 2024).

Um método muito popular de padronização é o StandardScaler do scikit-learn (Warmerdam, 2024). Nele, o score padrão (ou z-score) de uma amostra x é calculado pela fórmula 2.1, onde u é a média das amostras de treinamento (ou zero, se o parâmetro `with_mean=False` for definido), e s é o desvio padrão das amostras de treinamento (ou um, se `with_std=False`) (ScikitLearn, 2025d).

O z-score é uma forma de medir quantos desvios padrão cada valor está distante da média. Ele transforma dados brutos em dados que seguem uma distribuição z , que é uma

distribuição normal com média igual a 0 e desvio padrão igual a 1. Por isso, ela também é conhecida como distribuição normal padrão (Oracle, sd).

$$z = \frac{x - \mu}{s} \quad (2.1)$$

O StandardScaler é sensível a outliers (pontos fora do padrão), isso significa que valores muito distantes da média podem distorcer o escalonamento e fazer com que as variáveis sejam escaladas de forma desigual (ScikitLearn, 2025d). Por essa razão, após a padronização, precisamos selecionar os valores que ficaram muito distantes de zero, geralmente acima de +3 ou abaixo de -3, e remover do conjunto, pois estes são considerados outliers (Oracle, sd).

O desvio padrão é uma medida estatística que quantifica a dispersão dos valores em um conjunto de dados, indicando o quanto os valores se afastam da média desse conjunto (Gomes, 2024). O cálculo do desvio padrão amostral é dado por 2.2:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (2.2)$$

Para avaliar se um desvio padrão é bom ou ruim, é necessário analisar o contexto e objetivos da análise. No caso de uma pesquisa científica, por exemplo, um desvio padrão baixo pode indicar dados confiáveis e replicáveis (Gomes, 2024).

2.1.3 Redução de Dimensionalidade

No contexto do aprendizado de máquina, dimensões se referem aos recursos ou atributos dos dados e, conforme adicionamos mais dimensões, o volume do espaço aumenta exponencialmente, levando a dados mais esparsos, necessidade de mais recursos computacionais, maior tempo de processamento, perda de desempenho e dificuldades de visualização. Isso é conhecido como "maldição da dimensionalidade", e ocorre quando a complexidade dos dados aumenta sem aumentar a quantidade de exemplos (Datacamp, 2024b).

A redução da dimensionalidade é a principal solução para este problema (Datacamp, 2024b). Ela pode ser vista como a aplicação de uma transformação nos dados que estão em um espaço de alta dimensão para um espaço de baixa dimensão, assegurando a preservação da topologia do espaço dos dados originais (Medeiros e Costa, 2008).

Com a representação reduzida, tarefas como classificação e agrupamento são capazes de gerar resultados mais precisos e facilmente interpretáveis, com custos computacionais reduzidos (Cunningham, 2008). Alguns métodos de redução de dimensionalidade são citados a seguir.

2.1.3.1 Principal Component Analysis (PCA)

O PCA é um método linear de extração de atributos que busca projetar os dados em um espaço onde as variáveis não são correlacionadas. Ele descorrelaciona as variáveis originais por meio de uma transformação linear ortogonal e projeta os dados em um sistema de coordenadas onde os eixos, chamados de componentes principais, são independentes entre si (Bôa et al., 2022; Medeiros e Costa, 2008).

Dessa forma, a variância dos dados é maximizada em ordem decrescente e podemos visualizar e analisar o conjunto de dados com mais eficiência. (Bôa et al., 2022; Medeiros e Costa, 2008).

2.1.3.2 *t-Distributed Stochastic Neighbor Embedding (t-SNE)*

O t-SNE é uma técnica de redução de dimensionalidade não-linear não supervisionada, ou seja, que não depende de rótulos ou categorias previamente conhecidas, usada para explorar e visualizar dados de alta dimensão (Datacamp, 2024a).

Ele calcula a similaridade entre os pontos em um espaço de alta dimensão, onde os pontos próximos têm maior possibilidade de serem vizinhos. Em seguida, tenta representar esses pontos em um espaço de baixa dimensão, mantendo as relações entre eles. Por fim, usa descida de gradiente para minimizar a divergência entre as distribuições de similaridade (Datacamp, 2024a).

Esse processo permite a criação de clusters (agrupamentos) e subclusters (subagrupamentos) de pontos de dados semelhantes no espaço de menor dimensão, que, ao serem visualizados, ajudam a entender a estrutura e as relações do espaço de maior dimensão (Datacamp, 2024a).

2.1.4 Técnicas de Segmentação e Filtragem

Segmentação de imagens é uma técnica que divide uma imagem em regiões (segmentos) para facilitar a identificação de objetos e acelerar o processamento (IBM, sd).

Filtragem de imagens consiste em transformar os valores dos pixels considerando também seus vizinhos, utilizando máscaras (ou kernels) aplicadas sobre a imagem. A máscara funciona da seguinte forma: o pixel central (i, j) é substituído por um novo valor calculado a partir dos pixels vizinhos, conforme os pesos definidos pela máscara, e assim é gerada uma imagem processada (Spring - DPI/INPE, 2006).

Os filtros espaciais podem ser classificados em passa-baixa, passa-alta ou passa-banda. Os mais utilizados em processamento de imagens são o passa-baixa, um filtro linear que suaviza a imagem atenuando as altas frequências, e o passa-alta, também linear, realça detalhes, produzindo uma "agudização" ("sharpering") da imagem, ou seja, torna mais nítidas as transições entre regiões diferentes (Spring - DPI/INPE, 2006).

Nas subseções abaixo, temos a definição do que é um threshold, um exemplo de algoritmo que usa o threshold para segmentação, um exemplo de técnica de filtragem e uma breve explicação do que é a biblioteca Sitk, que implementa esses exemplos de algoritmos e nos permite realizar o processamento das imagens de maneira fácil.

2.1.4.1 *Thresholding*

Thresholding é o processo de conversão de uma imagem em escala de cinza para uma imagem binária, onde os pixels são representados por dois valores: 0 ou 255. Por exemplo, seja p um valor de pixel, todos os pixels com intensidades inferiores a p são definidos como 0 (preto) e todos os pixels com intensidades superiores a p são definidos como 255 (branco). Isso cria uma representação binária da imagem, destacando as áreas de interesse de acordo com o valor de threshold escolhido (Arcanjo, 2023).

É comum que esta técnica seja utilizada na segmentação de objetos ou áreas específicas de uma imagem, de forma que facilite a análise e extração de informações relevantes (Arcanjo, 2023).

A escolha do melhor valor de threshold varia conforme o problema a ser resolvido, e métodos simples, onde os valores são fornecidos manualmente, requerem bastante experimentação e ajustes. Métodos mais avançados como o thresholding adaptativo encontram um valor ótimo

de threshold ao analisar as vizinhanças dos pixels, permitindo ajustar o valor de forma adaptativa e levando em consideração o contexto local de cada pixel (Arcanjo, 2023).

2.1.4.2 Limiarização de Otsu

A Limiarização de Otsu é um algoritmo de limiarização automática, proposto por Nobuyuki Otsu, cujo objetivo é determinar o valor ideal de limiar (threshold) de uma imagem em tons de cinza, que separa os elementos de primeiro plano e de fundo em dois grupos, atribuindo a cada um deles a cor branca ou preta (Torok, sd).

Essa abordagem é eficaz em imagens com histogramas bimodais, pois permite uma divisão adequada usando um único valor de limiar (Torok, sd).

2.1.4.3 Gaussian Blur

Esta é uma técnica de processamento de imagens que faz uma média ponderada de pixels com base na função gaussiana. A origem deste nome é uma homenagem a Carl Friedrich Gauss, cuja função gaussiana serve de base para esse tipo de técnica. É considerado um filtro passa-baixa, ou seja, um filtro que reduz ruídos de alta frequência suavizando valores de pixels irregulares (Cloudinary, 2025).

