UNIVERSIDADE FEDERAL DE ALAGOAS-UFAL
CAMPUS A.C. SIMÕES
CIÊNCIA DA COMPUTAÇÃO

JOSÉ LUCAS LEITE CALHEIROS

MODEL FOR PULMONARY NODULES MACRO ENVIRONMENT CLASSIFICATION USING RADIOMIC FEATURES

MACEIÓ
2019

José Lucas Leite Calheiros

Model for pulmonary nodules macro environment classification using radiomic features

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal de Alagoas - UFAL, Campus A.C. Simões.

Orientador: Prof. Dr. Marcelo Costa Oliveira
Coorientador: Prof. Me. Lucas Benevides Viana de Amorim

Maceió
2019

José Lucas Leite Calheiros

Model for pulmonary nodules macro environment classification using radiomic features

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal de Alagoas - UFAL, Campus A.C. Simões.

Data de Aprovação: 30/08/2019

Banca Examinadora

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

Prof. Me. Lucas Benevides Viana de Amorim
UFAL - Instituto de Computação
Coorientador

Prof. Dr. Bruno Almeida Pimentel
UFAL - Instituto de Computação
Examinador

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

AGRADECIMENTOS

Gostaria de agradecer primeiramente aos meus pais, José Cícero e Leni, pelo apoio e suporte por todos esses anos.

Aos meus irmãos, Jalme e Laura, por estarem sempre presentes.

Aos três integrantes do grupo colegas que me acompanharam e me ajudaram em quase três destes quatro anos. Ao pessoal do toco++ como grandes amigos que eu conheci.

Aos colegas da 2015.1, nunca imaginei que uma turma poderia abrigar um conjunto de pessoas tão boas.

Às pessoas que conheci durante minha estadia no IM, pelos ótimos e divertidos momentos.

Ao meu orientador Marcelo, e ao meu coorientador Lucas, pela atenção e pelo tempo dedicado.

Aos professores Bruno e Thiago, por aceitarem compor minha banca.

À UFAL, onde tive a honra de estudar e conhecer pessoas fantásticas.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Alan Turing

RESUMO

O câncer pode ser descrito como um conjunto de doenças caracterizadas pelo crescimento desordenado de células que invadem e destroem tecidos e órgãos. Dentre os tipos de câncer o mais letal e mais comum é câncer de pulmão com 1,76 milhões de mortes estimadas em 2018. O diagnóstico precoce da doença representa a maior esperança de sobrevida para o paciente, a descoberta da doença em seus primeiros estágios eleva a taxa de sobrevida de 5 anos do paciente até 90%, enquanto a descoberta tardia resulta em uma taxa de até 15% para o período de 1 ano. A utilização Tomografia Computadorizada (TC) permitiu a realização do diagnóstico de câncer pulmonar de forma mais barata e menos invasiva, entretanto, a análise de imagens de tomografia computadorizada por radiologistas é uma tarefa complexa que pode ser afetada por condições como fadiga e iluminação, nessas condições, mesmo profissionais experientes estão sujeitos a erros. As Ferramentas de Auxílio Computacional (CAD) realizam o processo de classificação através de uma análise sobre atributos quantitativos da imagem, e tem a proposta de fornecer uma segunda opinião ao radiologista, tornando assim a tarefa de diagnóstico mais robusta. O uso de atributos radiômicos fornece uma grande quantidade de dados de representação da imagem, assim para o aprimoramento do modelo de classificação de sistemas CAD é necessário encontrar os atributos de maior relevância em meio aos dados. O objetivo deste trabalho foi desenvolver um modelo para predição de nódulos pulmonares utilizando atributos de diferentes regiões do macroambiente tumoral. Neste trabalho 897 nódulos pulmonares foram classificados por 15 modelos de classificação sob 3 diferentes formas de balanceamento. O modelo melhor avaliado obteve área sob a curva ROC (AUC) média de 0,915, sensibilidade de 82,2% e especificidade de 82,6%. Os resultados mostram que a utilização de atributos radiômicos extraídos da região do parênquima melhora efetivamente o desempenho de modelos preditivos para a classificação de nódulos pulmonares.

Palavras-chave: Atributos Radiômicos; Auxilio ao Diagnóstico por Computador; Câncer de Pulmão; Classificação; Nódulo Pulmonar; Parênquima.

ABSTRACT

Cancer can be described as a set of diseases characterized by disordered growth of cells that invade and destroy tissues and organs. Among the most lethal and most common types of cancer is lung cancer with an estimated 1.76 million deaths in 2018. Early diagnosis of the disease represents the patient's greatest hope for survival, the discovery of the disease in its early stages elevates the patient's 5-year survival rate is up to 90%, while late discovery results in a rate of up to 15% for the 1 year period. The use of computed tomography (CT) has made the diagnosis of lung cancer cheaper and less invasive. However, the analysis of computed tomography images by radiologists is a complex task that can be affected by conditions such as fatigue and illumination. conditions, even experienced professionals are subject to error. The Computational Assistance Tools (CAD) perform the classification process through an analysis of quantitative image features, and have the purpose of providing a second opinion to the radiologist, thus making the diagnostic task more robust. The use of radiometric features provides a large amount of image representation data, so in order to improve the CAD system classification model it is necessary to find the most relevant features in the data. The objective of this work was to develop a model for lung nodule prediction using features from different regions of the tumor macroenvironment. In this study 897 pulmonary nodules were classified by 15 classification models under 3 different forms of balancing. The best evaluated model obtained area under the ROC curve (AUC) average of 0.915, sensitivity of 82.2% and specificity of 82.6%. The results show that the use of radiomic features extracted from the parenchyma region effectively improves the performance of predictive models for lung nodule classification.

Keywords: Radiomic Features; Computer Aided Diagnosis; Lung Cancer; Ranking; Pulmonary nodule; Parenchyma.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AF	Atributos de Forma
AI	Atributos de Intensidade
AG	Algoritmo Genético
AT	Atributos de Textura
AM	Aprendizagem de Máquina
ANB	Atributos de Nitidez de Borda
AUC	Area Under the Curve
BNP	Banco de Nódulos Pulmonares
CAD	Computer-aided Diagnosis
GBM	Gradient Boost Machine
GI	Gini Importance
KNN	k-Nearest Neighbors
LIDC	Lung Image Database Consortium
RU	Random Under-sampling
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Over-sampling Technique
SMOTEENN	Synthetic Minority Over-sampling Technique + Edited Nearest Neighbor
SVM	Support Vectors Machine
TC	Tomografia Computadorizada
VP	Verdadeiros Positivos
VN	Verdadeiros Negativos
FP	Falso Positivos
FN	Verdadeiro Negativos
tfp	taxa de falsos positivos
tvp	taxa de verdadeiros positivos

SUMÁRIO

1 INTRODUÇÃO

O câncer pode ser descrito como um conjunto de doenças caracterizadas pelo crescimento desordenado de células que invadem e destroem tecidos e órgãos. Em 2018 o câncer foi causador de cerca de 9,6 milhões de mortes, sendo portanto a segunda maior causa de mortes no mundo, atrás apenas de doenças cardiovasculares (World Health Organisation, 2019). O câncer de pulmão, além de ser a forma mais frequente de câncer com 2,09 milhões de casos estimados em 2018, é também o maior causador de mortes pela doença, sendo responsável por 18,4% do total de mortes por câncer (BRAY et al., 2018).

O momento da detecção do câncer pulmonar é o principal fator para a sobrevivência do paciente. Se detectado em seus estágios iniciais a taxa de sobrevivência em 5 anos é de 70-90% (KNIGHT et al., 2017). Por outro lado quando a doença é detectada em estágios mais avançados a taxa de sobrevivência do paciente cai drasticamente, sendo reduzida a 15-19% em um 1 ano (BANNISTER; BROGGIO, 2016).

O principal método de detecção de câncer de pulmão é a tomografia computadorizada (TC). Com a recente popularização da TC houve um crescimento na detecção de nódulos pulmonares, entretanto esse crescimento não foi acompanhado pelo crescimento no diagnóstico de câncer de pulmão (GOULD et al., 2015). Alguns métodos de diagnósticos como a biópsia são bastante invasivos e caros, portanto a utilização desses métodos não se justifica como uma opção viável para a análise de todos os casos. Assim, faz-se necessário a utilização de métodos mais baratos e menos invasivos e dentre estes métodos está o diagnóstico através da análise da imagem obtida pela TC (ATWATER et al., 2016).

A tarefa de diagnosticar um paciente através da análise de imagens médicas é bastante complexa e um desafio mesmo para radiologistas experientes (AKGüL et al., 2010). Os resultados do diagnóstico são baseados em características semi-subjetivas da imagem e os radiologistas podem sofrer influência de fatores como estresse e fadiga, nessas condições, mesmo o especialista experiente está sujeito a erros (CHUQUICUSMA et al., 2018).

Ferramentas de Diagnóstico Auxiliado por Computador (CAD - Computer Aided Diagnosis) têm a proposta de fornecer apoio à decisão de diagnóstico pelo radiologista realizando uma análise quantitativa dos atributos da imagem e posteriormente a sua classificação. Os sistemas CAD tipicamente fornecem uma única resposta ao radiologista, agindo como uma segunda opinião (DILGER A. JUDISCH, 2015).

O termo radiômicos (do inglês *radiomics*) refere-se ao processo que compreende a

extração de atributos quantitativos em larga escala de uma imagem, convertendo-a em dados mineráveis, e a subsequente análise desses dados para o auxílio à tomada de decisão (GILLIES et al., 2016).

A extração de uma grande quantidade de atributos em geral também aumenta a quantidade de ruídos em meio aos dados extraídos, os ruídos não só podem ser irrelevantes para a classificação de um modelo preditivo, como também podem afetar negativamente o desempenho do modelo (ZHU; WU, 2004). No contexto de radiômicos, vários atributos quantitativos têm sido usados para caracterizar nódulos pulmonares, mas a questão de quais deste atributos quantitativos possuem maior relevância na classificação entre nódulos benignos e malignos continua em aberto (CHOI; CHOI, 2013; OLIVEIRA, 2013; DILGER A. JUDISCH, 2015).

O processo de classificação de nódulos pulmonares por ferramentas computacionais está limitado à quantidade de informações relevantes que podem ser extraídas da imagem. Na literatura, a maioria dos atributos usados para classificação são derivados apenas da análise da região do nódulo pulmonar, porém, trabalhos recentes indicam que a área em torno do nódulo, o parênquima, pode possuir informações relevantes para a classificação (DILGER A. JUDISCH, 2015; DILGER et al., 2015; FILHO et al., 2016). Assim faz-se necessário verificar o efeito que as informações derivadas dessa região têm sobre o processo de classificação, e quais seriam os seus atributos mais relevantes.

Outro fator que pode afetar o desempenho de classificação de um modelo preditivo é o desbalanceamento de seus dados (YEN; LEE, 2006). E em geral bases de dados de nódulos pulmonares possuem prevalência de nódulos benignos, alguns trabalhos obtiveram bons desempenhos com técnicas de balanceamento na tarefa de reconhecimento de nódulos pulmonares (MEHRE SUDIPTA MUKHOPADHYAY, 2016; SUI et al., 2015a).

A classificação é a etapa de decisão do sistema CAD, o conjunto de atributos é utilizado para a geração de um modelo preditivo que atribui uma probabilidade de malignidade do nódulo. Diversos algoritmos de aprendizagem de máquina vem sendo amplamente usados na literatura para classificação de nódulos pulmonares (FILHO et al., 2016; REEVES et al., 2016; DILGER et al., 2015; TAGHAVI et al., 2010; TARTAR et al., 2013; Kaucha et al., 2017).

1.1 OBJETIVOS

O objetivo principal deste trabalho foi desenvolver um modelo para a classificação de nódulos pulmonares usando atributos radiômicos extraídos do macroambiente tumoral com o intuito de verificar a importância do uso de atributos provenientes da região do parênquima para

a melhora de desempenho no processo de classificação do câncer pulmonar.

Como objetivos secundários investigamos o impacto de diferentes técnicas de balanceamento na classificação dos nódulos, também verificamos o impacto na classificação causado por diferentes conjuntos de atributos extraídos do macroambiente tumoral, e por fim, analisamos a importância dos atributos para a classificação.

