UNIVERSIDADE FEDERAL DO MARANHÃO – UFMA
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

ANTONINO CALISTO DOS SANTOS NETO

UMA ABORDAGEM RADIOMICS USANDO ÍNDICES DE DIVERSIDADE FILOGENÉTICA E FUNCIONAL PARA CLASSIFICAR NÓDULOS DE CÂNCER DE PULMÃO DE CÉLULAS NÃO PEQUENAS EM IMAGENS DE TOMOGRAFIA COMPUTADORIZADA

SÃO LUÍS – MA
2019

ANTONINO CALISTO DOS SANTOS NETO

UMA ABORDAGEM RADIOMICS USANDO ÍNDICES DE DIVERSIDADE FILOGENÉTICA E FUNCIONAL PARA CLASSIFICAR NÓDULOS DE CÂNCER DE PULMÃO DE CÉLULAS NÃO PEQUENAS EM IMAGENS DE TOMOGRAFIA COMPUTADORIZADA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da UFMA como requisito parcial para obtenção do grau de Mestre em Engenharia de Eletricidade.

Orientador: Prof. Dr. Aristófanes Corrêa Silva

Coorientador: Prof. Dr. André Borges Cavalcante

SÃO LUÍS – MA

2019

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

dos Santos Neto, Antonino Calisto.
UMA ABORDAGEM RADIOMICS USANDO ÍNDICES DE DIVERSIDADE FILOGENÉTICA E FUNCIONAL PARA CLASSIFICAR NÓDULOS DE CÂNCER DE PULMÃO DE CÉLULAS NÃO PEQUENAS EM IMAGENS DE TOMOGRAFIA COMPUTADORIZADA / Antonino Calisto dos Santos Neto. - 2019.
84 f.

Coorientador(a): André Borges Cavalcante.
Orientador(a): Aristófanes Corrêa Silva.
Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia de Eletricidade/ccet, Universidade Federal do Maranhão, São Luís, 2019.

1. Índice de diversidade filogenética. 2. Índice de diversidade funcional. 3. NSCLS. 4. NSCLS-Radiomics. 5. Radiomics. I. Cavalcante, André Borges. II. Silva, Aristófanes Corrêa. III. Título.

ANTONINO CALISTO DOS SANTOS NETO

UMA ABORDAGEM RADIOMICS USANDO ÍNDICES DE DIVERSIDADE FILOGENÉTICA E FUNCIONAL PARA CLASSIFICAR NÓDULOS DE CÂNCER DE PULMÃO DE CÉLULAS NÃO PEQUENAS EM IMAGENS DE TOMOGRAFIA COMPUTADORIZADA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da UFMA como requisito parcial para obtenção do grau de Mestre em Engenharia de Eletricidade.

Dissertação aprovada em ____ de _____ de _____.

Prof. Dr. Aristófanes Corrêa Silva
Orientador

Prof. Dr. André Borges Cavalcante
Coorientador

Prof. Dr. Antônio Oséas de Carvalho Filho
Membro da Banca Examinadora

Prof. Dr. Stelmo Magalhaes Barros Netto
Membro da Banca Examinadora

SÃO LUÍS – MA
2019

Aos meus pais, irmãos, família e amigos

AGRADECIMENTOS

A Deus.

A minha família pelo incentivo, carinho e confiança. Em especial aos meus pais, Neusa Maria Calisto Alves e Jenival Alves da Silva, e meus irmãos Paulyran Calisto Alves e Jenival Alves da Silva Júnior, pelos ensinamentos, conselhos e pelo esforço em sempre proporcionar a seus filhos tudo de melhor.

Ao orientador Aristófanes Silva e ao coorientador André Cavalcante por toda a paciência, confiança, conselhos e ensinamentos.

Aos professores por todo o conhecimento repassado. Em especial a Anselmo, João Dallyson e Geraldo.

Aos amigos que sempre me apoiaram e me fizeram crescer como pessoa e como profissional. Em especial a Otilio, João, Giovanni, Jonisson, Denes e Luana por sempre me ajudarem na realização deste trabalho.