Seu algoritmo calcula a intensidade dos pixels próximos e diminui os pesos gradualmente conforme a distância entre os pixels aumenta, resultando em um efeito de suavização na imagem, que pode ser ajustado em diferentes níveis (Cloudinary, 2025).

2.1.4.4 SimpleITK (Sitk)

O SimpleITK é uma interface de programação simplificada para os algoritmos e estruturas de dados do Insight Toolkit (ITK) para segmentação, registro e análise avançada de imagens (SimpleITK, sdc). Ele possui, inclusive, algoritmos que aplicam as técnicas de segmentação e filtragem citadas nas subseções acima; são eles SmoothingRecursiveGaussian e OtsuThresholdImageFilter (SimpleITK, sdb).

O SmoothingRecursiveGaussian é implementado usando filtros gaussianos recursivos e calcula a suavização da imagem por convolução, que aplica um filtro sobre a imagem ao combinar os valores dos pixels com uma matriz de pesos (kernel), resultando em uma média ponderada que suaviza variações locais. Os kernels gaussianos são implementados com filtros IIR (Infinite Impulse Response), que calculam a saída com base tanto nas amostras de entrada quanto em saídas anteriores, permitindo um controle mais preciso sobre a resposta em frequência (SimpleITK, sdb; ScienceDirect, sd).

Já o OtsuThresholdImageFilter limiariza uma imagem usando o limiar de Otsu. Ele cria uma imagem binária com o limiar e separa os elementos de primeiro plano e de fundo. O OtsuThresholdCalculator é quem calcula o limiar internamente e o BinaryThresholdImageFilter é quem aplica esse limiar à imagem de entrada (SimpleITK, sda).

2.2 RADIÔMICA

A radiômica é um processo que permite a extração automática e análise de dados numéricos. Diferente dos métodos tradicionais, que dependem da observação visual feita por especialistas, a radiômica transforma as imagens em números de forma padronizada e confiável. Esses dados incluem informações como intensidade, forma, tamanho, volume e textura, que

oferecem mais informações sobre o objeto analisado, como o tipo de tumor ou as condições do tecido ao redor (JATOBÁ, 2021; SCAPICCHIO et al., 2021).

A extração de características radiômicas, como implementada no PyRadiomics (van Griethuysen et al., 2017), segue um pipeline estruturado e padronizado, como mostrado na Figura 2.2, alinhado com as diretrizes da *Image Biomarker Standardisation Initiative* (IBSI) (Zwanenburg et al., 2020).

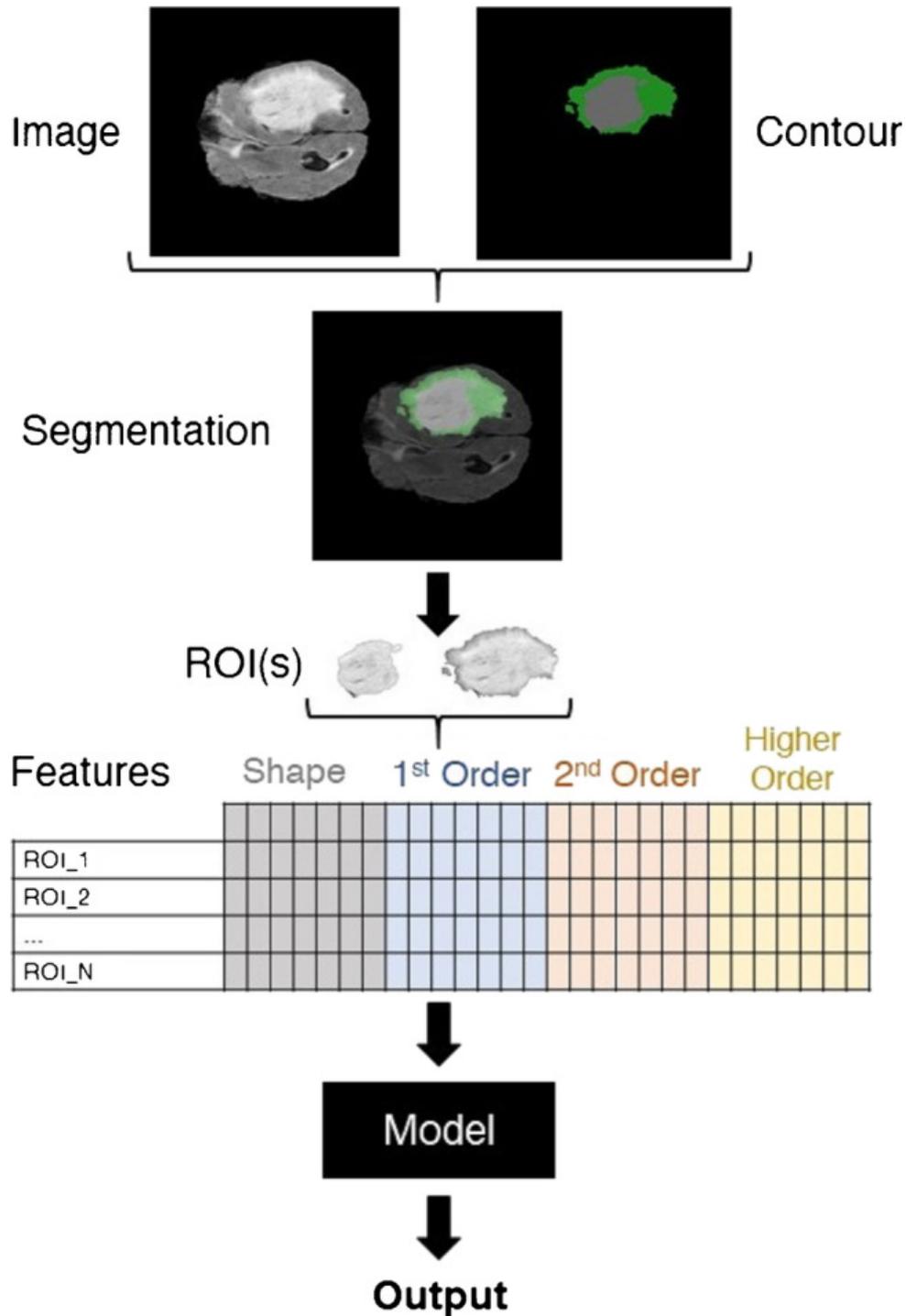


Figura 2.2: Exemplo de um pipeline radiômico. Fonte: (Wagner et al., 2021).

O processo se inicia ao receber como entrada uma imagem médica e uma máscara binária que delimita a Região de Interesse (ROI). Em seguida, passa por etapas de pré-processamento

como resampling, normalização e aplicação de filtros quando necessário. A partir disso, são extraídos diferentes tipos de características, como as que descrevem a distribuição estatística dos valores dos voxels, as características de forma, que descrevem a geometria tridimensional da ROI, e também as características de textura que capturam padrões espaciais na distribuição dos valores de intensidade através de matrizes como GLCM (*Gray Level Co-occurrence Matrix*), GLRLM (*Gray Level Run Length Matrix*), GLSZM (*Gray Level Size Zone Matrix*), NGTDM (*Neighbouring Gray Tone Difference Matrix*) e GLDM (*Gray Level Dependence Matrix*) (van Griethuysen et al., 2017; Zwanenburg et al., 2020).

A ferramenta de extração de características (ou *extractor*) é responsável por gerar atributos quantitativos a partir das regiões de interesse das imagens médicas. Além de extrair características da imagem original, o extractor também pode aplicar filtros ou transformações para gerar imagens derivadas, das quais são extraídas características adicionais, aumentando a quantidade de informações quantitativas obtidas (van Griethuysen et al., 2017). No final do processo, um vetor de características é gerado, podendo ser usado como base para análises posteriores, como modelagem estatística ou aprendizado de máquina (van Griethuysen et al., 2017).