1.2 ESTRUTURA DO TRABALHO

- **Capítulo 2 - Fundamentação Teórica:** apresenta os principais conceitos que fundamentam este trabalho;
- **Capítulo 3 - Materiais e Métodos:** descreve a metodologia empregada na preparação e execução dos experimentos;
- **Capítulo 4 - Resultados e Discussão:** apresenta e discute os resultados obtidos pelos experimentos e os compara com a literatura;
- **Capítulo 5 - Conclusão:** apresenta as conclusões obtidas pelo trabalho.

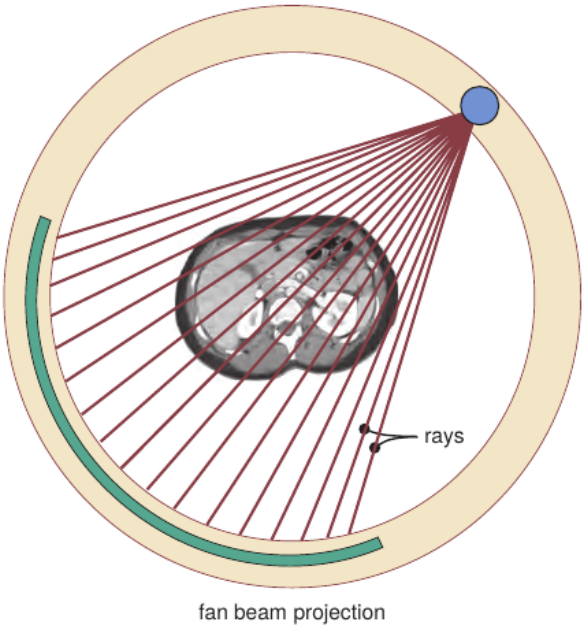
2 FUNDAMENTAÇÃO TEÓRICA

2.1 NÓDULOS PULMONARES EM IMAGENS DE TOMOGRAFIA COMPUTADORIZADA

A Tomografia Computadorizada (TC) é um processo no qual o paciente é envolto à um equipamento tubular enquanto é submetido a uma fonte emissora de raios X. A fonte é rotacionada durante a realização do exame e emite uma série de pulsos estreitos de raios X (Figura 1). Os raios são detectados por um detector digital localizado na direção oposta da fonte e transmitidos para um computador que utiliza técnicas matemáticas para construir uma fatia de imagem 2D do paciente. Esse processo continua até obter-se o número desejado de fatias.

Em geral as fatias 2D construídas durante o exame são montadas em volumes, mantendo-se um espaçamento e permitindo que seja gerada uma imagem 3D das estruturas escaneadas.

Figura 1 – Modelo de emissão de raios X em um tomógrafo.



Fonte: adaptado de (JERROLD et al., 2013)

A TC é a principal técnica de imagem para o diagnóstico do câncer de pulmão (JERROLD et al., 2013; Diciotti et al., 2010). Sua aplicação na investigação do câncer pulmonar forneceu uma nova perspectiva quando se comparada ao exame de Raio X digital, a TC fornece imagens 3D de alta resolução permitindo uma visualização nódulo mais ampla com maior quantidade de informações, possibilitando assim uma maior sensibilidade na detecção de nódulos.

A TC realiza o imageamento de diversas estruturas do corpo de diferentes densidades, possibilitando assim sua utilização na detecção de câncer pulmonar. A detecção da doença

em seus primeiros estágios é imprescindível para o aumento da taxa de sobrevida do paciente (KNIGHT et al., 2017), e graças a popularização da TC houve um aumento na detecção de nódulos pequenos, embora que, este crescimento não foi acompanhado pelo aumento no número de diagnósticos de câncer de pulmão (GOULD et al., 2015). O diagnóstico desses nódulos é de suma importância, e por conta da alta taxa de falso positivos, faz-se necessário a utilização de métodos de classificação mais baratos e menos invasivos (ATWATER et al., 2016), e dentre eles a utilização da TC para o auxilio ao diagnóstico.

2.2 BANCO DE NÓDULOS PULMONARES

2.2.1 LIDC-IDRI

O LIDC-IDRI (ARMATO et al., 2011) é um repositório público de imagens de TC de câncer pulmonar. Atualmente o repositório contém 1.018 exames de TC realizados em 1.010 pacientes e 244.559 imagens de TC de tórax. O LIDC-IDRI também contém dados de exames, marcações e classificações de nódulos pulmonares identificados por quatro radiologistas experientes. As classificações dos nódulos foram realizadas segundo algumas características subjetivas, dentre elas a probabilidade de malignidade em cinco níveis.

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

2.2.2 Banco de Nódulos Pulmonares

O Banco de Nódulos Pulmonares (BNP) foi desenvolvido no Laboratório de Telemedicina e Informática Medica (LaTIM) com o objetivo de organizar os dados do repositório LIDC-IDRI em um esquema de banco de dados. O BNP foi organizado por Ferreira Junior et al. (JUNIOR et al., 2016) e utiliza uma abordagem NoSQL orientada a documentos com o MongoDB.

Para evitar redundâncias, apenas a leitura do radiologista que identificou maior quantidade de nódulos do LIDC-IDR foi armazenada. Ao todo o BNP possui 752 exames e 1944 nódulos pulmonares. Dentro do banco de dados estão organizadas as marcações de cada nódulo realizadas

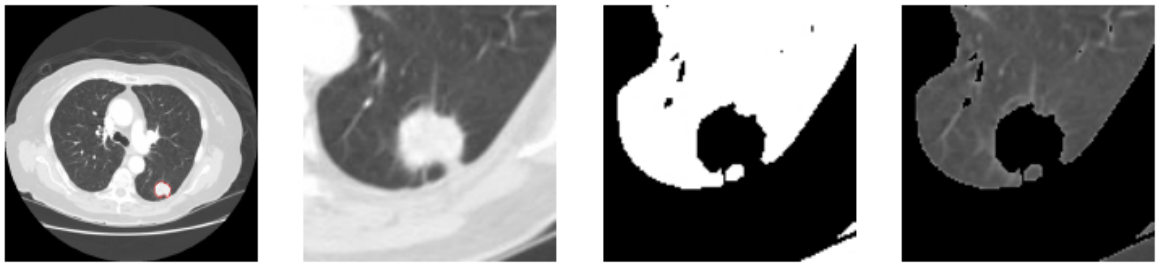
pelo radiologista, juntamente com suas imagens, e classificações segundo as 5 probabilidades de malignidade.

2.2.3 Atributos Radiômicos

O termo radiômicos, do inglês *radiomics*, refere-se ao processo no qual é realizada a extração de um grande volume de características quantitativas de imagens médicas, convertendo a imagem em dados mineráveis, e a subsequente análise desses dados para o auxilio da tomada de decisão (GILLIES et al., 2016).

A extração dos atributos utilizados neste trabalho foi realizada por Ferreira Júnior (JÚNIOR et al., 2015) e Lima Filho (FILHO et al., 2016) utilizando segmentações de nódulos e parênquimas. As segmentações dos nódulos utilizadas já estavam contidas no BNP (JUNIOR et al., 2016). As segmentações dos parênquimas foram realizadas de forma automática e armazenadas no BNP, a quantidade de parênquima incluído correspondeu a duas vezes o diâmetro do nódulo (FILHO et al., 2016). A Figura 2 ilustra o processo de segmentação.

Figura 2 – Ilustração do processo de segmentação do parênquima pulmonar.



(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.

Fonte: (FILHO et al., 2016)

As categorias de atributos extraídas correspondem a Atributos de Intensidade 3D (AI), Atributos de Forma 3D (AF), Atributos de Textura 3D (AT) e Atributos de Nitidez de Borda 3D (ANB). A Tabela 1 mostra as categorias de atributos e a região onde foram extraídas.

2.2.3.1 Atributos de Intensidade 3D

Os AI representam informações referentes ao histograma do valor de intensidade em escala de cinza dos pixeis das áreas segmentadas em todas as fatias da segmentação. Seguem

Tabela 1 – Categorias de atributos e regiões de extração. Os atributos de nitidez de borda são extraídos da área comum entre nódulo e parênquima.

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

abaixo as equações utilizadas para extração dos atributos.

$$\text{Energia} = \sum_{i=1}^{n} x_i^2, \tag{1}$$

$$\text{Intensidade média } (\bar{x}) = \sum_{i=1}^{n} x_i^2, \tag{2}$$

$$\text{Intensidade mediana}, \tag{3}$$

$$\text{Intensidade mínima } (I_m), \tag{4}$$

$$\text{Intensidade máxima } (I_M), \tag{5}$$

$$\text{Entropia} = -\sum_{k=1}^{N} p(x_k) log_2(p(x_k)), \tag{6}$$

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)s^4}, \tag{7}$$

$$\text{Skewness} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)s^3}, \tag{8}$$

$$\text{Desvio médio absoluto} = \frac{1}{n}\sum_{i=1}^{n} \left|x_i - \overline{x}\right|, \tag{9}$$

$$\text{Range} = \left|I_M - I_m\right|, \tag{10}$$

$$\text{Raiz quadrada média} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}, \tag{11}$$

$$\text{Desvio padrão} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{12}$$

$$\text{Uniformidade} = \sum_{k=1}^{N} p(x_k)^2, \tag{13}$$

$$\text{Variância} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{14}$$

onde x_i representa o valor de intensidade do i-ésimo pixel da imagem; *s* é o desvio padrão da intensidade; *n* é o total de pixeis da região.

Ao todo foram extraídos 28 atributos AI por Lima Filho (FILHO et al., 2016), 14 originados da segmentação do nódulo, e 14 obtidos a partir da segmentação do parênquima, a escolha dos atributos foi sugerida por Dilger (DILGER, 2013).

2.2.4 Atributos de Forma 3D

Os AF representam informações referentes às características geométricas da área segmentada. Seguem abaixo as equações utilizadas para extração dos atributos.

$$\text{Compacidade 1} = \frac{V}{\sqrt{\pi}A^{\frac{2}{3}}}, \tag{15}$$

$$\text{Compacidade 2} = 36\pi\frac{V^2}{A^3}, \tag{16}$$

$$\text{Desproporção esférica} = \frac{A}{4\pi R^2}, \tag{17}$$

$$\text{Esfericidade} = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}, \tag{18}$$

$$\text{Área (A)} = \sum^{i\in f} p_i, \tag{19}$$

$$\text{Área da superfície} = 4\pi R^2, \tag{20}$$

$$\text{Relação superfície volume} = \frac{A}{V}, \tag{21}$$

$$\text{Volume (V)} = (\sum^{f\in F}\sum^{i\in f} p_i) \cdot EspessuraDeF, \tag{22}$$

$$\text{Raio da esfera com volume V} = \sqrt[3]{\frac{3V}{4\pi}}, \tag{23}$$

$$\text{Diâmetro} = \max_{f\in F}(\max x_i - \min x_i, \max y_i - \min y_i) \tag{24}$$

onde p_i representa o i-ésimo pixel da fatia da segmentação; f representa o conjunto de pixeis (p) da segmentação de uma fatia; x e y são as coordenadas cartesianas para cada pixel i; F representa o conjunto total de fatias (f) do nódulo; $EspessuraDeF$ é o tamanho do *voxel* (pixel tridimensional).

Os AF ao todo contabilizam 9 atributos, que foram extraídos por Lima Filho (FILHO et al., 2016) através de adaptações das implementações propostas por Aerts (AERTS et al., 2014).