A CAPES pelo apoio financeiro durante o mestrado.

A todos vocês muito obrigado.

"A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê"

Arthur Schopenhauer

RESUMO

O câncer de pulmão é a maior causa de morte por câncer em todo mundo, representando mais de 17% do total de mortes relacionadas com câncer, sendo que o câncer de pulmão de células não pequenas (*Non Small Cell Lung Cancer* - NSCLC) corresponde a aproximadamente 85% das ocorrências do câncer pulmonar. Porém, seu diagnóstico precoce pode ajudar em uma queda acentuada nesta taxa de mortalidade. Devido o árduo processo na análise dos exames por imagens, surge um campo emergente em processamentos de imagens chamado de Radiomics. Esta abordagem permite caracterizar uma imagem quantitativamente, o que possibilita a definição muito mais precisa do fenótipo do tumor, utilizando técnicas de processamento de imagens e reconhecimento de padrões, provendo um diagnóstico precoce de NSCLC de forma rápida e ajudando na opinião do especialista. Diante disso, este trabalho propõe uma metodologia para a classificação de nódulos de NSCLC em exames de Tomografia Computadorizada (TC) utilizando índices de diversidade filogenética e funcional em uma abordagem Radiomics. Dividida em seis etapas, esta metodologia se inicia com a aquisição das imagens de nódulos de NSCLC da base pública de imagens NSCLC-Radiomics. Na segunda etapa, as lesões foram extraídas utilizando as marcações dos especialistas. Em seguida, na terceira etapa são feitas quantizações para criarem maiores diversidades de espécies. Na quarta fase, são extraídas características de textura baseadas em índices de diversidade filogenética e funcional. Em seguida, na quinta fase as características são submetidas aos classificadores *Suport Vector Machine*, e Random Forest. Por fim, na sexta etapa, a metodologia proposta é validada utilizando a área sob a curva *Receiver Operating Characteristic* (ROC), o índice Kappa e a acurácia. Os melhores valores achados para a classificação de nódulos de NSCLC na abordagem Radiomics, resultaram em um índice Kappa de 0,989, uma área sob a curva ROC de 0,999 e uma acurácia de 99,44%.

Palavras-chave: Radiomics, NSCLS-Radiomics, NSCLS, Índice de diversidade filogenética, Índice de diversidade funcional, Classificação de nódulos pulmonares.

ABSTRACT

Lung cancer is the world's largest cause of cancer death, accounting for more than 17% of all cancer-related deaths, with Non- Small Cell Lung Cancer (NSCLC) corresponds to approximately 85% of lung cancer occurrences. However, its early diagnosis may help in a sharp fall in this mortality rate. Due to the arduous process in the analysis of the imaging tests, an emerging field in image processing called Radiomics arises. This approach allows quantitative characterization of an image, which allows a much more precise definition of the tumor phenotype, using image processing and pattern recognition techniques, providing an early diagnosis of NSCLC quickly and helping the opinion of the specialist. Therefore, this work proposes a methodology for the classification of NSCLC nodules in Computed Tomography (CT) examinations using indexes of phylogenetic and functional diversity in a Radiomics approach. Divided into six steps, this methodology starts with the acquisition of NSCLC nodal images from the public NSCLC-Radiomics imaging base. In the second step, the lesions were extracted using the markings of the specialists. Then, in the third stage, quantizations are made to create greater diversity of species. In the fourth phase, texture characteristics are extracted based on phylogenetic and functional diversity indexes. Then, in the fifth phase, the characteristics are submitted to the classifications Suport Vector Machine, Random Forest and Random Tree. Finally, in the sixth step, the proposed methodology is validated using the area under the Receiver Operating Characteristic (ROC) curve, the Kappa index and the accuracy. The best values found for the classification of NSCLC nodules in the Radiomics approach resulted in a Kappa index of 0.990, an area under the ROC curve of 0.999 and an accuracy of 99.44

Keywords: Radiomics, NSCLS-Radiomics, NSCLS, Index of phylogenetic diversity, Functional diversity index, Classification of pulmonary nodules