Na literatura médica, diversos estudos já aplicam a radiômica em diferentes modalidades de imagens para investigar doenças. A maior parte dessas pesquisas se concentra na área de oncologia, utilizando técnicas como tomografia computadorizada e tomografia por emissão de pósitrons (PET/CT) para avaliar tumores de pulmão e rim. Também são comuns estudos que utilizam ressonância magnética (RM) na análise de câncer de próstata, mama, gliomas e carcinoma hepatocelular (Nogueira-Barbosa e Marchiori, 2019).

2.3 APRENDIZADO DE MÁQUINA EM RADIÔMICA

O aprendizado de máquina (ML) se tornou uma ferramenta essencial no contexto da radiologia, especialmente no apoio à análise de grandes volumes de imagens médicas. Esses algoritmos são capazes de identificar padrões complexos e sutis, muitas vezes imperceptíveis à avaliação humana, potencializando o diagnóstico e a tomada de decisão clínica (Wagner et al., 2021).

A radiômica complementa esse processo ao permitir a extração sistemática de características quantitativas das imagens, convertendo informações visuais em atributos mensuráveis que podem ser interpretados por modelos de aprendizado de máquina. Essas características descrevem propriedades como intensidade, forma, textura e volume das regiões de interesse (ROI), formando vetores numéricos utilizados como entrada nos modelos preditivos (JATOBÁ, 2021; SCAPICCHIO et al., 2021).

O desenvolvimento de modelos supervisionados de aprendizado de máquina exige uma abordagem metodológica bem estruturada, iniciando com a definição do problema clínico por uma equipe multidisciplinar — que inclui médicos especialistas, neurorradiologistas, cientistas de dados e profissionais com conhecimento técnico em processamento de imagens. A seguir, ocorre a coleta e a rotulação dos dados, que envolve seleção e anonimização das imagens, controle de qualidade e verificação por meio de laudos ou desfechos clínicos (Wagner et al., 2021).

Para a construção e avaliação dos modelos, os dados são divididos em conjuntos de treinamento, validação e teste. O conjunto de treinamento permite que o modelo aprenda padrões presentes nos dados, o de validação é utilizado para ajustar hiperparâmetros e evitar o overfitting, e o conjunto de teste avalia o desempenho do modelo. Quando o número de amostras é reduzido, técnicas como a validação cruzada se tornam fundamentais para garantir uma avaliação mais confiável (Wagner et al., 2021).

A escolha do classificador mais adequado depende das características dos dados e do objetivo do estudo, e o desempenho dos modelos é avaliado utilizando métricas como acurácia, sensibilidade, especificidade, precisão, sensibilidade, F1-score, etc (Wagner et al., 2021).

Os classificadores escolhidos neste trabalho, bem como as estratégias de validação aplicadas, são apresentados nas subseções a seguir, escolhidos com base no cenário em que são melhor aplicados.

2.3.1 Support Vector Machines (SVM)

São um conjunto de métodos de aprendizado supervisionado, ou seja, que dependem de rótulos ou categorias previamente conhecidas, usados para classificação, regressão e detecção de outliers (ScikitLearn, 2025a). Tem como objetivo encontrar um hiperplano que maximize a margem e defina o limite de decisão que será aplicado na tarefa de classificação (Bôa et al., 2022).

A margem é a distância entre o hiperplano e os vetores de suporte, que são os pontos mais próximos do limite em cada classe. Quando os dados não são linearmente separáveis, admite-se uma variável de folga, representada pela letra grega ξ . Essa variável permite que algumas restrições sejam violadas, ou seja, alguns pontos fiquem fora da margem. O valor de ξ representa a distância entre um ponto que ultrapassou a margem e o vetor de suporte da sua própria classe (Bôa et al., 2022).

Entre suas vantagens estão sua eficácia em espaços de alta dimensão, mesmo quando o número de dimensões é maior que o de amostras; a eficiência em termos de memória, pois utilizam um subconjunto de pontos de treinamento na função de decisão (chamados vetores de suporte); e também sua versatilidade, já que diferentes funções de kernel podem ser especificadas para a função de decisão (ScikitLearn, 2025a).

A função de núcleo ou kernel pode assumir formas como linear, polinomial, sigmoide ou de base radial. E além da escolha do kernel, alguns parâmetros do modelo podem ser ajustados para melhorar o desempenho em casos específicos; por exemplo, em problemas onde se deseja dar mais importância a certas classes ou amostras individuais, os parâmetros `class_weight` e `sample_weight` podem ser usados (ScikitLearn, 2025a).

2.3.2 Balanced Random Forest (BRF)

Balanced Random Forest é uma variação do Random Forest, cujo objetivo é lidar com conjuntos de dados desbalanceados (Chen et al., 2004).

O algoritmo Random Forest constrói cada árvore a partir de uma amostra bootstrap dos dados de treinamento, ou seja, uma reamostragem aleatória com reposição, mas ao lidar com conjuntos de dados altamente desbalanceados, é possível que uma amostra bootstrap contenha poucas ou nenhuma instância da classe minoritária, o que resulta em árvores com desempenho ruim na previsão da classe minoritária (Chen et al., 2004).

Inspirado nisso, foi proposta uma abordagem que constrói um conjunto de árvores a partir de dados balanceados por down-sampling, denominada Balanced Random Forest (BRF) (Chen et al., 2004).

O BRF executa uma amostragem bootstrap balanceada em cada iteração e, em seguida, seleciona aleatoriamente, com reposição, o mesmo número de amostras da classe majoritária (ImbalancedLearn, 2024). Cada árvore de decisão é construída utilizando os dados balanceados e cresce até o tamanho máximo, sem poda, através da aplicação do algoritmo CART com seleção aleatória de variáveis em cada divisão (`mtry`) (Chen et al., 2004).

Essa estratégia tem como objetivo reduzir o viés causado pelo desbalanceamento, melhorando o desempenho na predição da classe minoritária sem recorrer ao oversampling, e mitigando a perda de informação normalmente associada ao downsampling tradicional (Chen et al., 2004).

2.3.3 Estratégias de Validação

Validar um modelo é uma etapa importante para avaliar a capacidade de predição do mesmo (Cunha, 2019). A validação cruzada, por exemplo, é uma técnica que avalia a capacidade de generalização de um modelo. Ela divide os dados em diferentes subconjuntos e, iterativamente, treina e testa o modelo usando diferentes combinações (Lucas, 2024).

Encontrar o equilíbrio ideal entre um modelo que se ajusta excessivamente aos dados de treinamento (overfitting) e um modelo que não captura os padrões subjacentes dos dados (underfitting) é um desafio constante do aprendizado de máquina, e a validação cruzada é uma técnica capaz de suavizar esses problemas, pois permite uma avaliação mais robusta e confiável do desempenho de um modelo (Lucas, 2024).

Um exemplo de validador cruzado é o K-Fold. Nele, o conjunto de dados é dividido em K subconjuntos, chamados de folds. A cada iteração, um fold é usado como conjunto de validação e os $K-1$ restantes são usados como conjunto de treinamento (ScikitLearn, 2025b; Lucas, 2024).

Em cada uma das K iterações, são calculadas métricas para medir a performance, e no fim, uma média delas é feita para obter uma estimativa final do desempenho do modelo. Se K for igual a n , ou seja, o número de folds é igual ao total de amostras, significa que é uma estratégia Leave One Out, onde o conjunto de dados é dividido de forma que cada fold é usado uma vez como conjunto de validação, enquanto todos os outros são usados como conjunto de treinamento (ScikitLearn, 2025b; Lucas, 2024).

Essa estratégia de Leave-One-Out Cross-Validation é comum em análise de dados biológicos e médicos, especialmente para conjuntos de dados pequenos (Lucas, 2024).

Outros exemplos de técnicas são: Holdout, onde o conjunto de testes é dividido aleatoriamente em um subconjunto de treinamento e um de teste, e o modelo de machine learning é treinado usando apenas o conjunto de treinamento; Stratified K-Fold Cross-Validation, que é similar ao K-fold, com a diferença na divisão dos dados, pois divide aleatoriamente o conjunto em K partes, e a estratificação é que garante a preservação da proporção das classes em cada fold durante a divisão; entre outras (Lucas, 2024; ScikitLearn, 2025b).