2.2.5 Atributos de Textura 3D

Os AT contém informações referentes ao aspecto e repetição de padrões em diversas regiões do nódulo e suas variações. Seguem abaixo as equações utilizadas para extração dos atributos.

$$\text{Energia} = \sum_i \sum_j C^2(i,j), \tag{25}$$

$$\text{Entropia} = -\sum_i \sum_j C(i,j) log C(i,j), \tag{26}$$

$$\text{Momento da diferenca inverso} = \sum_i \sum_j \frac{C(i,j)}{1+(i-j)^2}, \tag{27}$$

$$\text{Contraste} = \sum_i \sum_j (i-j)^2 C(i,j), \tag{28}$$

$$\text{Variância} = \sum_i \sum_j (i-\mu)^2 C(i,j), \tag{29}$$

$$\text{Matiz} = \sum_i \sum_j (i+j-\mu_x-\mu_y)^3 C(i,j), \tag{30}$$

$$\text{Proeminência} = \sum_i \sum_j (i + j - \mu_x - \mu_y)^4 C(i,j), \tag{31}$$

$$\text{Correlação} = \sum_i \sum_j \frac{(i - \mu_x)(j - \mu_y)}{\sqrt{\sigma_x \sigma_y}} C(i,j), \tag{32}$$

$$\text{Homogeneidade} = \sum_i \sum_j \frac{C(i,j)}{1 + |i - j|}, \tag{33}$$

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:

$$C_x(i) = \sum_j C(i,j), \tag{34}$$

$$C_y(j) = \sum_i C(i,j), \tag{35}$$

$$\mu_x = \sum_i i C_x(i), \tag{36}$$

$$\mu_y = \sum_j j C_y(j), \tag{37}$$

$$\sigma_x = \sum_i (i - \mu_x)^2 \cdot \sum_j C(i,j), \tag{38}$$

$$\sigma_y = \sum_j (j - \mu_y)^2 \cdot \sum_i C(i,j), \tag{39}$$

Os AT foram escolhidos por Ferreira Junior (JÚNIOR et al., 2015) segundo sugestão de Haralick (Haralick et al., 1973). Os atributos foram obtidos através da utilização de uma matriz de co-ocorrência, uma técnica estatística de segunda ordem, ou seja obtém informações relacionadas aos pixeis da imagem. Os AT foram obtidos a partir da aplicação dos resultados das equações 25-33 à matriz de co-ocorrência nas orientações 0°, 45°, 90° e 135°, e distância de 1 *voxel*, o que resultou em um conjunto formado por 36 atributos de textura extraídos do nódulo.

A extração dos AT foi primeiramente realizada por Ferreira Júnior (JÚNIOR et al., 2015) sobre a região dos nódulos, posteriormente os AT da região do parênquima foram extraídos por Lima Filho (FILHO et al., 2016).

Ao todo foram extraídos 72 AT, 36 referentes a região do nódulo, e 36 referentes a região do parênquima.

2.2.6 Atributos de Nitidez de Borda 3D

Os ANB são importantes para identificar o potencial de malignidade visto que tumores de câncer crescem em tecidos vizinhos aos nódulos (LEVMAN; MARTEL, 2011). Seguem abaixo as equações dos ANB.

$$\text{Diferença entre os extremos} = x_n - x_1, \tag{40}$$

$$\text{Soma dos valores} = \sum_{i=1}^{n} x_i, \tag{41}$$

$$\text{Soma dos quadrados} = \sum_{i=1}^{n} x_i^2, \tag{42}$$

$$\text{Soma dos logs} = \sum_{i=1}^{n} log x_i, \tag{43}$$

$$\text{Média aritmética } (\mu) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{44}$$

$$\text{Média geométrica} = \sqrt[n]{\prod_{i=1}^{n} x_i}, \tag{45}$$

$$\text{Variância da população} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{46}$$

$$\text{Variância da amostra } (v) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{47}$$

$$\text{Desvio padrão } (s) = \sqrt{v}, \tag{48}$$

$$\text{Medida de kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{s^4}, \tag{49}$$

$$\text{Medida de skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{s^3}, \tag{50}$$

$$\text{Segundo momento central} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}{s^2}, \tag{51}$$

onde x é o valor de intensidade do vetor de pixeis 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 (FILHO et al., 2016).

Ferreira Júnior (JÚNIOR et al., 2015) utilizou 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 à borda do nódulo, para então extrair atributos estatísticos sobre as intensidades dos pixeis dessas linhas ortogonais em todo volume da imagem. Ao todo foram extraídos 12 ANB.

2.3 REDUÇÃO DE DIMENSIONALIDADE

Em aprendizagem de máquina a extração e utilização de uma grande quantidade de atributos pode ser justificada pela maior capacidade de representação e quantidade de informações disponíveis para a classificação. Entretanto uma grande quantidade de atributos também aumenta a probabilidade de ocorrência de atributos irrelevantes ou ruído. A presença dessas características numa base de dados acarreta no aumento da dificuldade de aprendizagem e na perda de desempenho de classificação, principalmente de modelos mais sensíveis (NETTLETON et al., 2010).

As técnicas de redução de dimensionalidade visam selecionar os atributos mais relevantes, diminuindo ou removendo o ruído e atributos de menor importância.

2.3.1 Algoritmo Genético

O Algoritmo Genético (AG) é um método de otimização amplamente utilizado (YANG; HONAVAR, 1998; Il-Seok Oh et al., 2004; LEARDI et al., 1992) que se mostra eficiente na busca de soluções ótimas, ou próximas das ótimas.

O AG utiliza técnicas baseadas em conceitos da biologia evolutiva tais quais seleção natural, hereditariedade, mutação e cruzamento. O funcionamento do AG começa por meio da geração aleatória de uma população de indivíduos, cada indivíduo sendo composto por um determinado número de cromossomos, cada cromossomo é uma representação de uma característica a ser otimizada, em geral possuindo valor binário indicando se a característica deve ser mantida (1), ou não (0).

Após a população ser iniciada cada individuo passa por uma função de avaliação (do inglês *fitness*), os valores resultantes dessa avaliação são utilizados para identificar os indivíduos

com maior aptidão, os indivíduos então passam por um processo de seleção, que mantém tradicionalmente uma alta porcentagem dos melhores indivíduos.

O resultado da seleção então é utilizado para compor uma nova população de indivíduos através do cruzamento (criação de uma nova geração). A nova geração é composta pelos indivíduos selecionados da geração anterior juntamente com os indivíduos resultantes do cruzamento. O cruzamento é realizado através da troca aleatória de diversos conjuntos de cromossomos pertencentes aos indivíduos pais, e durante este processo podem ocorrer trocas aleatórias de valores de determinados cromossomos (mutação), cada cruzamento resulta por fim em dois novos indivíduos.

O processo de seleção e geração da nova população é repetido até que se atinja algum objetivo desejado, seja por número de gerações ou valor alcançado.

2.4 BALANCEAMENTO DE BASE

O problema do desbalanceamento ocorre quando, dentro de uma base de dados há uma predominância de uma classe sobre a outra, ou seja, certa classe A possui consideravelmente mais amostras do que a classe B. Tipicamente o problema do desbalanceamento pode afetar negativamente a classificação priorizando a acurácia de uma única classe denegrindo o desempenho da classificação de outras classes.

As três principais abordagens para realizar o balanceamento de uma base de dados são, o *over-sampling* que consiste no aumento da base de dados, seja pela repetição de indivíduos da classe minoritária da base ou gerando novos indivíduos artificialmente, o *under-sampling* que consiste na remoção de indivíduos da classe dominante, nivelando pelo valor da classe minoritária, e a combinação das duas abordagens expostas anteriormente.

2.4.1 Synthetic Minority Over-sampling

Synthetic Minority Over-sampling Technique (SMOTE) é uma técnica de *over-sampling* em classes minoritárias que utiliza conceitos de proximidade e vetores para a criação de amostras sintéticas.

O funcionamento do SMOTE é descrito no Algoritmo 1.

Algoritmo 1: Pseudo código SMOTE

Entrada: Base desbalanceada, K
Saída: Base balanceada
início
  repita
    Selecione uma amostra minoritária A da base;
    Encontre os K vizinhos minoritários mais próximos de A;
    Escolha um vizinho B dentre os K;
    $AB = B - A$ (AB é o vetor partindo de A até B);
    $Na = AB * random(1) + A$ (Na é a nova amostra sintética);
    Adicione Na à base;
  até *base balanceada*;
fim

2.4.2 Random Under-sampling

Random Under-sampling (RU) é uma técnica de *under-sampling* na classe majoritária. O RU realiza descartes de amostras aleatórias na classe majoritária até que o número de amostras minoritárias seja igual ao número de amostras majoritárias.

2.4.3 SMOTEENN

Synthetic Minority Over-sampling Technique + Edited Nearest Neighbor Rule (SMOTEENN) é uma técnica que combina a técnica de *over-sampling* SMOTE com a regra *Edited Nearest Neighbor* proposta por Wilson (Wilson, 1972).

O objetivo do SMOTEENN além de balancear a base de dados é realizar uma limpeza profunda de dados (BATISTA et al., 2004). O SMOTEENN inicialmente funciona da mesma forma que o SMOTE, após realizar o povoamento de amostras sintéticas, a segunda parte do algoritmo se inicia. Seu funcinamento pode ser descrito da seguinte forma: para cada amostra E_i na base de dados seus N vizinhos mais próximos são encontrados. A classe que E_i pertence deve ser a maioria dentre os seus N vizinhos, caso contrário E_i é removido.

2.5 APRENDIZAGEM DE MÁQUINA

2.5.1 Árvore de Decisão

A Árvore de Decisão (do inglês Decision Tree) é um modelo de classificação que utiliza a estrutura de dados de uma árvore para a realização da tomada de decisão.

As Árvores de Decisão montam seus modelos de predição através da divisão de atributos em múltiplas camadas de decisões condicionais. O treinamento do modelo inicia-se pelo nó raiz, e é realizada uma escolha do atributo de divisão e do ponto de divisão, a escolha é realizada via diferentes critérios sendo um dos mais populares o cálculo da entropia. Após a escolha do atributo e ponto de corte, são gerados dois nós correspondentes as amostras que foram divididas. Esse processo continua até um critério de parada.

Para a classificação de uma amostra de dados a árvore é percorrida e cada nó representa um condicional a ser verificado.

2.5.2 Regressão Logística

A Regressão Logística (do inglês *Logistic Regression*) é um modelo estatístico que, em sua forma básica, usa uma função logística para modelar uma variável dependente binária. Na análise de regressão, a regressão logística estima os parâmetros de um modelo logístico. O cálculo se dá pela Equação 52.

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}. \tag{52}$$

onde p_i é a probabilidade de ocorrência da amostra i, $x_{j,i}$ é o atributo j da amostra i, e β_k são os coeficientes lineares a serem otimizados.

2.5.3 K-Nearest Neighbor

K-Nearest Neighbor (KNN) é um algoritmo de classificação que utiliza o conceito de proximidade espacial para classificar uma amostra. Para classificar uma amostra A o KNN procura os k vizinhos mais próximos de A pertencentes à base de treinamento, a classe predominante nos k vizinhos é o resultado da classificação.

2.5.4 Floresta Aleatória

A Floresta Aleatória (do inglês Random Forest) é um algoritmo que funciona de modo similar às Arvores de Decisão. A Floresta Aleatória funciona através da geração de N conjuntos aleatórias de amostras pertencentes à base de dados, esses conjuntos são utilizadas para a criação de N Árvores de Decisão. A classificação é feita por uma votação dos resultados das árvores geradas.

2.5.5 Máquina de Suporte de Vetores

A metodologia da Máquina de Suporte de Vetores (do inglês Support Vector Machine - SVM) vem da aplicação da teoria de aprendizagem estatística para separar hiperplanos para problemas de classificação binária. A ideia central do SVM é ajustar uma função discriminativa para que faça o uso ideal da informação de separabilidade dos casos de fronteira. Dado um conjunto de casos que pertencem a uma das duas classes, o treinamento de um SVM linear consiste em procurar o hiperplano que deixa o maior número de casos da mesma classe no mesmo lado, enquanto maximiza a distância de ambas as classes do hiperplano (CUSANO et al., 2003).

Além do hiperplano outras superfícies podem ser aplicadas à metodologia da SVM através de uma função *kernel* na equação de otimização da separabilidade do modelo (Equação 53).

$$max(\sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i k(\vec{x}_i, \vec{x}_j) y_j c_j) \tag{53}$$

onde c_i são multiplicadores de Lagrange para $\sum_{i=1}^{n} c_i y_i = 0$ e $0 \leq c_i \leq C$, x_i o vetor de atributos de uma amostra i no conjunto de treinamento, $y \in [-1, 1]$ a classe de i.

Seguem abaixo as funções *kernel* mais populares:

- Linear: $k(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j})$
- Polinomial: $k(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d$
- RBF: $k(\vec{x_i}, \vec{x_j}) = \exp(-\gamma \|\vec{x_i} - \vec{x_j}\|^2)$

2.5.6 Máquina de Aumento de Gradiente

A Máquina de Aumento de Gradiente (do inglês Gradient Boost Machine - GBM) é um algoritmo de Aprendizado de Máquina baseado em árvore de decisão que usa uma estrutura de aprimoramento de gradiente, o procedimento de aprendizado se ajusta consecutivamente a novos modelos para fornecer uma estimativa mais precisa da variável de resposta. (FRIEDMAN, 2000)

2.6 MÉTRICAS

As métricas são parte essencial do processo de avaliação de um modelo preditivo, são realizadas através de análises estatísticas sobre os dados verdadeiros e predistos pelo modelo de classificação, e permitem quantificar o comportamento do modelo em diferentes aspectos.