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC	Acurácia
AUC	*Area Under the Curve*
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnosis*
FN	Falso Negativo
FP	Falso Positivo
INCA	Instituto Nacional do Câncer
SVM	*Suport Vector Machine*
NSCLS	*Non Small Cell Lung Cancer*
ROI	Região de Interesse
ROC	*Receiver Operating Characteristic*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da tecnologia, diversas áreas têm sido beneficiadas na simplificação de soluções para seus problemas. Uma dessas áreas é a área da saúde, destacando as pesquisas relacionadas à detecção do tumor e diagnóstico do câncer. O câncer é caracterizado pelo crescimento desordenado das células, que invadem tecidos e células vizinhas, tornando-se, às vezes, agressivas e incontroláveis (BAILEY; KLEIN; LEEF, 2000).

Assim, é de suma importância a detecção e diagnóstico do câncer (nódulos malignos), com o intuito de aumentar as chances de sobrevivência do paciente sendo grande desafio em imagens médicas.

Geralmente, os nódulos pulmonares malignos seguem os seguintes padrões: eles se espalham por diversas partes do corpo, podendo, ás vezes, aparecerem como nódulos isolados. As bordas dos nódulos pulmonares são irregulares, possuem texturas semelhantes ao restante do exame, em sua maioria das vezes são pequenas, dificultando a análise por parte do especialista (AZEVEDO; SIQUEIRA, 2017).

Um nódulo é caracterizado como uma pequena massa de tecido que se forma sob um determinada estrutura (pele, órgão, entre outras), geralmente devido a lesão. Os nódulos podem ser benignos, não sendo necessária maior intervenção médica, ou podem ser malignos, sendo necessário um auxílio médico em seu tratamento (MACMAHON et al., 2017). Além das classificações de nódulos em benignos e malignos, os nódulos malignos podem ser classificados quanto ao tipo de células que os compõe: células pequenas e células não pequenas. O cãncer de pulmão de células pequenas se espalham de forma mais rápida pelo pulmão. Esse tipo de câncer de pulmão corresponde a aproximadamente 15% de todos os casos de câncer de pulmão.

Já o câncer de pulmão de células não pequenas (*Non Small Cells Lung Cancer* - NSCLC) é o mais comum entre os tipos de câncer de pulmão. Esse tipo é caracterizado pelo crescimento desordenado das células epiteliais, responsável por grande parte dos erros nas análises dos especialistas, já que a textura desse crescimento é bastante semelhante com a textura do tecido saudável.

A maioria dos nódulos pulmonares de NSCLC surgem nas paredes dos brônquios, atribuindo ao câncer de pulmão mais comum o nome de broncogênico, podendo também

ocorrer nas paredes dos pulmões. Às vezes, demoram vários anos para se desenvolverem, causando dificuldades na análise, pois visivelmente a textura é semelhante às paredes.

Devido a importância da análise de textura, somados aos avanços no desenvolvimento tecnológico e na detecção e diagnóstico do câncer de pulmão, tornou-se comum a extração de alto desempenho de características quantitativas que resultaram na conversão de imagens em dados, e na utilização destes dados para suporte na decisão. Esta prática é chamada de Radiomics (NETO et al., 2018).

A abordagem Radiomics, visa a extração de características a partir de imagens, assumindo que as características irão descrever os genótipos e fenótipos dos tecidos, da mesma maneira (ou melhor) que uma biópsia feita em um laboratório clínico, com a principal característica de que não seria invasiva, pois toda a análise seria feita por imagens (GILLIES; KINAHAN; HRICAK, 2015).

Este trabalho atua diretamente sobre o volume dos nódulos de NSCLC com a aplicação dos índices de diversidade filogenética e funcional, sobre uma base que se enquadre no contexto Radiomics, classificando-os (ou definindo seu tipo histológico) em adenocarcinoma, células grandes, adenocarcinoma células escamosas, mutação negativa e não especificados.