2.3.4 Métricas de Desempenho

Avaliar a qualidade de um modelo de machine learning é fundamental para entender sua capacidade de acerto e erro. Para isso, utilizam-se funções matemáticas que quantificam seu desempenho de acordo com o objetivo da tarefa (Filho, 2018).

O módulo sklearn.metrics, por exemplo, possui diversas implementações de funções de perda, pontuação e outras funções auxiliares para medir o desempenho da classificação (ScikitLearn, 2025c). As métricas, que são o resultado do cálculo destas funções, são utilizadas em tarefas de classificação binária, de múltiplas classes e de múltiplos labels, sempre com o objetivo de prever a qual categoria uma amostra pertence (Filho, 2018).

A Matriz de Confusão é uma representação que permite visualizar a distribuição das previsões feitas pelo modelo em relação aos casos reais como na Figura 2.3 e, com ela, é possível calcular algumas das métricas mais populares para classificação como acurácia, precisão,

sensibilidade, especificidade e F1-score (Awari, 2023). As definições de cada uma dessas métricas serão apresentadas a seguir.

		Classe esperada	
		Nódulo	Não nódulo
Classe prevista	Nódulo	TP	FP
	Não nódulo	FN	TN

Figura 2.3: Matriz de Confusão. TP, TN, FP e FN são, respectivamente, verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Fonte: A autora (2025).

- **Acurácia:** É o número de acertos dividido pelo número total da amostra. Pode ser calculado usando a função `accuracy_score` do ScikitLearn e é representado pela equação 2.3 (Pádua, 2020). Um modelo perfeito teria zero falsos positivos e zero falsos negativos, resultando em uma acurácia de 1.0 ou 100% (GoogleDeveloperProgram, sd).

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

- **Precisão:** É número de exemplos que o modelo previu que pertenciam a uma classe e realmente pertenciam, ou seja, os verdadeiros positivos, dividido pelo total de exemplos positivos previstos daquela classe. A função `precision_score` do ScikitLearn auxilia no cálculo desta métrica. A equação que representa esse cálculo é mostrado em 2.4 (Filho, 2018; Pádua, 2020). Um modelo perfeito teria zero falsos positivos resultando em uma precisão de 1.0 (GoogleDeveloperProgram, sd).

$$\text{Precisão} = \frac{TP}{TP + FP} \quad (2.4)$$

- **Sensibilidade (*Recall*):** Conhecido também como TPR (*True Positive Rate*), é a proporção de amostras positivas que foram classificadas corretamente como positivas, ou seja, de todos os exemplos realmente positivos, quantos ele acertou (GoogleDeveloperProgram, sd; Filho, 2018). Sua definição matemática é mostrada na figura 2.5 e ele pode ser calculado usando a função `recall_score` do ScikitLearn (Pádua, 2020). Um modelo perfeito teria zero falsos negativos, resultando em um recall de 1.0, que representaria uma taxa de detecção de 100% (GoogleDeveloperProgram, sd).

$$\text{recall} = \frac{TP}{TP + FN} \quad (2.5)$$

- **Especificidade:** Também conhecido como TNR (*True Negative Rate*), é a proporção de todos os negativos reais que foram classificados incorretamente como positivos (GoogleDeveloperProgram, sd). Representado matematicamente pela equação 2.6, pode ser calculado com auxílio da biblioteca *confusion_matrix* do ScikitLearn que conta a incidência de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, assim, após extrair essa contagem, basta usar a representação matemática para efetuar o cálculo (ScikitLearn, 2025c; de Medicina Laboratorial, 2024).

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

- **F1-score:** É a média harmônica entre a precisão e o recall, mostrada na equação 2.7 (Pádua, 2020). Pode ser útil quando temos um dataset com classes desproporcionais e queremos que o recall e a precisão estejam em equilíbrio (Filho, 2018). É possível usar a função *f1_score* do ScikitLearn para calcular esta métrica.

$$\text{F1-score} = 2 * \frac{\text{precisão} * \text{recall}}{\text{precisão} + \text{recall}} \quad (2.7)$$

- **Desvio padrão:** É uma forma de expressar a variabilidade dos dados mesmo que estejam em escalas diferentes. Para isso, utiliza-se o coeficiente de variação (CV), dado pela fórmula 2.8 que analisa a dispersão em termos relativos e é expresso em porcentagem. Quanto menor for o valor do coeficiente de variação, menor será a dispersão em torno da média. De forma geral, CV menor ou igual a 15% significa que os dados tem baixa dispersão, CV entre 15% e 30% representa uma dispersão média e acima de 30% é uma alta dispersão (Rigonatto, sd).

$$\text{CV} = \frac{\text{desvio padrão}}{\text{média dos dados}} * 100 \quad (2.8)$$

3 METODOLOGIA

Para analisar os resultados da classificação de imagens tomográficas utilizando características radiômicas, obtidas pelos modelos SVM e Random Forest, as etapas descritas a seguir foram realizadas.

3.1 BASE DE DADOS

Utilizando a base de dados LUNA16 (LUNG Nodule Analysis 2016), que disponibiliza publicamente imagens de tomografia de tórax. A base de dados contém 888 exames de tomografia computadorizada provenientes da base de dados LIDC-IDRI, também acessível ao público. Os exames incluídos no LUNA16 têm espessura de corte inferior a 2,5 mm e estão organizados em 10 subconjuntos, preparados para a validação cruzada na análise dos resultados (Van Ginneken e Jacobs, 2020).

Além disso, os dados fornecidos pelo co-orientador Fernando, utilizados por ele em sua tese "*Classifier Ensemble Based on Computed Tomography Attenuation Patterns for Computer-Aided Detection System*" (Pereira et al., 2021), forneceram as coordenadas de corte para criação de sub-imagens, que possibilitaram que a quantidade de amostras fosse de 888 para 4986.

3.2 CANDIDATOS A NÓDULO

Para selecionar os candidatos a nódulo, foi necessário obter as informações contidas no arquivo "annotations.csv" da base de dados do LUNA16, cujo arquivo disponibiliza as coordenadas dos nódulos $\geq 3\text{mm}$ (Van Ginneken e Jacobs, 2020). A partir disso, foi calculada a distância euclidiana entre as coordenadas das imagens pré-corte e as coordenadas do arquivo de anotações, definindo como verdadeira a seguinte lógica: se a distância for < 10 , a imagem é um candidato a nódulo.

3.3 PRÉ-PROCESSAMENTO

Nesta etapa, acontece a normalização dos dados, suavização e segmentação das imagens usando funções da interface Sitk.

A normalização foi realizada usando a função `RescaleIntensity`, que reescala a intensidade dos pixels na imagem para o intervalo $[0, 255]$ e, em seguida, a função `Cast` converte a imagem para o tipo de pixel `UInt8`, fazendo com que os valores dos pixels também fiquem entre 0 e 255.

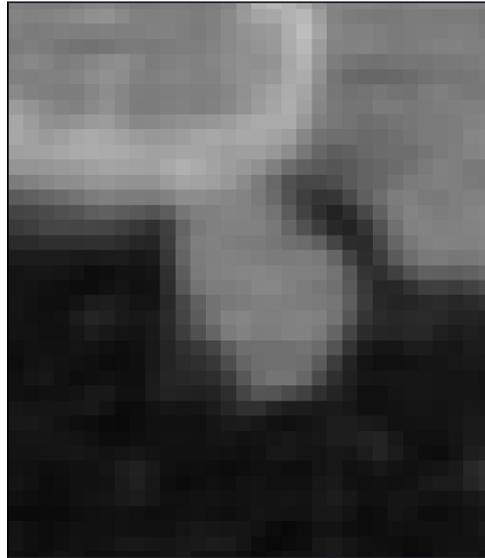


Figura 3.1: Visualização de um corte bidimensional de uma imagem normalizada. Fonte: A autora (2025).