A Tabela 2 é chamada matriz de confusão e apresenta uma forma de representação dos dados obtidos por um modelo de classificação binária, as colunas representam a quantidade de amostra preditas como verdadeiras ou falsas, e as linhas representam o valor real dessas amostras.

Tabela 2 – Matriz de Confusão

Real\Predito	**Positivo**	**Negativo**
Positivo	Verdadeiros Positivos (VP)	Falso Negativos (FN)
Negativo	Falso Positivos (FP)	Verdadeiro Negativos (VN)

A partir das entradas da matriz de confusão podemos elaborar diferentes métricas que descrevem características especificas do modelo.

$$Acurácia = \frac{VP + VN}{VP + FN + FP + VN} \tag{54}$$

$$Sensibilidade = cobertura = \frac{VP}{VP + FN} \tag{55}$$

$$Especificidade = \frac{VN}{VN + FP} \tag{56}$$

$$Precisão = \frac{VP}{VP + FP} \tag{57}$$

$$F_1score = \frac{2 * precisão * cobertura}{precisão + cobertura} \tag{58}$$

- **Acurácia:** taxa de acerto geral do classificador.
- **Sensibilidade ou cobertura:** taxa de acerto de casos positivos do classificador. Indica quão bem o classificador identifica casos positivos.
- **Especificidade:** taxa de acerto de casos negativos do classificador. Indica o quão bem o classificador identifica casos negativos.
- **Precisão:** proporção dos casos classificados como positivos que realmente são positivos. É uma medida da confiabilidade do preditor para classificações positivas.

- **F1 score:** média harmônica da precisão e cobertura.

2.6.1 Curva ROC e AUC

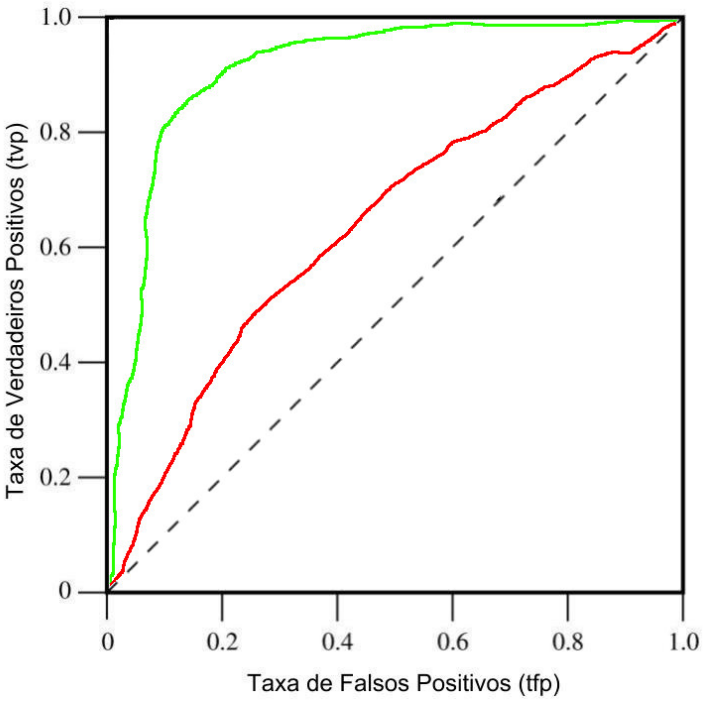
Análise ROC (do inglês Receiver Operating Characteristics) é um método gráfico para avaliação, organização e seleção de sistemas de predição. Na comunidade médica a curva ROC tem extensa utilização na avaliação de sistemas CAD (ZOU, 2019).

O espaço do gráfico ROC (Figura 3) é definido por um plano bidimensional onde o eixo das abcissas é representado pela taxa de falso-positivos (Equação 59) avaliados pelo classificador e o eixo das ordenadas é representado pela taxa de verdadeiro-positivos (Equação 60)

$$taxa\ de\ FP\ (tfp) = \frac{FP}{FP + TN} \tag{59}$$

$$taxa\ de\ VP\ (tvp) = \frac{TP}{TP + FN} \tag{60}$$

Figura 3 – Espaço ROC



Fonte: adaptado de (FAWCETT, 2006)

Em um espaço ROC (Figura 3) certos elementos apresentam características importantes. A linha $x = y$ representa um classificador aleatório onde a chance de classificação é 50% para cada classe. O ponto $(0, 0)$ representa um modelo que nunca classifica uma amostra como

positiva. O ponto $(1, 1)$ representa um modelos onde todas as amostras são classificadas como positivas. O ponto $(0, 1)$ representa um classificador perfeito. Ao se comparar duas curvas em um espaço ROC, aquela que mais se aproxima de do ponto $(0, 1)$ é a de melhor desempenho. Na Figura 3 o modelo representado pela curva verde pode ser considerado melhor que o modelo representado pela curva vermelha.

2.6.2 Importância de Atributos

A Importância de Gini (do inglês *Gini Importance* - GI) é uma métrica pertencente à modelos baseados em árvores e mensura a relevância de um atributo para o processo de classificação de um modelo. Também conhecida como Redução Média de Impureza, a GI calcula a importância de cada atributo através do número de vezes que um recurso é dividido, ponderando pelo número de amostras que ele divide (Equação 61).

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{61}$$

onde ni_j é a importância do nó j, w é o número de amostras divididas pelo nó j, C é o número de divisões do valor do nó j.

As equações abaixo especificam o cálculo da importância de uma árvore (Equação 62), seu valor normalizado (Equação 63), e o cálculo da importância para múltiplas árvores (Equação 64).

$$fi_i = \frac{\sum_{j:node \in feature_i} ni_j}{\sum_{k \in nodes}} \tag{62}$$

$$normfi_i = \frac{fi_j}{\sum_{j \in features} fi_j} \tag{63}$$

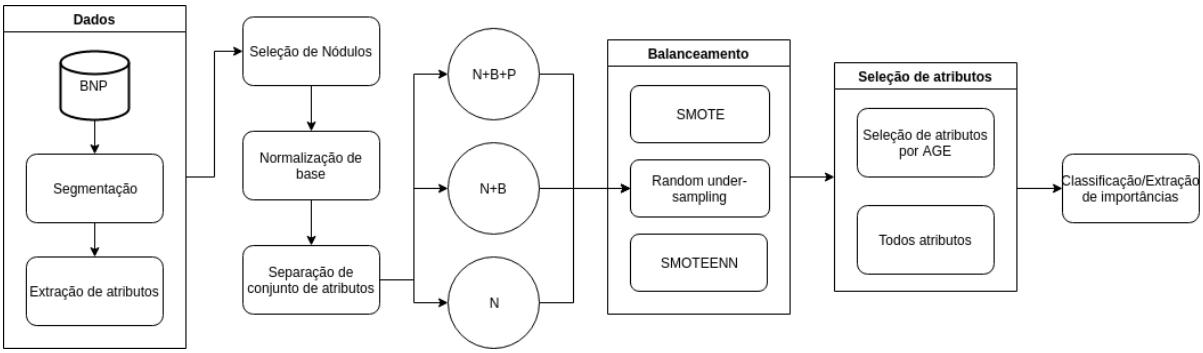
$$MTfi_i = \frac{\sum_{j \in trees} normfi_{ij}}{T} \tag{64}$$

onde T é o número de árvores.

3 MATERIAIS E MÉTODOS

Uma visão geral da metodologia aplicada neste trabalho pode ser vista na figura 4. Os dados de nódulos pulmonares utilizados foram extraídos através de imagens pertencentes ao BNP (seção 2.2). Os nódulos foram segmentados e tiveram seus atributos extraídos (seção 2.2.3) e então selecionados (seção 3.1). Foi realizada uma normalização (seção 3.2) na base de dados e então a partir da união de diferentes categorias de atributos, foram originados três diferentes conjuntos de dados (nódulo, nódulo+borda e parênquima+nódulo+borda) (seção 3.3). As bases de treinamento dos conjuntos de dados foram balanceadas (seção 3.4). Foi realizada uma etapa de seleção de atributos para cada base de dados (seção 3.5), e por fim, o treinamento e teste dos modelos de classificação (seção 3.6).

Figura 4 – Esquema geral da metodologia utilizada neste trabalho.



Os experimentos deste trabalho foram realizados em um PC GNU/Linux Ubuntu 18.04 LTS, CPU Intel Core i5 3.30GHz e 8GB de RAM. Neste trabalhos utilizamos os softwares: python (versão 3.7.3), sklearn (versão 0.21.2), imblearn (versão 0.5.1).

3.1 SELEÇÃO DE NÓDULOS

Os nódulos do BNP com probabilidade de malignidade 3 foram descartados devido a incerteza definida pelo especialista, restando 1.171 nódulos. Devido a complexidade em segmentar o parênquima no entorno de nódulos não sólidos, foram selecionados apenas nódulos de natureza sólida, após esta etapa restaram 897 nódulos.

Neste trabalho os nódulos com malignidade 1 e 2 foram considerados benignos, e nódulos com malignidade 4 e 5 malignos.

Tabela 3 – Número de nódulos e suas classificações.

	Benigno		**Maligno**		
Probabilidade de malignidade	1	2	4	5	**Total**
Número de nódulos	252	364	154	127	897
Somatório	616		281		

3.2 PRÉ-PROCESSAMENTO

Os atributos foram normalizados usando a técnica Z-score (Equação 65), removeu-se a média e o desvio-padrão de cada atributo resultando em uma distribuição de atributos aproximadamente normal.

$$z_i = \frac{x_i - \mu}{\sigma} \tag{65}$$

Onde σ representa o desvio padrão do atributo, μ representa a média do atributo, e por fim z_i e x_i representam respectivamente o valor pós-normalização e o valor pré-normalização.

3.3 CONJUNTOS DE DADOS

Objetivando a comparação e medição das importâncias de diferentes categorias de atributos, foram montados três diferentes conjuntos contendo subconjuntos do total de atributos extraídos.

O primeiro conjunto de dados, identificado como conjunto Nódulo (N) contém os atributos de intensidade, forma e textura referentes à região do nódulo. O segundo conjunto é identificado como Nódulo+Borda (N+B) e contém todos os atributos do conjunto N acrescido dos atributos de nitidez de borda. Por fim, o terceiro conjunto de dados identificado como Nódulo+Borda+Parênquima (N+B+P) contém os atributos do conjunto N+B acrescido dos atributos de intensidade e de textura referentes a região do parênquima. A tabela 4 mostra o número de atributos para cada conjunto de dados.

Tabela 4 – Total de atributos para cada conjunto de dados.

	N	**N+B**	**N+B+P**
Total de atributos	59	72	122

A utilização desses três conjuntos de dados nos permitirá avaliar a importância dos atributos extraídos das regiões da borda do nódulo e do parênquima em relação a apenas atributos extraídos do nódulo no processo de classificação de nódulos pulmonares.

3.4 BALANCEAMENTO

O balanceamento foi realizado com objetivo de diminuir o viés de classificação dos modelos preditivos para a classe majoritária. Foram aplicados três métodos de balanceamento SMOTE, RU e SMOTEENN (seção 3.4). No SMOTE o parâmetro de geração de amostras foi definido para as 5 mais próximas amostras. No SMOTEENN a etapa de geração de amostras foi definida da mesma forma do SMOTE e o parâmetro de remoção de amostras foi definido para as 3 amostras mais próximas. O balanceamento foi efetuado durante a etapa de validação cruzada, e somente para os conjuntos de treinamento dos modelos preditivos.

3.5 SELEÇÃO DE ATRIBUTOS

Com o objetivo de reduzir o ruído e selecionar os atributos mais relevantes de cada conjunto para a classificação, foi aplicado um algoritmo genético evolutivo.

O AG foi definido com uma população de 50 indivíduos e um número máximo de 100 gerações, também foi definido como critério de parada que o número máximo de gerações sem mudanças de melhor individuo é 10. A seleção de indivíduos para a etapa de cruzamento foi feita através do método de torneio. O método de recombinação utilizado foi o *uniform crossover* com taxa de 60%. A taxa de mutação foi definida como 5%.