1.1 Justificativa para o tipo histológico

Atualmente, o câncer de pulmão, é o câncer com maior taxa de mortalidade mundial no homem e o segundo na mulher, ficando atrás apenas para o câncer de mama (TSUKAZAN et al., 2017). No Brasil, as taxas para o câncer de pulmão estão aumentando principalmente entre as mulheres, sendo o principal motivo, o aumento do uso do tabaco, devido o maior contato da população feminina com esse produto. Na população brasileira, esse contato representa em média 80% dos casos de câncer de pulmão (SCHMID et al., 2018).

Na detecção desses tipos de nódulos pulmonares, a espessura das fatias das imagem Tomografia Computadorizada (TC) torna-se cada vez mais fina. Com isso, as análises das imagens de TC feitas pelos especialistas em radiologia para detectar nódulos pulmonares menores, ou seja, podem se localizar em áreas de difícil detecção (como próximos às estruturas de textura e forma semelhantes), principalmente quando forem nódulos de NSCLC, além da possibilidade de conter estruturas conexas a ele, como por

exemplo os vasos sanguíneos. Com o intuito de distinguir entre nódulos e vasos, são necessários comparações de muitas imagens de TC para a retirada de dados relevantes a devido esse procedimento ser exaustivo e repetitivo, podendo ocasionar um desvio de atenção por parte dos especialistas, podendo haver algum equívoco na análise, o que faz com que ocorram erros nas suas análises frequentemente (ROCHA; ISHIKAWA; SZARF, 2017).

Além disso, o NSCLC corresponde de 85% a 90% dos casos de câncer de pulmão, sendo responsável pela maior parte das mortes por este câncer contabilizadas no mundo, necessitanto uma atenção especial para sua detecção e diagnóstico (DOMENICO et al., 2018).

É notável a grande importância na detecção e diagnóstico precoce de nódulos pulmonares, principalmente pelo fato de que a lesão pode ser precursora de câncer pulmonar, com maiores chances de ser NSCLC, provendo uma grande probabilidade de cura do câncer e um aumento de sobrevida do paciente.

Por essas razões, é de grande importância o desenvolvimento de ferramentas que auxiliem os profissionais da saúde na detecção e acompanhamento dos nódulos pulmonares.

1.2 Objetivo

O objetivo geral é desenvolver uma metodologia para uma abordagem Radiomics utilizando descritores de textura baseado em índices de diversidade filogenética e funcional para classificação de nódulos de NSCLC, em exames de TC.

1.2.1 Objetivos Específicos

Para alcançar o objetivo geral deste trabalho, os seguintes objetivos específicos deverão ser contemplados:

- Avaliar e adaptar índices de diversidade filogenética e funcional para descrever a textura das imagens;
- Utilizar técnicas de reconhecimento de padrões para testar as características produzidas em relação à sua capacidade de discriminar os tipos histológicos dos nódulos de NSCLC;

- Aplicar o algoritmo *Greedy Stepwise* para otimizar o processo de classificação selecionando o conjunto mínimo de características que melhor discriminam as amostras;

- Avaliar os métodos propostos através de experimentos, utilizando uma base de imagens pública de TC com nódulos de NSCLC que se adequem à proposta da abordagem Radiomics;

1.3 Contribuições do Trabalho

Este trabalho contribui diretamente para diversas áreas do meio científico, podendo-se destacar as seguintes:

- Índices de diversidade filogenética já utilizados no contexto de redução de falso positivo e diagnósticos de nódulos pulmonares, sendo a primeira vez utilizados no contexto de Radiomics;

- Utilização de indices de diversidade funcional em uma abordagem 3D, já utilizados no contexto de classificação do câncer de mama, porém em 2D, sendo a primeira vez utilizados no contexto de Radiomics;

- Novos índices de diversidade funcional adaptados para o contexto de processamento de imagens;

1.4 Organização do Trabalho

Os demais capítulos desta dissertação foram organizados em:

O Capítulo 2 apresenta um resumo dos trabalhos relacionados com o tema da pesquisa utilizando análise de textura.