A suavização da imagem é feita aplicando um filtro Gaussiano recursivo, o `SmoothingRecursiveGaussian`, que aplica recursivamente um desfoque (Gaussian Blur) na imagem.

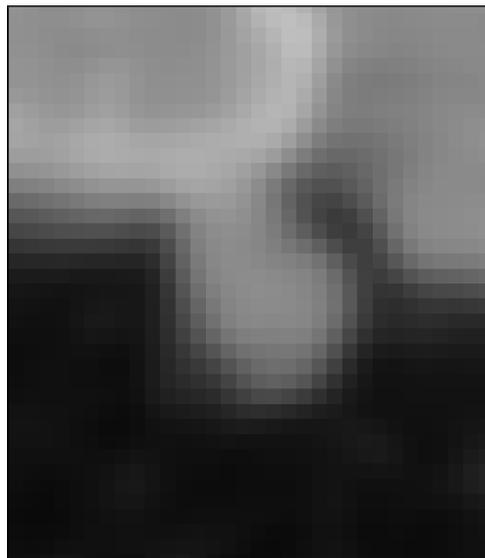


Figura 3.2: Imagem 3.1 com Gaussian Blur. Fonte: A autora (2025).

Por fim, esta imagem foi segmentada usando o filtro de Limiarização de Otsu do Sitk, o `OtsuThresholdImageFilter`, que resulta em uma imagem binária onde o objeto está separado do fundo.

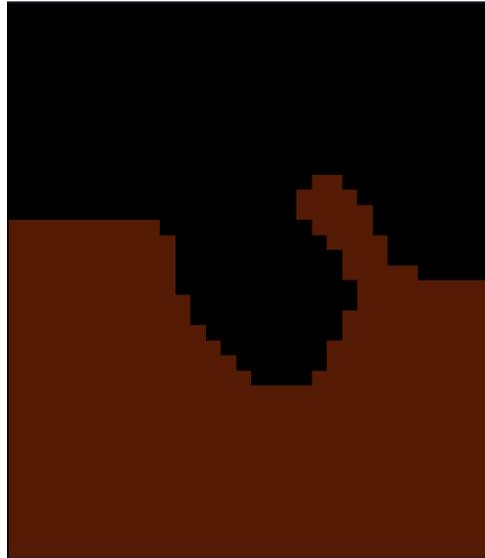


Figura 3.3: Imagem 3.1 após aplicação da limiarização de Otsu, com fundo em vermelho e objeto em preto. Fonte: A autora (2025).

A Figura 3.4 mostra os cortes de uma imagem durante três fases do processamento de uma imagem: antes da aplicação dos filtros, após a aplicação do filtro gaussiano e depois do Otsu. Nelas, podemos observar como cada filtro impacta a imagem: o gaussiano suaviza ruídos, deixando a imagem mais fosca, e o Otsu segmenta binariamente a imagem com base nas intensidades.

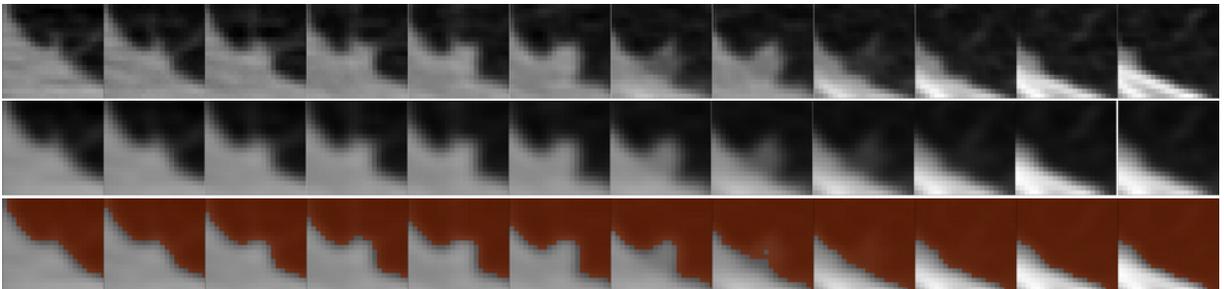


Figura 3.4: Visualização dos 12 cortes axiais ao longo do eixo z de uma imagem, comparando a imagem original, suavizada com Gaussian Blur, e segmentada por Otsu onde a parte vermelha é o fundo. Fonte: A autora (2025).

3.4 EXTRAÇÃO DE CARACTERÍSTICAS RADIÔMICAS

O primeiro passo para extrair as características radiômicas das imagens foi configurar os parâmetros do extrator. Para isso, foi obtido do GitHub o arquivo "Params.yaml" do AIM-Harvard/pyradiomics, repositório da biblioteca pyradiomics utilizada, que contém os parâmetros de configuração padrão para a extração de características radiômicas (Van Griethuysen et al., 2017). As configurações adicionais utilizadas foram "label: 1", para especificar que a região de interesse é a que foi identificada como 1 pela limiarização de Otsu, e "correctMask: True", que informa ao extrator que ele deve ajustar a máscara automaticamente para garantir que ela seja válida, como mostrado a seguir:

```

1 !wget -c https://raw.githubusercontent.com/AIM-Harvard/pyradiomics/
   master/examples/exampleSettings/Params.yaml
2 params = 'Params.yaml'
3 settings = {'label': 1, 'correctMask': True}

```

```
4 | extractor = featureextractor.RadiomicsFeatureExtractor(params,
    | additionalInfo=True, **settings)
```

Após essa etapa, a função "extractor.execute" recebe a imagem cortada original e a máscara obtida no processamento. Os resultados obtidos para cada imagem através desta extração foram armazenados em um arquivo CSV.

3.5 CLASSIFICAÇÃO

Os modelos usados para a classificação foram *Support Vector Machines* e *Balanced Random Forest*, que foram comparados para analisar o desempenho de cada um ao lidar com classes altamente desbalanceadas.

Inicialmente, foram eliminadas das colunas do arquivo CSV gerado na extração as características menos relevantes como versões, hash, e quase todas as que iniciavam com "diagnostics" e "original", pois continham mais informações sobre os dados do que as análises em si e poderiam introduzir viés ou causar *overfitting* caso fossem mantidas.

Para a validação cruzada, o método *KFold* foi utilizado, ele implementa internamente a estratégia de *Leave One Out* no conjunto de dados, e como o conjunto está dividido em 10 subconjuntos, divisão recebida desta forma e mantida conforme recomendado pela competição LUNA16, o K-fold com k=10 foi definido. Isso significa que a cada iteração um subconjunto será selecionado como conjunto de teste e o restante como conjunto de validação, garantindo que cada subconjunto seja selecionado como teste uma vez pela validação cruzada.

Após a seleção do conjunto de teste e treino, mais uma correção nas colunas do CSV foi feita, garantindo que a coluna alvo "is_candidate" fosse removida do conjunto de treinamento, e que colunas com dados não numéricos, como tuplas, tivessem seu valor quebrado e dividido em nome_da_coluna_parte1 por exemplo, removendo a coluna original em seguida.

Uma padronização foi aplicada utilizando o *StandardScaler* e os outliers foram removidos com base no *z-score*, em amostras cujas features tinham valores de z-score maiores que 3.

O número total de amostras era de 4986, sendo 840 positivas e 4146 negativas. Após a remoção de outliers, o número total de amostras foi para 3957, com 639 positivas e 3318 negativas. A projeção das amostras pode ser observada na figura 3.5, que mostra a distribuição das classes em duas dimensões usando a técnica t-SNE (*t-Distributed Stochastic Neighbor Embedding*).