3.6 CLASSIFICAÇÃO

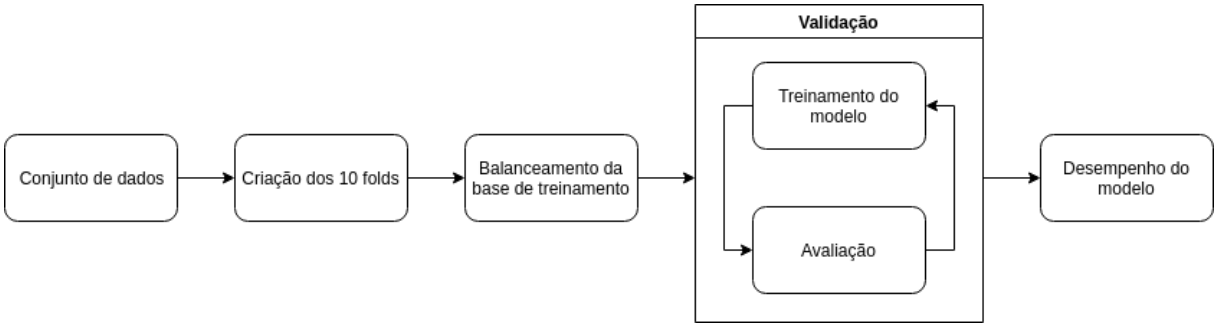
A classificação foi realizada sob 3 diferentes métodos de balanceamento: SMOTE; RU; SMOTEENN. Como classificadores foram utilizados 6 algoritmos de aprendizagem de máquina: *Decision Tree*; *Logistic Regression*; KNN; *Random Forest*; SVM; *XGBoost* (TARTAR et al., 2013; TAGHAVI et al., 2010; Kaucha et al., 2017). Foram utilizadas duas diferentes configurações para o algoritmo KNN, três diferentes configurações para os algoritmos *Random Forest* e SVM, e cinco diferentes para o *XGBoost*, resultando em 15 diferentes modelos de classificadores.

Os parâmetros utilizados para o modelo de *Logistic Regression* foram, regularização *l2*, tolerância 0,0001, coeficiente *C* 1,0, e o otimizador utilizado foi o *liblinear*. Para o modelo *Decision Tree*, o critério de avaliação foi a entropia, o critério para divisão foi o melhor valor. As configurações utilizadas para o modelo SVM-Linear foram, coeficiente *C* 1,0, *gamma* $\frac{1}{n_{atributos}}$, *kernel* linear. Para o modelo SVM-RBF foram as mesmas do SVM-Linear com exceção que o *kernel* utilizado foi o RBF. O modelo SVM-Poly possui as mesmas configurações dos modelos SVM-Linear e SVM-RBF com exceceções do kernel polinomial de terceiro grau. Os modelos de

KNN, KNN-10 e KNN-20 utilizaram pesos inversamente proporcionais a distância, a métrica utilizada foi *Minkowski* e o número de vizinhos foi setado para 10 e 20 respectivamente. Para os três modelos de *Random Forest* o critério de divisão foi entropia e os números de árvores foram 100, 500 e 1000. Para os cinco modelos do algoritmo XGBoost foram utilizados o *booster gbtree* com profundidade máxima 3, *gamma* 0 e taxa de aprendizado 0,1, o número de estimadores foram 15, 18, 20, 25, 50.

A avaliação dos modelos foi realizada utilizando uma validação cruzada com 10 *folds* para os 897 nódulos e os 3 conjuntos de dados: N; N+B e N+B+P. O esquema geral da classificação pode ser visto na figura 5.

Figura 5 – Esquema geral de classificação.



3.7 MÉTRICAS

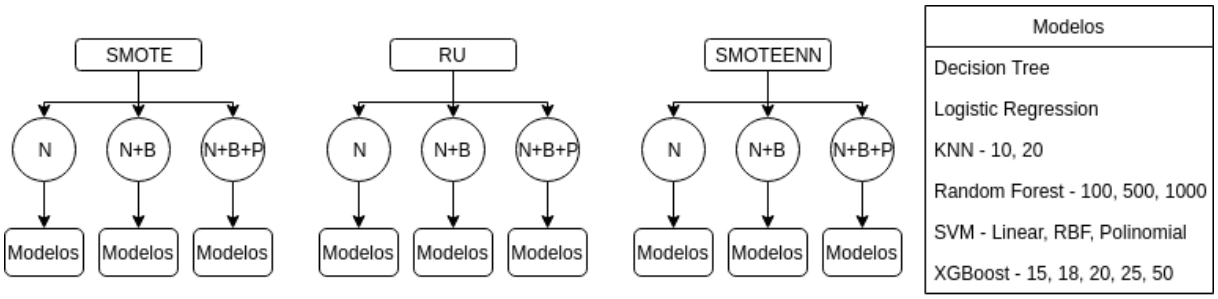
A avaliação dos modelos de classificação foi feita através da avaliação de 6 métricas, as métricas avaliadas foram: acurácia, F1-score, precisão, especificidade, sensibilidade, e área sobre a curva ROC. Tais métricas possuem extensa utilização em trabalhos similares na literatura (DILGER et al., 2015; SUI et al., 2015b). Cada métrica foi avaliada para cada etapa da validação cruzada, ao fim da validação foram calculadas e contabilizadas a média e desvio-padrão das métricas para cada modelo.

Para avaliação dos atributos foram utilizadas as métricas de importância e frequência de atributos esta última utilizada como uma métrica geral de importâncias. A métrica de importância de atributos foi realizada apenas em modelos de classificação baseados em árvores via seu modo de cálculo relacionado a características inerentes a modelos do tipo. A métrica de frequência de atributos foi realizada apenas nos modelos que resultaram nos melhores resultados de classificação.

4 RESULTADOS E DISCUSSÃO

Os resultados deste trabalho foram obtidos através da composição de 3 técnicas de balanceamento a 3 conjuntos de dados com e sem seleção de atributos, para 15 modelos de classificação. A Figura 6 mostra o esquema dos resultados.

Figura 6 – Esquema da composição dos resultados.



A métrica utilizada como critério de avaliação dos modelos foi a área sob a curva ROC (AUC), e, como critério de desempate foi utilizada a métrica F1-score. A escolha desses critérios se baseia na utilização da métrica em trabalhos similares (DILGER et al., 2015; FILHO et al., 2016; WAY et al., 2009) no caso da AUC, e na capacidade da métrica F1-score de mensurar o equilíbrio entre as medidas de Especificidade e Sensibilidade. As tabelas 5-6, mostram os resultados da AUC dos métodos de seleção de atributos, para cada técnica de balanceamento, em todos os conjuntos de dados dos modelos de classificação avaliados.

Na próxima seção os modelos são apresentados inicialmente de forma geral, e posteriormente de forma específica para os mais bem avaliados. Após a apresentação, são mostradas as importâncias dos melhores modelos baseados em árvore, e por fim a frequência geral dos atributos selecionados para a classificação.

4.1 DESEMPENHO DA CLASSIFICAÇÃO

Dentre os modelos avaliados sem seleção de atributos (Tabela 5) o melhor resultado, usando o critério de desempate, foi obtido pelo modelo RandomForest-1000 com o balanceamento RU e conjunto de dados N+B+P, com AUC média de 0,897.

Para os modelos avaliados com seleção de de atributos por algoritmo genético (Tabela 6) o melhor resultado foi obtido pelo algoritmo SVM-Linear com balanceamento por SMOTE no conjunto N+B+P, o modelo obteve AUC média de 0,915. Para este resultado foram selecionados 35 atributos.

Tabela 5 – Resultados AUC para a classificação utilizando diferentes estratégias de balanceamento em diferentes conjuntos de dados, sem seleção de atributos.

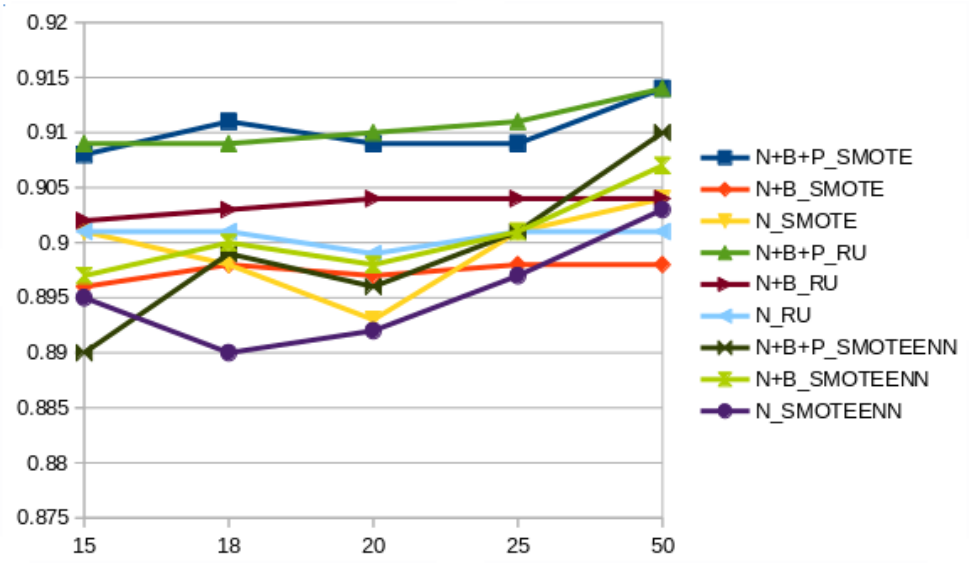
	SMOTE			RU			SMOTEEN		
Model	N+B+P	N+B	N	N+B+P	N+B	N	N+B+P	N+B	N
DecisionTree	0,739	0,735	0,710	0,765	0,746	0,740	0,789	**0,790**	0,787
KNN-10	0,866	0,863	0,863	0,866	**0,882**	0,871	0,839	0,849	0,848
KNN-20	0,872	0,876	0,875	0,867	**0,886**	0,877	0,857	0,864	0,860
LogisticRegression	0,884	0,889	0,888	0,884	0,885	0,887	0,889	**0,890**	0,889
RandomForest-100	**0,893**	0,892	0,886	0,895	0,892	0,888	0,892	0,889	0,886
RandomForest-500	**0,897**	0,893	0,889	**0,897**	0,892	0,887	0,893	0,893	0,888
RandomForest-1000	**0,897**	0,894	0,889	0,896	0,893	0,888	0,894	0,893	0,888
SVM-Linear	0,876	0,883	**0,887**	0,881	0,878	0,881	0,886	**0,887**	0,884
SVM-Poly	0,857	0,862	0,856	0,805	0,810	0,836	0,856	0,862	**0,863**
SVM-RBF	**0,893**	0,883	0,885	0,886	0,885	0,884	0,878	0,874	0,877
XGBoost-15	**0,889**	0,883	0,886	0,887	0,883	0,882	0,874	0,875	0,863
XGBoost-18	**0,892**	0,882	0,887	0,890	0,884	0,884	0,875	0,882	0,869
XGBoost-20	**0,894**	0,884	0,888	0,891	0,883	0,883	0,876	0,881	0,870
XGBoost-25	**0,894**	0,885	0,887	0,892	0,884	0,882	0,882	0,882	0,873
XGBoost-50	**0,894**	0,886	0,885	0,892	0,889	0,886	0,894	0,892	0,884

Tabela 6 – Resultados AUC para a classificação utilizando diferentes estratégias de balanceamento em diferentes conjuntos de dados, seleção por AG.