O Capítulo 3 trata da fundamentação teórica necessária para a construção desta pesquisa. São abordados os conceitos referentes a classificação de nódulos de NSCLC, índices de diversidade filogenética, índices de diversidade funcional, seleção de características com algoritmo *Greedy Stepwise*, classificação e validação de resultados.

O Capítulo 4 descreve e detalha todas as etapas da metodologia proposta por esta dissertação.

O Capítulo 5 são mostrados e discutidos os resultados alcançados e um estudo comparativo com os trabalhos relacionados.

O Capítulo 6 são apresentados as considerações finais e sugestões de trabalhos futuros.

2 TRABALHOS RELACIONADOS

Neste capítulo, serão apresentados trabalhos disponíveis na literatura, que tratam sobre o diagnóstico de NSCLC.

No trabalho de (LI et al., 2014) é feito a classificação de nódulos de NSCLC em adenocarcinoma (ACA) e carcinoma de células escamosas (SCC). são utilizados 208 exames de ACA e 93 exames de SCC. Foram extraídas 1000 características. Como resultado, obteve-se uma acurácia de 86,05%.

No trabalho de (AERTS et al., 2014) são extraídas 440 características Radiomics de textura com o intuito de conterem grande significância clínica (poder de prognóstico) e de mostrar que através de imagens médicas, podem-se extrair diversas características. Quatro características foram selecionadas baseadas na estabilidade e na perfomace de prognóstico. Os testes foram feitos em uma base de 225 casos de NSCLC, em 136 casos de Carcinoma de células escamosas de cabeça e pescoço (*Head and neck squamous cell carcinoma* - HNSCC), 95 casos de HNSCC e em 89 casos de outra base de NSCLC. Como resultado, foi obtido um índice Kappa entre 0,65 e 0,69. Além disso, o trabalho mostra o potencial das imagens em fornecer características independentemente do tipo de conjunto testado.

O trabalho de (LUKE et al., 2014) é dividido em duas partes. A primeira parte tem como objetivo mostrar que as características das imagens de TC poderiam ser utilizadas para prever a redução do tamanho do tumor em resposta à terapia. Para obter esse objetivo, os pacientes com NSCLC de 64 estágios com tratamentos similares foram submetidos a um mesmo *scanner* e a um mesmo protocolo de TC. As características quantitativas das imagens foram extraídas e o método denominado de "regressão de componentes principais com seleção de subconjuntos de recozimento"foi utilizado para selecionar as características mais relevantes, criando uma matriz de características utilizadas para previsão das mudanças do tumor em resposta ao tratamento. Os testes de validação utilizaram o *cross-validation*. A segunda parte do trabalho, foi identificar o conjunto de características das imagens de TC de NSCLC que seria reprodutível para discriminar os nódulos de NSCLC. Para isso, foram utilizadas imagens de 56 pacientes de três *scanners* diferentes, e uma seleção de características que retornou 40 características de textura. Como resultado, obteve-se um índice de concordância *Kappa* de 0,90.

No traballho de (PATIL; MAHADEVAIAH; DEKKER, 2016) são extraídas 1000 características de forma, de textura, estatísticas de primeira ordem e de *wavelet* de nódulos de NSCLC em exames de TC da base pública NSCLC-Radiomics de 317 pacientes. Como resultado obteve uma acurácia de 88%.

Seguindo o conceito de Radiomics, o trabalho de (COROLLER et al., 2016) avalia se as características Radiomics propostas são capazes de discriminar os nódulos e prever a resposta patológica após a quimioterapia neoadjuvante em 127 pacientes com NSCLC. Quinze características Radiomics foram avaliadas quanto à sua capacidade de prever e discriminar a resposta patológica e discrimínção de nódulos de NSCLC. Características convencionais de imagem, como volume e diâmetro dos tumores, foram utilizadas para comparação com os resultados obtidos. Quanto à discriminação de tipos histológicos de nódulos de NSCLC obteve-se uma área sob a curva *Receiver Operating Characteristic* (ROC) de 0,63.