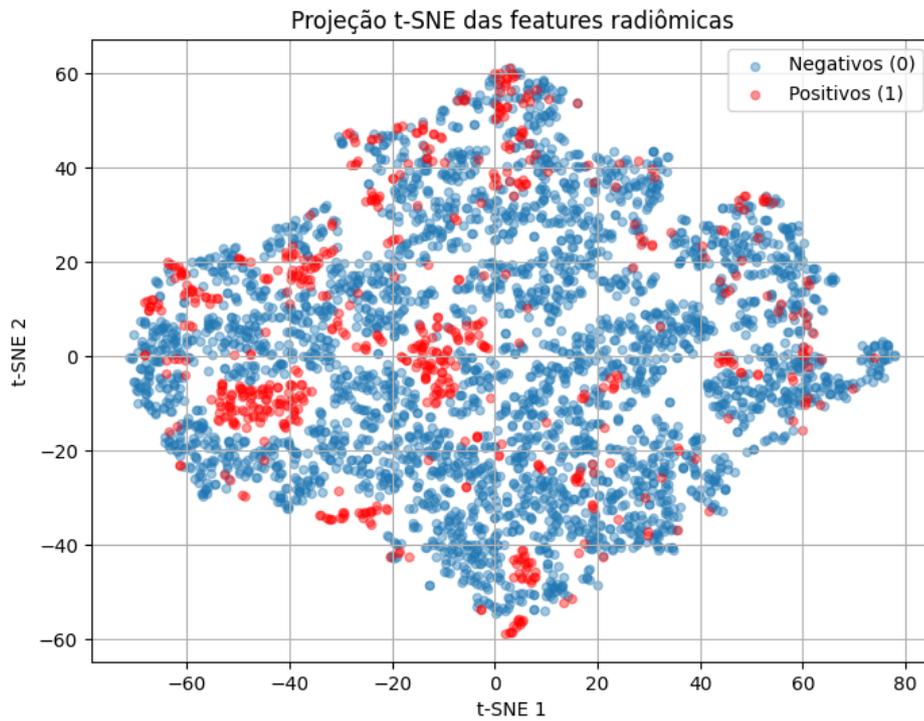


Figura 3.5: Projção t-SNE das classes positivas e negativas em duas dimensões. Fonte: A autora (2025).

A técnica de PCA (*Principal Component Analysis*) também foi aplicada, escolhendo o número de componentes de acordo com a variância explicada acumulada, mostrada na figura 3.6. A escolha foi de 14 componentes e o mapa de calor com os pesos ou cargas atribuídos a cada variável pode ser visualizado na figura 3.7. Podemos observar que algumas das características que se destacaram foram *original_shape_Elongation* e *original_glcm_Id*.

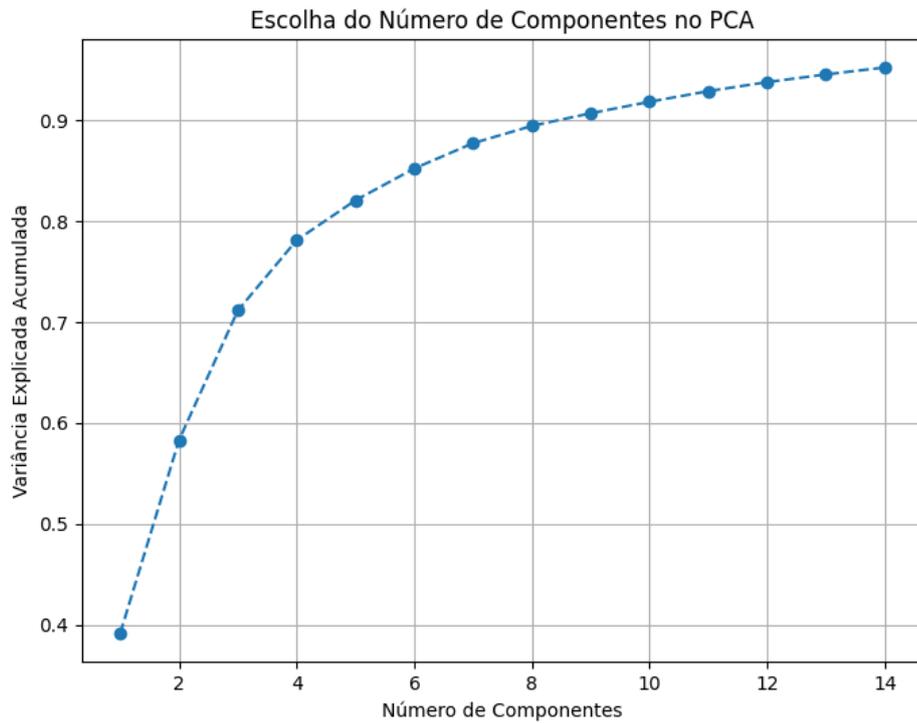


Figura 3.6: Gráfico da Variância Acumulada Explicada para escolha do número de componentes ideal para aplicação da técnica de PCA. Fonte: A autora (2025).

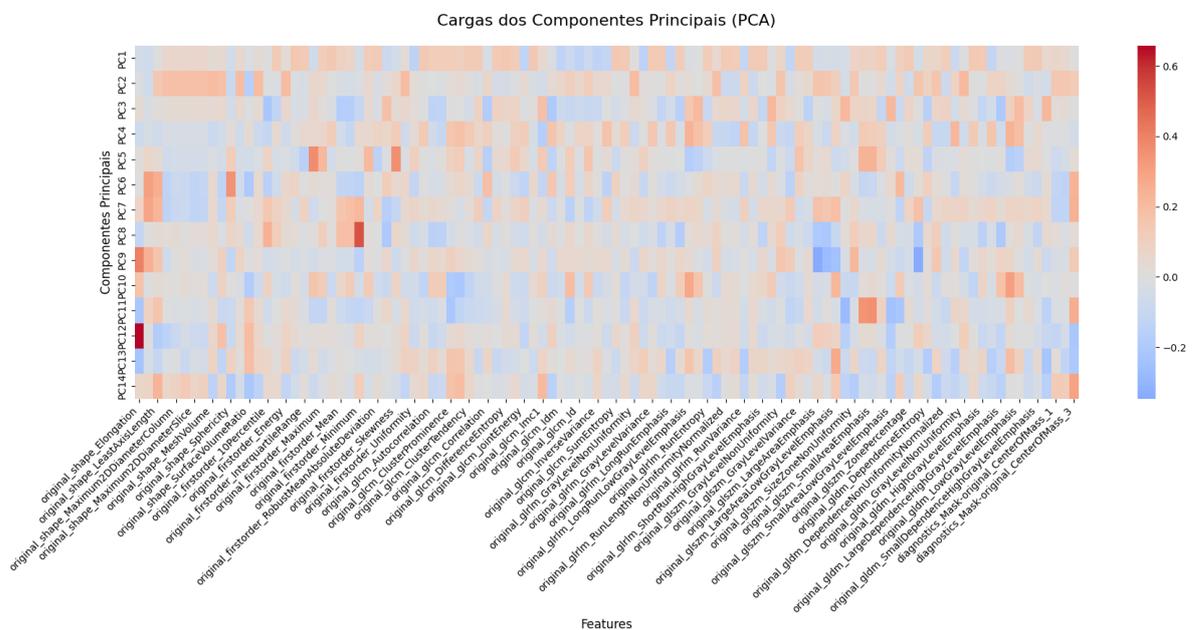


Figura 3.7: Mapa de calor com os pesos atribuídos a cada variável de cada componente principal do PCA. Fonte: A autora (2025).

Por fim, o treinamento com os modelos SVM e Random Forest Balanceado foi realizado, ambos com o parâmetro "*class_weight='balanced'*" definido, e no caso do SVM, com o *kernel* linear. As predições foram feitas e os resultados estão disponíveis na seção a seguir.

3.6 RESULTADOS

O resultado da classificação com os modelos SVM e BRF pode ser visualizado nas figuras a seguir.