	SMOTE			RU			SMOTEEN		
Model	N+B+P	N+B	N	N+B+P	N+B	N	N+B+P	N+B	N
DecisionTree	0.781	0.766	0.784	0.815	0.815	0.781	0.817	**0.826**	0.813
KNN-10	0.908	0.901	0.897	0.906	**0.911**	0.901	0.895	0.887	0.887
KNN-20	0.907	0.905	0.902	**0.912**	**0.912**	0.903	0.902	0.897	0.897
LogisticRegression	0.908	0.908	0.901	0.911	0.905	0.900	**0.914**	0.905	0.904
RandomForest-100	**0.905**	0.903	0.901	0.902	0.904	0.900	0.903	0.900	0.896
RandomForest-500	**0.907**	0.904	0.901	**0.907**	0.903	0.901	**0.907**	0.901	0.900
RandomForest-1000	0.911	0.906	0.901	0.908	0.905	0.901	**0.912**	0.906	0.899
SVM-Linear	**0.915**	0.906	0.906	0.908	0.905	0.902	0.913	0.904	0.902
SVM-Poly	0.899	0.900	0.892	0.895	**0.902**	0.890	0.866	0.900	0.900
SVM-RBF	**0.911**	0.909	0.907	0.907	0.910	0.903	0.898	0.901	0.904
XGBoost-15	0.908	0.896	0.901	**0.909**	0.902	0.901	0.890	0.897	0.895
XGBoost-18	**0.911**	0.898	0.898	0.909	0.903	0.901	0.899	0.900	0.890
XGBoost-20	0.909	0.897	0.893	**0.910**	0.904	0.899	0.896	0.898	0.892
XGBoost-25	0.909	0.898	0.901	**0.911**	0.904	0.901	0.901	0.901	0.897
XGBoost-50	**0.914**	0.898	0.904	**0.914**	0.904	0.901	0.910	0.907	0.903

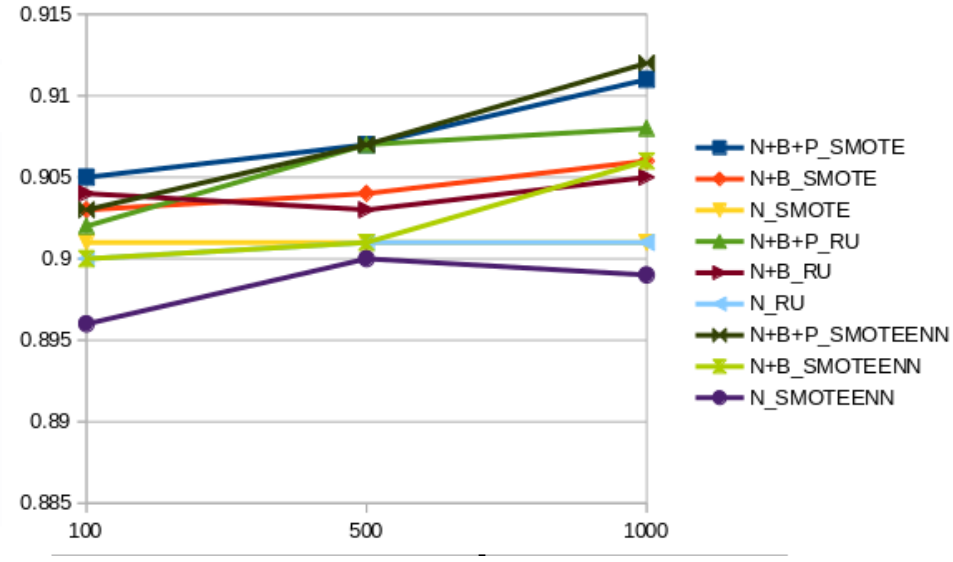
A Figura 7 apresenta graficamente o desempenho do algoritmo XGBoost em diferentes configurações da seleção de atributos por AG. O conjunto N+B+P apresentou maior AUC média com os balanceamentos RU e SMOTE no valor de 0,914. No conjunto N+P a maior AUC média foi obtida pelo balanceamento SMOTEENN com valor de 0,907, o balanceamento SMOTEENN também apresentou a maior variação de resultados neste conjunto com diferença de 0,010 entre a maior e menor AUC média. No conjunto N a maior AUC média foi obtida pelo balanceamento SMOTE com valor de 0,904, a diferença entre o maior e menor resultado deste conjunto foi de 0,011. Em geral a melhor média de resultados foi obtida com o balanceamento RU.

Figura 7 – Desempenho do algoritmo XGBoost com variação de número de árvores.



A Figura 8 mostra o desempenho das configurações do algoritmo Random Forest com seleção de atributos por AG. No conjunto N+B+P a maior média AUC foi obtida com o balanceamento SMOTEENN, com valor de 0,912. O conjunto N+B obteve maior valor AUC com a utilização do balanceamento SMOTE, com valor de 0,906. No conjunto N a maior AUC média foi obtida pelo balanceamento SMOTE, com valor de 0,901. Para todos os casos a AUC média foi maior no conjunto N+B+P em relação ao conjunto N+B e maior no conjunto N+B em relação ao conjunto N, evidenciando a superioridade do conjunto N+B+P sobre os conjuntos N+B e N.

Figura 8 – Desempenho do algoritmo Random Forest com variação de número de árvores.



As Tabelas 7-9 mostram os modelos com melhor desempenho para os conjuntos de dados balanceados com SMOTE. Os resultados mostram um diferença absoluta de 0,006 comparando o

melhor caso do conjunto N+B+P com o conjunto N+B, e 0,008 quando comparado com o melhor caso do conjunto N. Também nota-se uma predominância do algoritmo SVM nas configurações SVM-Linear e SVM-RBF indicando que talvez esses modelos performem melhor com este tipo de balanceamento.

Tabela 7 – Três melhores classificações conjunto N+B+P, balanceado com *SMOTE*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
SVM-Linear	0,826	0,866	**0,913**	0,828	**0,822**	**0,915±0,021**
XGBoost-50	0,816	0,860	0,899	0,828	0,789	0,914±0,025
SVM-RBF	**0,836**	**0,877**	0,901	**0,859**	0,786	0,911±0,037

Tabela 8 – Três melhores classificações conjunto N+B, balanceado com *SMOTE*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
SVM-RBF	0,807	0,852	0,903	0,808	0,804	**0,909±0,032**
LogisticRegression	0,819	0,861	**0,913**	0,817	**0,825**	0,908±0,031
SVM-Linear	**0,824**	**0,869**	0,887	**0,856**	0,754	0,906±0,030

Tabela 9 – Três melhores classificações conjunto N, balanceado com *SMOTE*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
SVM-RBF	**0,832**	**0,871**	0,916	**0,833**	0,829	**0,907±0,034**
SVM-Linear	0,809	0,851	**0,920**	0,794	**0,843**	0,906±0,033
XGBoost-50	0,817	0,860	0,905	0,821	0,807	0,904±0,028

As Tabelas 10-12 mostram os modelos com melhor desempenho para os conjuntos de dados balanceados com RU. O desempenhos dos conjuntos N+B+P e N+B foram similares, a diferença absoluta entre os seus maiores valores de AUC foi de 0,002. O conjunto N obteve 0,903 como melhor performance, 1,20% abaixo do conjunto N+B+P e 0,99% do conjunto N+B.

Tabela 10 – Três melhores classificações conjunto N+B+P, balanceado com *RU*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
XGBoost-50	0,815	0,854	0,928	0,794	0,861	**0,914±0,022**
KNN-20	0,819	0,859	0,922	0,807	0,847	0,912±0,024
LogisticRegression	**0,831**	**0,868**	**0,932**	**0,815**	**0,864**	0,911±0,030

Tabela 11 – Três melhores classificações conjunto N+B, balanceado com *RU*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
KNN-20	**0,815**	0,854	**0,928**	0,796	**0,857**	**0,912±0,022**
KNN-10	0,806	0,846	0,921	0,786	0,850	0,911±0,020
SVM-RBF	**0,815**	**0,856**	0,920	**0,804**	0,840	0,910±0,026

Tabela 12 – Três melhores classificações conjunto N, balanceado com *RU*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
SVM-RBF	**0,810**	**0,852**	0,916	**0,799**	0,836	**0,903±0,036**
KNN-20	0,797	0,839	0,915	0,777	0,840	**0,903±0,022**
SVM-Linear	0,804	0,845	**0,925**	0,779	**0,857**	0,902±0,030

As Tabelas 13-15 mostram os modelos com melhor desempenho para os conjuntos de dados balanceados com SMOTEEN. O maior valor AUC atingido foi de 0,914 com o modelo de Logistic Regression. A diferença para o maior valor obtido no conjunto N+B foi de 0,007, e para o modelo melhor avaliado do conjunto N foi de 0,010

Tabela 13 – Três melhores classificações conjunto N+B+P, balanceado com *SMOTEENN*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
LogisticRegression	0,803	0,841	0,943	0,762	0,893	**0,914±0,026**
SVM-Linear	**0,804**	0,841	**0,947**	0,760	**0,900**	0,913±0,029
RandomForest-1000	**0,804**	**0,844**	0,930	**0,774**	0,868	0,912±0,028

Tabela 14 – Três melhores classificações conjunto N+B, balanceado com *SMOTEENN*.

	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
XGBoost-50	0,789	**0,832**	0,916	0,766	0,839	**0,907±0,034**
RandomForest-1000	**0,803**	0,724	0,916	**0,787**	0,836	0,906±0,031
LogisticRegression	0,788	0,829	**0,929**	0,752	**0,868**	0,905±0,033

Tabela 15 – Três melhores classificações conjunto N, balanceado com *SMOTEENN*.

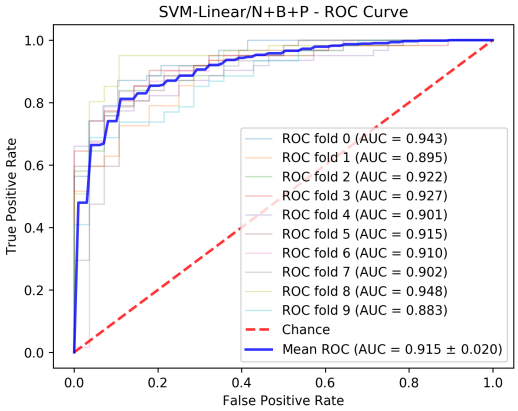
	Acurácia	F1-score	Precisão	Sens.	Espec.	AUC
SVM-RBF	0,800	0,842	0,924	0,774	**0,857**	**0,904±0,039**
LogisticRegression	0,785	0,827	**0,925**	0,750	0,861	**0,904±0,032**
RandomForest-500	**0,803**	**0,847**	0,907	**0,799**	0,811	0,903±0,032

O melhor desempenho AUC para o conjunto N+B+P é obtido pelo modelo SVM-Linear com balanceamento SMOTE (Tabela 7), o melhor desempenho para o conjunto N+B é obtido pelo modelo KNN-20 com balanceamento RU (Tabela 11), e o melhor desempenho para o conjunto N é obtido pelo modelo SVM-RBF (Tabela 9). As curvas ROC desses modelos são apresentadas na Figura 9.

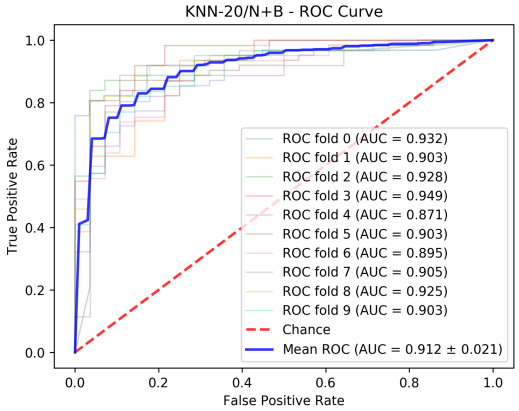
A linha azul representa a curva ROC média para os 10 *folds* da classificação. As outras linhas representam o desempenho para cada *fold*.

Figura 9 – Curvas ROC para os melhores modelos para cada conjunto de dados.