Em (HUYNH et al., 2016) é proposto a extração de características em uma base de exames de TC de pulmão tratados com terapia de radiação de corpo estereotáxico em nódulos de NSCLC. Foram extraídas 12 características de textura selecionadas pela relevância e estabilidade. Como resultado, obteve-se um índice *Kappa* de 0,67.

No trabalho de (SHEN et al., 2017) é feita uma comparação das imagens de TC de pulmão por fatia (2D) e por volume de exame de paciente (3D) quanto à eficiência das características, onde é apresentado os resultados das discriminações das características de forma e textura nesses dois tipos de imagens. No trabalho é analisado 588 pacientes, sendo extraídas 1014 características Radiomics (507 características para as fatias e 507 característica para os volumes). Foi utilizado a área sob a curva *Receiver Operating Characteristic* (ROC) para avaliar o desempenho de previsão dos classificadores treinados (a máquina de vetor de suporte e regressão logística). Dois grupos de imagens foram separados para treinamento (463 pacientes tanto para a análise 2D, sendo que cada paciente tem entre 120 e 160 fatias, quanto para a análise 3D do volume) e outro para validação (125 pacientes para cada tipo de análise 2D e 3D). No grupo de treinamento, o melhor resultado nas análises 2D foi uma área sobre a curva ROC de 0,653 e para as análises 3D foi uma área sobre curva ROC de 0,671. No grupo de validação, o melhor resultado nas análises 2D foi uma área sobre a curva ROC de 0,755 e nas análises 3D foi uma área sobre a curva ROC de 0,663.

No trabalho de (ZHANG et al., 2017), são extraídas características quantitativas

de imagens radiológicas que demonstraram fornecer um valor promissor de prognóstico na previsão dos resultados clínicos em vários estudos. No entanto, vários desafios, incluindo redundância de características, dados desequilibrados e tamanhos pequenos de amostras, levaram a uma precisão preditiva relativamente baixa Imagens de TC de 112 pacientes (idade média de 75 anos) com nódulos de NSCLC foram submetidos a radioterapia estereotáxica do corpo, com a intenção de prever recorrência, morte e sobrevida livre de recidiva usando uma análise com as características de textura. Para abordar a redundância das características, uma análise abrangente (seleção de classificadores) indicou que os modelos *Random Forest* e a *Principal Component Analysis* foram os métodos de modelagem preditiva e de seleção de características ideais. Uma análise completa de variância mostrou que os desfechos de dados, técnicas de seleção de características e classificadores foram fatores significativos para afetar a acurácia preditiva, sugerindo que esses fatores devem ser investigados quando se constroem modelos preditivos baseados em Radiomics para o prognóstico do câncer. O *Random Forest* obteve o melhor resultado com uma curva ROC de 0,79 para a predição de nódulos de NSCLC.

Em (HASSAN et al., 2018) informam que as características Radiomics são potenciais biomarcadores de imagem para avaliação da resposta terapêutica em oncologia. No entanto, a robustez das características em relação aos parâmetros de imagem não está bem estabelecida. Identificaram-se que as características mais eficientes das imagens previamente identificados são intrinsecamente dependentes do tamanho do *voxel* e do número de níveis de cinza em uma recente investigação de textura. Neste trabalho, validaram-se o tamanho do *voxel* e normalizações de níveis de cinza em tumores de pulmão. Dezoito pacientes com NSCLS de diferentes volumes tumorais foram analisados. Para comparar com os dados do paciente, os exames foram adquiridos em oito *scanners* diferentes. Vinte e quatro características previamente identificadas foram extraídas de nódulos de NSCLC. A classificação de Spearman (rs) e o índice Kappa foram usados como métricas. Oito das 10 características apresentaram correlações altas (rs$>$ 0,9) e baixas (rs $<$0,5) com o número de *voxels* antes e após as normalizações, respectivamente. Da mesma forma, as características da textura foram instáveis (Kappa $<$0,6) e altamente estáveis (Kappa $>$ 0,8) antes e após normalizações dos níveis de cinza, respectivamente. Conclui-se que o tamanho do *voxel* e normalizações de níveis de cinza derivadas de um estudo de textura também se aplicam a tumores de pulmão. Este estudo destaca a importância e a utilidade de investigar a robustez das características Radiomics em relação aos parâmetros

de imagem de TC em bases que utilizam a abordagem Radiomics.