SVM	Rodada 1	Rodada 2	Rodada 3	Rodada 4	Rodada 5
F1-score	0,481	0,494	0,459	0,503	0,461
Precisão	0,366	0,364	0,321	0,374	0,341
Sensibilidade	0,703	0,767	0,807	0,769	0,713
Especificidade	0,0.723	0,741	0,672	0,712	0,695

SVM	Rodada 6	Rodada 7	Rodada 8	Rodada 9	Rodada 10
F1-score	0,496	0,432	0,531	0,426	0,49
Precisão	0,348	0,310	0,402	0,306	0,353
Sensibilidade	0,861	0,716	0,783	0,703	0,8
Especificidade	0,611	0,733	0,719	0,741	0,738

Figura 3.8: Resultado das métricas de desempenho do classificador SVM em cada iteração da validação cruzada. Fonte: A autora (2025).

BRF	Rodada 1	Rodada 2	Rodada 3	Rodada 4	Rodada 5
F1-score	0,495	0,443	0,466	0,467	0,395
Precisão	0,484	0,431	0,428	0,434	0,389
Sensibilidade	0,505	0,456	0,511	0,505	0,402
Especificidade	0,878	0,884	0,869	0,852	0,86

BRF	Rodada 6	Rodada 7	Rodada 8	Rodada 9	Rodada 10
F1-score	0,505	0,419	0,532	0,403	0,506
Precisão	0,412	0,383	0,568	0,4	0,494
Sensibilidade	0,653	0,463	0,5	0,405	0,52
Especificidade	0,775	0,875	0,908	0,901	0,905

Figura 3.9: Resultado das métricas de desempenho do classificador BRF em cada iteração da validação cruzada com seu respectivo desvio padrão. Fonte: A autora (2025).

As figuras 3.8 e 3.9 mostram o resultado das métricas para cada uma das 10 iterações da validação cruzada. Cada rodada usa um subconjunto para teste e os demais para treinamento. Usando o `np.mean` do *numpy*, as médias da Figura 3.10 foram calculadas.

Média das métricas	SVM	BRF
F1-score	0,477 ± 0,031	0,463 ± 0,045
Precisão	0,349 ± 0,029	0,442 ± 0,054
Sensibilidade	0,762 ± 0,050	0,492 ± 0,067
Especificidade	0,709 ± 0,039	0,871 ± 0,036

Figura 3.10: Média das métricas de desempenho obtidas pelos classificadores SVM e BRF. Fonte: A autora (2025).

Podemos observar na Figura 3.10 cujos resultados se baseiam na média dos classificadores, o resultado do SVM apresentou maior sensibilidade (0,762), indicando que o modelo identificou melhor os casos positivos, porém seu valor de precisão ficou bem baixo (0,349), o que indica que teve muitos falsos positivos. Já o BRF apresentou maior especificidade (0,871) e uma precisão um pouco maior que o SVM (0,442), mostrando que sua prioridade era reduzir os falsos positivos, embora tenha falhado em detectar a maioria dos nódulos.

Quanto ao desvio padrão, podemos observar no resultado dos cálculos da figura 3.11 que todas as métricas avaliadas apresentaram uma variação de baixa dispersão, ou seja, abaixo de 15% (Rigonatto, sd). Isso indica que os resultados são semelhantes em cada fold da validação cruzada, e que a variabilidade dos dados não é um problema.

Cálculo Coeficiente de Variação	
SVM (%)	BRF (%)
6,5	9,72
8,3	12,22
6,56	13,62
5,5	4,13

Figura 3.11: Resultado do cálculo do coeficiente de variação para os resultados das médias das métricas de desempenho com desvio padrão. Fonte: A autora (2025).

4 CONCLUSÃO

O câncer de pulmão permanece como uma das principais causas de mortalidade no mundo, e o diagnóstico precoce continua sendo o principal fator capaz de aumentar as taxas de sobrevivência dos pacientes.

Este trabalho buscou avaliar o desempenho dos algoritmos Support Vector Machine (SVM) e Balanced Random Forest (BRF) na classificação de imagens tomográficas utilizando a radiômica para a extração de características quantitativas.

Os resultados mostraram que ambos os modelos apresentaram desempenhos limitados. O SVM apresentou uma alta sensibilidade (0,762) indicando uma melhor capacidade em identificar corretamente os casos positivos; porém, uma precisão baixa (0,349), que indica maior ocorrência de falsos positivos. O BRF atingiu uma especificidade alta (0,81) e uma precisão superior ao SVM, embora ainda baixa (0,442). Juntamente com uma menor sensibilidade (0,492), demonstrou ter dificuldade em identificar corretamente os nódulos. Dessa forma, é possível concluir que os modelos demonstraram desempenho inferior ao esperado. Um dos principais problemas observados foi que a escolha do classificador SVM com as configurações atuais não foi adequada, dadas as projeções de dimensionalidade de dados.

Fica em evidência a necessidade de enfrentar desafios persistentes, como o desbalanceamento de classes das bases de dados médicas, ao mesmo tempo que se reconhece o potencial da inteligência artificial como uma aliada do diagnóstico clínico, oferecendo suporte nas tomadas de decisão e aprimorando a precisão dos resultados.

Para trabalhos futuros, recomenda-se buscar outras abordagens que auxiliem no controle do desbalanceamento de classes, além dos modelos aqui escolhidos com o direcionamento para bases de dados escassas. Outras técnicas de preparação das imagens para aplicação da radiômica, tratamento de dados e tipo de obtenção das imagens também são relevantes.

Por fim, espera-se que os avanços na aplicação da IA na detecção de nódulos pulmonares possam auxiliar os processos diagnósticos e contribuir para a redução da mortalidade e a melhoria da qualidade de vida dos pacientes.

REFERÊNCIAS

- Arcanjo, J. (2023). Segmentação de imagens utilizando técnicas de thresholding. <https://medium.com/data-hackers/segmenta%C3%A7%C3%A3o-de-imagens-utilizando-t%C3%A9cnicas-dethresholding-1ee031562c63>. Acessado em 15/06/2025.
- Awari (2023). Quais os principais critérios da avaliação de desempenho em machine learning? <https://awari.com.br/quais-os-principais-criterios-da-avaliacao-de-desempenho-em-machine-learning/>. Acessado em 09/06/2025.
- Bôa, G. M. d. M. C. F., Saldanha, M. A. e Rodrigues, E. S. (2022). Análise de componentes principais e máquina de vetores de suporte com núcleo rbf aplicados a um sistema de posicionamento interior: Principal component analysis and support vector machine with rbf kernel applied to an indoor system positioning. *Brazilian Journal of Development*, 8(10):69910–69922.
- Chen, C., Liaw, A. e Breiman, L. (2004). Using random forest to learn imbalanced data. Relatório técnico, Department of Statistics, University of California, Berkeley. Technical Report.
- Cloudinary (2025). Gaussian blur. <https://cloudinary.com/glossary/gaussian-blur>. Acessado em 15/06/2025.
- Cunha, J. P. Z. (2019). Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. Dissertação de Mestrado, Mestrado em Ciências - Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo - SP.
- Cunningham, P. (2008). *Dimension Reduction*, páginas 91–112. Springer Berlin Heidelberg.
- Datacamp (2024a). Introduction to t-sne. <https://www.datacamp.com/tutorial/introduction-t-sne>. Acessado em 15/06/2025.
- Datacamp (2024b). A maldição da dimensionalidade no aprendizado de máquina: Desafios, impactos e soluções. <https://www.datacamp.com/pt/blog/curse-of-dimensionality-machine-learning>. Acessado em 15/06/2025.
- Datacamp (2025). Pré-processamento de dados: Um guia completo com exemplos em python. <https://www.datacamp.com/pt/blog/data-preprocessing>. Acessado em 18/06/2025.
- de Medicina Laboratorial, I. N. (2024). Entendendo os pilares avaliativos de um bom teste diagnóstico. <https://inml.com.br/sensibilidade-e-specificidade/>. Acessado em 11/06/2025.
- Ferlay J, Ervik M, L. F. L. M. C. M. M. L. P. M. Z. A. S. I. B. F. (2024). Global cancer observatory: Cancer today. <https://gco.iarc.who.int/today>. Acessado em 18/06/2025.
- Filho, M. (2018). As métricas mais populares para avaliar modelos de machine learning. <https://mariofilho.com/as-metricas-mais-populares-para-avaliar-modelos-de-machine-learning/>. Acessado em 07/06/2025.