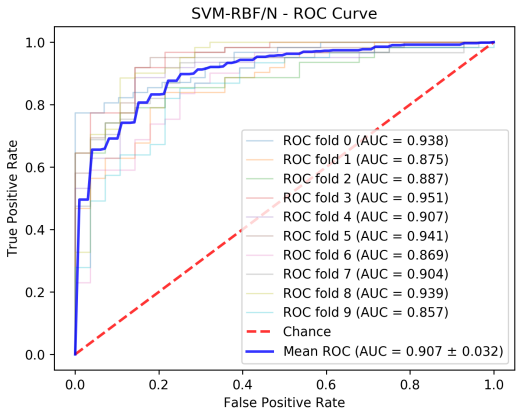
(a) SVM-Linear N+B+P SMOTE ($AUC = 0,915$)



(b) KNN-20 N+B RU ($AUC = 0,912$)



(c) SVM-RBF N SMOTE ($AUC = 0.907$)

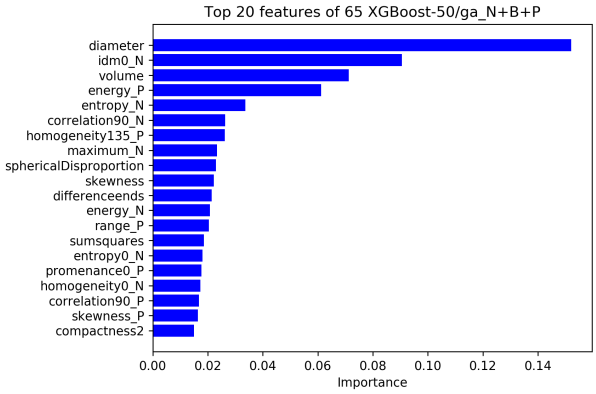


A Figura 10 apresenta os 20 atributos com maior importância nos dois melhores modelos de classificação baseados em árvores do conjunto N+B+P. A Tabela 16 mostra os atributos utilizados na classificação dos melhores modelos para os balanceamentos SMOTE e RU do conjunto N+B+P. A Tabela 17 mostra os resultados dos três melhores algoritmos classificados para cada técnica de balanceamento.

Na Figura 10 podemos ver o alto grau de importância de alguns atributos que pertencem às duas figuras como a Energia do parênquima (energy_P), Entropia do nódulo em 0^o (entropy0_p). Também é interessante notar que 30% dos 20 melhores atributos pertencentes ao modelo da Figura 10a são originados do parênquima, 45% se contabilizados os atributos de nitidez borda, esse número atinge 50% para os 20 melhores atributos da Figura 10b. Assim é evidenciada a importância do parênquima na classificação.

Figura 10 – Importância dos atributos dos dois melhores modelos baseados em árvore, conjunto N+B+P.

(a) XGBoost-50 SMOTE



(b) XGBoost-50 RU

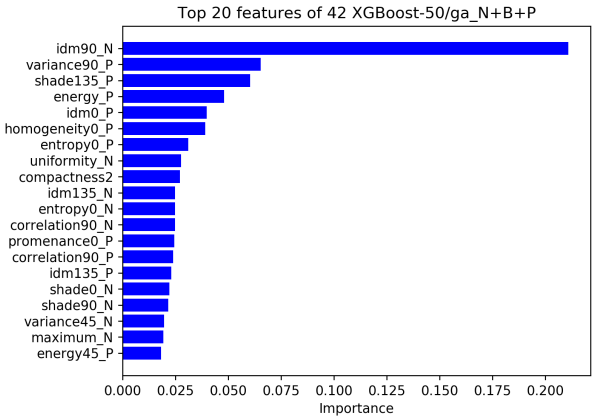


Tabela 16 – Lista de atributos dos dois melhores modelos preditivos.

Conjunto de Dados N+B+P	SMOTE	RU
Atributo	SVM-Linear	XGBoost-50
Intensidade de Nódulo	Entropia Intensidade média Skewness Variância	Intensidade máxima e média Raiz quadrada média Uniformidade
Intensidade de Parenquima	Intensidade máxima Skewness Desvio padrão Uniformidade Variância	Energia Intensidade média e mediana Desvio médio absoluto Intervalo
Forma	Esfericidade Relação superfície-volume	Compacidade 2 Desproporção esférica
Textura de Nódulo	Entropia em 90 Correlação em 0 e 90 Contraste em 0 e 135 Matiz em 90 Proeminência em 90 Variância em 90 e 135	Entropia em 0 Correlação em 0, 45 e 90 Contraste em 0, 90 e 135 MDI em 90 e 135 Matiz em 90 Proeminência em 45 Variância em 45 e 90
Textura de Parenquima	Entropia em 0 e 90 Correlação em 0 e 135 MDI em 90 e 135 Homogeneidade em 45 e 90	Energia em 135 Entropia em 0 e 90 Correlação em 90 e 135 Contraste em 45 e 135 MDI em 0 e 135 Homogeneidade em 0 Proeminência em 0 e 135 Variância em 0, 45 e 90
Nitidez de Borda	Diferença de extremos Soma dos valores Soma dos quadrados Média geométrica Desvio padrão Média de Skewness Segundo momento central	Diferença de extremos Soma dos logs Variância da amostra Média de Skewness

Segundo a Tabela 17 os atributos mais recorrentes (Ocorrência > 4) para o conjunto

Tabela 17 – Frequência de atributos dos 3 melhores algoritmos avaliados para cada técnica de balanceamento. Escala de 0 a 9.

Tipo	**Atributo**	**Frequência**		
		N+N+P	**N+B**	**N**
Intensidade Nódulo	Energia	3	6	1
	Entropia	5	6	6
	Kurtosis	1	2	4
	Intensidade máxima	4	5	6
	Intensidade média	7	4	3
	Desvio médio absoluto	3	1	1
	Intensidade mediana	2	2	1
	Intensidade mínima	5	5	0
	Intervalo de intensidade	0	3	4
	Raiz quadrada média	3	3	4
	Skewness	4	6	6
	Desvio padrão	2	2	2
	Uniformidade	2	4	4
	Variância	3	3	2
Intensidade Parênquima	Energia	5	-	-
	Entropia	3	-	-
	Kurtosis	0	-	-
	Intensidade máxima	3	-	-
	Intensidade média	4	-	-
	Desvio médio absoluto	5	-	-
	Intensidade mediana	4	-	-
	Intensidade mínima	3	-	-
	Intervalo de intensidade	2	-	-
	Raiz quadrada média	2	-	-
	Skewness	5	-	-
	Desvio padrão	2	-	-
	Uniformidade	5	-	-
	Variância	3	-	-
Forma	Compactness1	0	2	1
	Compactness2	3	2	0
	Desproporção esférica	7	7	6
	Esfericidade	6	9	7
	Área	2	3	0
	Área da superficie	5	3	2
	Relação superfície-volume	6	7	4
	Volume	1	3	2
	Diâmetro	4	2	3
Textura Nódulo	Energia em 0, 45, 90 e 135	1, 2, 1, 0	0, 3, 6, 3	2, 0, 2, 0
	Entropia em 0, 45, 90 e 135	2, 0, 2, 2	5, 2, 3, 2	3, 3, 2, 3
	Matiz em 0, 45, 90 e 135	2, 2, 2, 2	1, 2, 4, 2	1, 2, 3, 4
	Homogeneidade em 0, 45, 90 e 135	4, 4, 4, 2	7, 4, 1, 4	5, 3, 3, 5
	Correlação em 0, 45, 90 e 135	2, 4, 7, 1	3, 4, 5, 4	5, 4, 3, 4
	Contraste em 0, 45, 90 e 135	6, 4, 3, 5	3, 1, 1, 1	1, 2, 4, 3
	Proeminência em 0, 45, 90 e 135	7, 4, 3, 3	2, 1, 3, 1	5, 2, 5,4
	Variância em 0, 45, 90 e 135	3, 7, 6, 4	2, 2, 1, 2	4, 4, 4, 3
	MDI em 0, 45, 90 e 135	3, 2, 3, 2	5, 2, 4, 2	4, 4, 4, 1
Textura Parênquima	Energia em 0, 45, 90 e 135	3, 3, 3, 0	-	-
	Entropia em 0, 45, 90 e 135	6, 3, 6, 3	-	-
	Matiz em 0, 45, 90 e 135	2, 3, 4, 3	-	-
	Homogeneidade em 0, 45, 90 e 135	3, 3, 4, 1	-	-
	Correlação em 0, 45, 90 e 135	3, 4, 4, 3	-	-
	Contraste em 0, 45, 90 e 135	6, 6, 3, 3	-	-
	Proeminência em 0, 45, 90 e 135	4, 5, 5, 4	-	-
	Variância em 0, 45, 90 e 135	5, 4, 7, 5	-	-
	MDI em 0, 45, 90 e 135	3, 5, 4, 5	-	-
Nitidez de Borda	Diferença entre extremos	6	3	-
	Soma dos valores	6	4	-
	Soma dos quadrados	6	3	-
	Soma dos logs	5	2	-
	Média aritmética	1	3	-
	Média geométrica	3	4	-
	Variância da população	1	1	-
	Variância da amostra	1	1	-
	Desvio padrão	3	1	-
	Média de kurtosis	3	5	-
	Média de skewness	9	7	-
	Segundo momento central	6	4	-
	Diâmetro euclidiano	3	3	-

melhor avaliado são:

- Intensidade do nódulo: entropia, intensidade média e mínima;
- Intensidade do parênquima: energia, desvio médio absoluto, skewness, e uniformidade;
- Forma: desproporção esférica, esfericidade, área da superfície, e relação superfície-volume,;
- Textura do nódulo: correlação em 90^o, contraste em 0^o e 135^o, proeminência em 0^o, variância em 45^o, 90^o;
- Textura do parênquima: entropia em 0^o e 90^o, contraste em 0^o e 45^o, proeminência em 45^o e 90^o, variância em 0^o, 90^o e 135^o, MDI em 45^o e 135^o;
- Nitidez de borda: diferença entre extremos, soma dos valores, soma dos quadrados, soma dos logs, média de skewness e segundo momento central.

Os resultados deste trabalho evidenciam a importância dos atributos do parênquima e nitidez de borda para maior performance de classificação. Em geral os resultados dos modelos de classificação avaliados com o conjunto de atributos N+B+P apresentam superioridade significativa em relação aos demais conjuntos. Os modelos avaliados com o conjunto N+B também apresentam, em geral, melhor desempenho em relação ao conjunto N. Para as formas de balanceamento, o melhor desempenho foi obtido pelo conjunto balanceado por SMOTE com o algoritmo SVM-Linear.

4.2 TRABALHOS RELACIONADOS

Os resultados do melhor modelo obtido pela metodologia proposta neste trabalho são comparados à resultados oriundos de trabalhos similares na literatura na Tabela 18.

Tabela 18 – Comparação de trabalhos relacionados na classificação de nódulos pulmonares.

Trabalho	Modelo	Ac.	F1-score	Prec.	Sens.	Espec.	AUC	Validação
(FILHO et al., 2016)	Multi Layer Perceptron	-	-	-	-	-	0.875	VC 10-fold
(DILGER et al., 2015)	Artificial Neural Network	92%	91,45%	92,86%	92%	90,91%	0,935	LOO VC 10-fold
(WAY et al., 2009)	Linear Discriminant Analysis	-	-	-	-	-	0,857	LOO Two-Loop
Proposto	**SVM - Linear**	**82,6%**	**86,6%**	**91,3%**	**82,8%**	**82,2%**	**0,915**	**VC 10-fold**

Os resultados, apesar de poderem ser comparados aos trabalhos apresentados na Tabela 18, é importante ressaltar que foram feitos sob diferentes condições e conjuntos de dados.

Lima Filho (FILHO et al., 2016) propôs um modelo para classificação de nódulos pequenos combinando atributos do nódulo e parênquima. O conjunto de dados utilizado conteve 214 nódulos com diâmetros entre 5mm e 10mm provenientes do BNP (Seção 2.2), foram utilizados três algoritmos Random Forest, KNN e Multi Layer Perceptron (MLP), os atributos foram selecionados através do algoritmo genético. Os atributos utilizados por Lima Filho foram os mesmos atributos utilizados por este trabalho. A diferença de 0,040 entre o modelo proposto neste trabalho e o modelo proposto por Lima Filho se dá pela maior dificuldade na classificação de nódulos pequenos.

Dilger (DILGER et al., 2015) investigou o efeito de diferentes categorias de atributos do parênquima e do nódulo ao serem combinados com os classificadores Artificial Neural Network e Linear Discriminant Analysis. Foram selecionados 47 atributos utilizados para a classificação de 50 nódulos com diâmetros entre 4mm-30mm. O modelo alcançou AUC de 0,935, a classificação foi superior em 0,020 ao modelo proposto neste trabalho. Esta diferença pode se dar por conta do baixo número de nódulos presentes tanto na base de teste quanto da de treinamento.

Way (WAY et al., 2009) investigou a utilização de atributos radiais, do campo gradiente, forma e textura para classificação de nódulos. O conjunto de dados utilizados possuiu 256 nódulos com diâmetro entre 3mm-37.5mm, e os algoritmos utilizados foram Linear Discriminant Analisys e SVM. O modelo alcançou valor AUC de 0,857, a diferença para o modelo proposto neste trabalho foi de 0,058.