No trabalho de (LU et al., 2019) os nódulos de NSCLC são classificados em 3 classes (adenocarcinoma, adenocarcinoma de células escamosas e câncer de pulmão de células pequenas) utilizando uma base de imagens com 229 exames, extraindo 1160 características em exames de TC. Como melhor resultado, obteve-se uma AUC ROC de 0,857.

Estes foram alguns resumos dos trabalhos mais recentes relacionados a extração de características em exames de TC de nódulos de NSCLC no contexto da abordagem Radiomics. Pode-se citar dois pontos em comum nos trabalhos selecionados: os índices de validação obtiveram valores baixos, utilizaram características tradicionais para discriminação e os trabalhos utilizam bases de dados pequenas de nódulos de NSCLC. No método proposto, serão explorados essas deficiências com o objetivo de melhorar as métricas de validação, utilizar um maior número de casos, utilizar características ainda não incorporadas à abordagem Radiomics e conseguir discriminar os nódulos de NSCLC com uma menor quantidade de características.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são detalhados os tópicos necessários para compreensão das técnicas utilizadas na elaboração do metodologia proposto. As seções seguintes abordam conceitos sobre o câncer de pulmão e o exame de TC, métodos de análise de textura baseado em diversidade filogenética e funcional, seleção de características, técnicas de reconhecimento de padrões e as métricas de validação dos resultados.

3.1 Fisiologia, Imagem e Patologia Pulmonar

A relação da estrutura da fisiologia, imagens e patologia pulmonar é de suma importância para o entendimento da grande variedade e diferenças sutis que podem representar as categorias presentes no câncer de pulmão.

3.1.1 Nódulo Pulmonar

Em (TAN; DESAI, 2003) é definido que um nódulo pulmonar solitário (NPS) é uma lesão esférica ou oval, menor ou igual a 3 centímetros de diâmetro rodeado pelo parênquima pulmonar. Esses nódulos podem possuir texturas semelhantes entre si, que são relativamente específicos para a textura do câncer, mas também podem possuir uma textura que se assemelha ao tecido normal (JÚNIOR et al., 2004).

Na Figura 1, podem-se observar imagens de TC de nódulos pulmonares, mais especificamente, a grande semelhança de textura entre diferentes nódulos pulmonares, responsável por grande parte dos erros nos diagnósticos. Na Figura 1A têm-se a imagem de um nódulo maligno denominado carcinoma de mutação negativa, e na Figura 1B têm-se a imagem de outro nódulo maligno, denominado de adenocarcinoma. Analisando-os, observa-se a grande semelhança de textura, sendo bastante passiva de erros, necessitanto de uma atenção e preocupação especial para o seu diagnóstico preciso.

A preocupação do diagnóstico de um nódulo pulmonar reside na possibilidade do mesmo ser um câncer (30% de chance) ou um câncer proveniente de outro lugar que se disseminou para o pulmão (10% de chance) (MACKEIVICZ, 2017), de modo que as condições de cada ambiente influenciam nesses percentuais.

A análise de ambiente que o paciente está incluído, refere-se a análise dos dados clínicos, como por exemplo, históricos de tabagismo ou de tuberculose, sendo fundamental

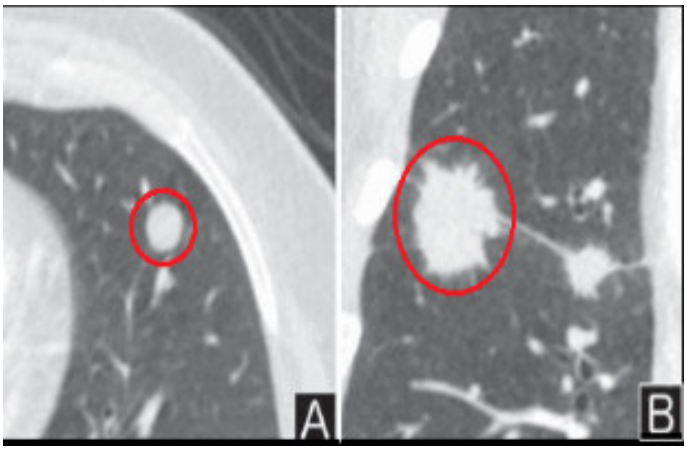

Figura 1 – Exemplos de nódulos.