- Gomes, P. C. T. (2019). Conheça as etapas do pré-processamento de dados. <https://www.datageeks.com.br/pre-processamento-de-dados/>. Datacamp. Acessado em 18/06/2025.
- Gomes, P. C. T. (2024). Desvio padrão: Conceito, cálculo e aplicações práticas. <https://www.datageeks.com.br/desvio-padrao/>. Datageeks. Acessado em 24/06/2025.
- GoogleDeveloperProgram (sd). Classification: Accuracy, recall, precision, and related metrics. <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=pt-br>. Sem data. Acessado em 11/06/2025.
- Guerreiro, T., Aguiar, P. e Araújo, A. (2024). Evidências atuais para um programa de triagem de câncer de pulmão. *Revista Portuguesa de Saúde Pública*, 42(2):133–158.
- IBM (sd). O que é segmentação de imagem? <https://www.ibm.com/br-pt/think/topics/image-segmentation>. Sem data. Acessado em 16/06/2025.
- ImbalancedLearn (2014-2024). Balancedrandomforestclassifier. <https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html>. The imbalanced-learn developers. Acessado em 16/06/2025.
- JATOBÁ, A. E. A. d. A. (2021). Multimodality ct/mri radiomics for lung nodule malignancy suspiciousness classification. Dissertação de Mestrado, Universidade Federal de Alagoas.
- Lucas, H. B. (2024). A importância da validação cruzada em machine learning. <https://iacomcafe.com.br/importancia-validacao-cruzada-machine-learning/>. Acessado em 12/06/2025.
- Mahmud Sujon, K., Binti Hassan, R., Tusnia Towshi, Z., Othman, M. A., Abdus Samad, M. e Choi, K. (2024). When to use standardization and normalization: Empirical evidence from machine learning models and xai. *IEEE Access*, 12:135300–135314.
- Marques, P. M. d. A. e Oliveira, L. F. d. (sd). Diagrama de processamento de imagens. Material didático do Centro de Ciências das Imagens e Física Médica (CCIFM), Faculdade de Medicina de Ribeirão Preto – FMRP/USP. Sem data. Acessado em 18/06/2025.
- Medeiros, C. J. F. d. e Costa, J. A. F. (2008). Uma comparação empírica de métodos de redução de dimensionalidade aplicados à visualização de dados. *Revista Linguagem e Instrumentação Musical (LNLN)*, 6(2).
- Nogueira-Barbosa, M. H. e Marchiori, E. (2019). Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. *Radiologia Brasileira*, 52(6):387–396.
- Oracle (sd). Z-score. https://docs.oracle.com/cloud/help/pt_BR/pbcs_common/PFUSU/insights_metrics_Z-Score.htm#PFUSU-GUID-640CEBD1-33A2-4B3C-BD81-EB283F82D879. Sem data. Acessado em 14/06/2025.
- Pereira, F. R., De Andrade, J. M. C., Escuissato, D. L. e De Oliveira, L. F. (2021). Classifier ensemble based on computed tomography attenuation patterns for computer-aided detection system. *IEEE Access*, 9:123134–123145.

- Pádua, M. (2020). Machine learning -métricas de avaliação: Acurácia, precisão e recall, f1-score. <https://medium.com/@mateuspdua/machine-learning-m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-e-recall-d44c72307959>. Acessado em 09/06/2025.
- Rigonatto, M. (sd). Coeficiente de variação. <https://mundoeducacao.uol.com.br/matematica/coeficiente-variacao.htm>. Mundo Educação. Sem data. Acessado em 24/06/2025.
- Salmi, M., Atif, D., Oliva, D. et al. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57:273.
- SCAPICCHIO, C., GABELLONI, M. e BARUCCI, A. e. a. (2021). A deep look into radiomics. *Radiol med*, 126:1296–1311.
- ScienceDirect (sd). Infinite impulse response filter. <https://www.sciencedirect.com/topics/computer-science/infinite-impulse-response-filter>. Digital Video and HD (Second Edition), 2012. Sem data. Acessado em 15/06/2025.
- ScikitLearn (2007 - 2025a). 1.4. support vector machines. <https://scikit-learn.org/stable/modules/svm.html>. Acessado em 16/06/2025.
- ScikitLearn (2007 - 2025b). 3.1. cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html. Acessado em 13/06/2025.
- ScikitLearn (2007 - 2025c). Metrics and scoring: quantifying the quality of predictions. https://scikit-learn.org/stable/modules/model_evaluation.html. Acessado em 07/06/2025.
- ScikitLearn (2007 - 2025d). StandardScaler. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>. Acessado em 14/06/2025.
- Silva, A. M. M. d., Patrocínio, A. C. e Schiabel, H. (2019). Processamento e análise de imagens médicas. *Revista Brasileira de Física Médica*, 13(1):34–48. Artigo de Revisão.
- Silva, W. F. d. (2009). Processamento de imagens médicas. Relatório de Projeto de Disciplina, IA369P - Tópicos em Engenharia de Computação VI: Visualização de Informação, Universidade Estadual de Campinas (UNICAMP). Orientadora: Profa. Wu Shin Ting. Acessado em 18/06/2025.
- SimpleITK (sda). `itk::simple::otsuthresholdimagefilter` class reference. https://simpleitk.org/doxygen/latest/html/classitk_1_1simple_1_1OtsuThresholdImageFilter.html. Sem data. Acessado em 15/06/2025.
- SimpleITK (sdb). `itk::simple::smoothingrecursivegaussianimagefilter` class reference. https://simpleitk.org/doxygen/latest/html/classitk_1_1simple_1_1SmoothingRecursiveGaussianImageFilter.html. Sem data. Acessado em 15/06/2025.
- SimpleITK (sdc). Overview. <https://simpleitk.org/TUTORIAL/>. Sem data. Acessado em 15/06/2025.

- Spring - DPI/INPE (2006). Filtragem - tutorial de geoprocessamento. <http://www.dpi.inpe.br/spring/portugues/tutorial/filtragem.html>. INPE - Instituto Nacional de Pesquisas Espaciais. Acessado em 16/06/2025.
- Torok, L. (sd). Método de otsu. Instituto de Computação – Universidade Federal Fluminense (UFF). Sem data. Disponível em: <http://profs.ic.uff.br/~aconci/OtsuTexto.pdf>. Acessado em 15/06/2025.
- Van Ginneken, B. e Jacobs, C. (2020). Luna16.
- van Griethuysen, J. J., Fedorov, A., Parmar, C. e et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107.
- Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillion-Robin, J. C., Pieper, S. e Aerts, H. J. W. L. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107.
- van Timmeren, J., Cester, D., Tanadini-Lang, S., Alkadhi, H., Baessler, B., Boellaard, R., Bogowicz, M., Buchali, A., Ciompi, F., Crimi, A., Hanbury, A., Horvat, N., Lambin, P., Leijenaar, R. T. H., Loeffler, C. U., Martinelli, M., Orhac, F., Pfaehler, E. A. G., Refaee, T., Schwyzer, H. K., Zwanenburg, A. e Donati, O. F. (2020). Radiomics in medical imaging—“how-to” guide and critical reflection. *Insights into Imaging*, 11(1):91.
- Wagner, M. W., Namdar, K., Biswas, A., Monah, S., Khalvati, F. e Ertl-Wagner, B. B. (2021). Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology*, 63(12):1957–1967.
- Warmerdam, V. D. (2024). The standardscaler is not standard. <https://papers.probabl.ai/the-standardscaler-is-not-standard>. Acessado em 14/06/2025.
- Zwanenburg, A., Vallières, M., Abdalah, M. A. e et al. (2020). The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338.