Ferreira Júnior (FERREIRA et al., 2017) efetuou uma seleção de atributos de textura e nitidez de borda do nódulo, seus resultados incluíram o número de ocorrências do atributo e sua significância estatística. O resultados foram 8 atributos de textura e nitidez de borda, para cada resultado exposto aqui será atribuído seus valores de ocorrência da Tabela 17. Os atributos selecionados foram: diferença entre extremos (6, 6), média kurtosis (3, 5), matiz em 0^o (2, 1, 1), MDI em 45^o (2, 2, 4), proeminência em 90^o (3, 3, 5), variância em 90^o (6, 1, 4), MDI em 90^o (3, 4, 4) e MDI em 135^o (2, 2, 1).

Lima Filho (FILHO et al., 2016) também incluiu em seu trabalho os atributos selecionados para o seu melhor modelo de classificação. Ao todo foram selecionados 65 atributos, estando 33,84% destes atributos com ocorrência maior que 5 em uma escala de 0 a 9 na Tabela 17. Ressalta-se novamente que o trabalho de Lima Filho é específico para nódulos pequenos, e portanto, os mesmos atributos podem ter diferentes relevâncias para a classificação.

5 CONCLUSÃO

Este trabalho apresentou o desenvolvimento de um modelo de classificação de nódulos pulmonares usando atributos radiômicos extraídos da região do macroambiente tumoral contendo as regiões do nódulo e do parênquima, sob diferentes técnicas de balanceamento.

Os resultados obtidos através da comparação de diferentes algoritmos de aprendizagem de máquina mostram que o melhor modelo para classificação de nódulos pulmonares com o conjunto de atributos utilizados neste trabalho é o SVM *kernel* linear balanceado com SMOTE. O modelo obteve AUC média de 0,915, especificidade de 82,8%, sensitividade de 82,2% e acurácia de 82,6%. A comparação entre diferentes técnicas de balanceamento evidencia a baixa performance geral da técnica SMOTEENN portanto não recomendamos sua utilização para balanceamento de nódulos pulmonares.

Com base nos resultados obtidos neste trabalho demonstramos que utilização dos atributos do parênquima e de nitidez de borda do nódulo melhoram efetivamente o desempenho de classificação de nódulos pulmonares. Destacamos também a alta taxa de importância dos atributos do parênquima, que chegou a ocupar 50% na lista dos 20 atributos mais importantes do nosso melhor modelo de classificação baseado em árvore. Evidencia-se assim que existe uma interação significativa entre o nódulo pulmonar e seu parênquima e que este deve ser amplamente utilizado no processo de classificação de nódulos pulmonares.

REFERÊNCIAS

AERTS, H.; VELAZQUEZ, E. R.; LEIJENAAR, R.; PARMAR, C.; GROSSMANN, P.; CAVALHO, S.; BUSSINK, J.; MONSHOUWER, R.; HAIBE-KAINS, B.; RIETVELD, D.; HOEBERS, F.; RIETBERGEN, M. M.; LEEMANS, C.; DEKKER, A.; QUACKENBUSH, J.; GILLIES, R.; LAMBIN, P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. **Nature communications**, v. 5, p. 4006, 08 2014.

AKGüL, C.; RUBIN, D.; NAPEL, S.; BEAULIEU, C.; GREENSPAN, H.; ACAR, B. Content-based image retrieval in radiology: Current status and future directions. **Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology**, v. 24, p. 208–22, 04 2010.

ARMATO, S. G.; MCLENNAN, G.; BIDAUT, L.; MCNITT-GRAY, M. F.; MEYER, C. R.; REEVES, A. P.; ZHAO, B.; ABERLE, D. R.; HENSCHKE, C. I.; HOFFMAN, E. A. 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**, Wiley Online Library, v. 38, n. 2, p. 915–931, 2011.

ATWATER, T.; COOK, C. M.; MASSION, P. P. The pursuit of noninvasive diagnosis of lung cancer. **Seminars in respiratory and critical care medicine**, v. 37, n. 5, p. 670–680, 2016.

BANNISTER, N.; BROGGIO, J. Cancer survival by stage at diagnosis for england (experimental statistics): Adults diagnosed 2012, 2013, 2014 and followe d up to 2015. **Produced in collaboration with Public Health England**, 2016.

BATISTA, G.; PRATI, R.; MONARD, M.-C. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explorations**, v. 6, p. 20–29, 06 2004.

BRAY, F.; FERLAY, J.; SOERJOMATARAM, I.; SIEGEL, R. L.; TORRE, L.; JEMAL, A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries: Global cancer statistics 2018. **CA: A Cancer Journal for Clinicians**, v. 68, 09 2018.

CHOI, W.; CHOI, T.-S. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. **Computer methods and programs in biomedicine**, v. 113, 09 2013.

CHUQUICUSMA, M. J.; HUSSEIN, S.; BURT, J.; BAGCI, U. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In: IEEE. **2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)**. [S.l.], 2018. p. 240–244.

CUSANO, C.; AB, G.; CIOCCA, G.; SCHETTINI, R. Image annotation using svm. **Internet Imaging V**, v. 5304, 12 2003.

Diciotti, S.; Lombardo, S.; Coppini, G.; Grassi, L.; Falchini, M.; Mascalchi, M. 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 A. JUDISCH, J. U. E. H. J. D. N. M. J. C. S. S. K. Improved pulmonary nodule classification utilizing lung parenchyma texture features. v. 9414, 2015. Disponível em: <https://doi.org/10.1117/12.2081397>.

DILGER, S.; UTHOFF, J.; JUDISCH, A.; HAMMOND, E.; MOTT, S.; SMITH, B.; NEWELL JR, J.; HOFFMAN, E.; SIEREN, J. (de R. Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. **Journal of Medical Imaging**, v. 2, p. 041004, 09 2015.

DILGER, S. K. N. The use of surrounding lung parenchyma for the automated classification of pulmonary nodules. In: . [S.l.: s.n.], 2013.

FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FERREIRA, J. R.; 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**, v. 12, n. 3, p. 509–517, Mar 2017. ISSN 1861-6429. Disponível em: <https://doi.org/10.1007/s11548-016-1471-7>.

FILHO, A. L.; MACHADO, A. P.; OLIVEIRA, M. **Modelo para Classificação de Nódulos Pulmonares Pequenos usando Descritores Radiomics**. Dissertação (Mestrado) — University of Alagoas - (UFAL), 2016.

FRIEDMAN, J. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, 11 2000.

GILLIES, R. J.; KINAHAN, P. E.; HRICAK, H. Radiomics: Images are more than pictures, they are data. In: **Radiology**. [S.l.: s.n.], 2016.

GOULD, M. K.; TANG, T.; LIU, I.-L.; LEE, J.; ZHENG, C.; DANFORTH, K.; KOSCO, A. E.; FIORE, J. L. D.; SUH, D. E. Recent trends in the identification of incidental pulmonary nodules. **American journal of respiratory and critical care medicine**, v. 192, 07 2015.

Haralick, R. M.; Shanmugam, K.; Dinstein, 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.

Il-Seok Oh; Jin-Seon Lee; Byung-Ro Moon. Hybrid genetic algorithms for feature selection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 26, n. 11, p. 1424–1437, Nov 2004. ISSN 0162-8828.

JERROLD, B.; SEIBERT, J.; LEIDHOLDT, E. M.; BOONE, J. M.; MAHESH, M. The essential physics of medical imaging, third edition. **Medical physics**, v. 40, p. 077301, 07 2013.

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**, Springer, v. 29, n. 6, p. 716–729, 2016.

JÚNIOR, J. R. F. et al. Auxílio computadorizado ao diagnóstico do câncer de pulmão otimizado por gpu. Universidade Federal de Alagoas, 2015.

Kaucha, D. P.; Prasad, P. W. C.; Alsadoon, A.; Elchouemi, A.; Sreedharan, S. Early detection of lung cancer using svm classifier in biomedical image processing. In: **2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)**. [S.l.: s.n.], 2017. p. 3143–3148.

KNIGHT, S.; CROSBIE, P.; BALATA, H.; CHUDZIAK, J.; HUSSELL, T.; DIVE, C. Progress and prospects of early detection in lung cancer. **Open Biology**, v. 7, p. 170070, 09 2017.

LEARDI, R.; BOGGIA, R.; TERRILE, M. Genetic algorithms as a strategy for feature selection. **Journal of Chemometrics**, v. 6, n. 5, p. 267–281, 1992. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.1180060506>.

LEVMAN, J.; MARTEL, A. A margin sharpness measurement for the diagnosis of breast cancer from magnetic resonance imaging examinations. **Academic radiology**, v. 18, p. 1577–81, 09 2011.

MEHRE SUDIPTA MUKHOPADHYAY, A. D. N. C. H. A. K. D. N. K. S. A. An automated lung nodule detection system for ct images using synthetic minority oversampling. v. 9785, 2016. Disponível em: <https://doi.org/10.1117/12.2216357>.

NETTLETON, D. F.; ORRIOLS-PUIG, A.; FORNELLS, A. A study of the effect of different types of noise on the precision of supervised learning techniques. **Artificial Intelligence Review**, v. 33, n. 4, p. 275–306, Apr 2010. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-010-9156-z>.

OLIVEIRA, M. A bag-of-tasks approach to speed up the lung nodules retrieval in the bigdata age. In: . [S.l.: s.n.], 2013.

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, Jan 2016. ISSN 1861-6429. Disponível em: <https://doi.org/10.1007/s11548-015-1245-7>.

SUI, Y.; WEI, Y.; ZHAO, D. Computer-aided lung nodule recognition by svm classifier based on combination of random undersampling and smote. **Computational and Mathematical Methods in Medicine**, v. 2015, p. 1–13, 05 2015.

SUI, Y.; WEI, Y.; ZHAO, D. Computer-aided lung nodule recognition by svm classifier based on combination of random undersampling and smote. **Computational and Mathematical Methods in Medicine**, v. 2015, p. 1–13, 05 2015.

TAGHAVI, S.; MOGHADDAM, H.; JAFARI, R.; ESMAEILZADEH, M.; GITY, M. Automated detection and classification of pulmonary nodules in 3d thoracic ct images. In: . [S.l.: s.n.], 2010. p. 3774–3779.

TARTAR, A.; KILIC, N.; AKAN, A. A new method for pulmonary nodule detection using decision trees. **Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference**, v. 2013, p. 7355–7359, 07 2013.

WAY, T. W.; SAHINER, B.; CHAN, H.-P.; HADJIISKI, L.; CASCADE, P. N.; CHUGHTAI, A.; BOGOT, N.; KAZEROONI, E. Computer-aided diagnosis of pulmonary nodules on ct scans: Improvement of classification performance with nodule surface features. **Medical Physics**, v. 36, n. 7, p. 3086–3098, 2009. Disponível em: <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3140589>.

Wilson, D. L. Asymptotic properties of nearest neighbor rules using edited data. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-2, n. 3, p. 408–421, July 1972. ISSN 0018-9472.

World Health Organisation. **Cancer**. 2019. <http://www.who.int/mediacentre/factsheets/fs282/fr/>, Last accessed on 2019-08-22.

XU, J.; NAPEL, S.; GREENSPAN, H.; BEAULIEU, C.; AGRAWAL, N.; RUBIN, D. Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval. **Medical physics**, v. 39, p. 5405–18, 09 2012.

YANG, J.; HONAVAR, V. Feature subset selection using a genetic algorithm. In: ____. **Feature Extraction, Construction and Selection: A Data Mining Perspective**. Boston, MA: Springer US, 1998. p. 117–136. ISBN 978-1-4615-5725-8. Disponível em: <https://doi.org/10.1007/978-1-4615-5725-8_8>.

YEN, S.-J.; LEE, Y.-S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In: ____. **Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 731–740. ISBN 978-3-540-37256-1. Disponível em: <https://doi.org/10.1007/978-3-540-37256-1_89>.

ZHU, X.; WU, X. Class noise vs. attribute noise: A quantitative study. **Artif. Intell. Rev.**, v. 22, p. 177–210, 11 2004.

ZOU, K. H. **Receiver Operating Characteristic (ROC) Literature Research**. 2019. Disponível em: <https://www.spl.harvard.edu/archive/spl-pre2007/pages/ppl/zou/roc.html>.