Adaptado de (FINAK et al., 2016)

no favorecimento da presença de NPS, principalmente de nódulos de NSCLC (FAIAL, 2017).

Porém, nas últimas décadas, principalmente com o advento da TC helicoidal e mais modernamente de múltiplos detectores (*multi-slice*), equipamentos que têm permitido cada vez mais o diagnóstico de nódulos antes invisíveis ao Raio X simples.

Para (COLELLA et al., 2017) uma TC visualiza todas as características importantes do nódulo, que o distingue de alterações da parede toráxica, podendo identificar lesões adicionais, analisar melhor os ápices, regiões peri-hilares e ângulos costofrênicos, permitindo um melhor estudo do mediastino.

3.2 Radiomics

Antes do surgimento de novas tecnologias relacionadas aos métodos de diagnóstico por imagem, a única forma de se avaliar as imagens de TC, ressonância magnética (RM) e tomografia por emissão de prótons (PET) era por meio da interpretação do radiologista. No entando, hoje é possível converter imagens em dados e utilizá-los de forma a tornar o diagnóstico e o prognóstico de determinadas doenças mais precisos.

Por meio da computação de alto desempenho, é possível extrair rapidamente inúmeras características quantitativas de imagens tomográficas (imagens de TC ou tomografia por emissão de pósitrons - PET). A possibilidade de extração de características por meio de imagens, a fim de conseguir dar um diagnóstico, é chamada de Radiomics.

Os dados Radiomics podem ser usados para construir modelos descritivos e

preditivos que relacionam as características da imagem com os fenótipos dos tecidos testados no exame de biopsia feito em laboratório (LIU et al., 2017).

O potencial do Radiomics vem crescendo sua contribuição na oncologia à medida que o conhecimento, as ferramentas e as técnicas de análises evoluíram. Características de imagem quantitativa com base na intensidade, forma, tamanho ou volume oferecem informações sobre o fenótipo do tumor e microambiente, de maneira diferente dos fornecidos nos relatórios clínicos, já que irão trabalhar com valores quantitativos.

A Figura 2 mostra um exemplo de um processo Radiomics e o uso do Radiomics no suporte a decisão. A investigação do paciente requer informações de fontes diferentes para serem combinadas em um modelo coerente para descreverem onde é a lesão, como ela é classificada. Na etapa 1, Radiomics começa com a aquisição de imagens de alta qualidade. A partir dessas imagens, na etapa 2, uma região de interesse (ROI) que contém o tumor inteiro ou sub-regiões (isto é, habitats) dentro do tumor pode ser identificada. Na etapa 3, os nódulos são segmentados (com um software ou técnica) e são renderizados em três dimensões (3D). E por fim, na etapa 4, características quantitativas são extraídas desses volumes renderizados para gerarem relatórios (etapa 4B), que são colocados em um banco de dados junto com outros dados, como os dados clínicos (identificado como *Clínico* na etapa 4A da Figura 2), as características extraídas (identificado como *Radiomics* na etapa 4A da Figura 2) e genômicos (identificado como *genomics* na etapa 4A da Figura 2). Assim, esses dados são gerados para desenvolverem modelos de diagnósticos, preditivos ou prognósticos para resultados de interesse, e provarem que a partir das imagens, podem-se chegar a resultados sastifatórios.

3.3 Tomografia Computadorizada

A Tomografia Computadorizada (TC) surgiu em 1969, por meio dos trabalhos produzidos por Sir God Hounsfield, engenheiro elétrico da firma britânica Eletric and Musical Industries (EMI), e pelo professor de matemática da univerisdade de Tufts (Medford, Massachusets, EUA), Alan M. Cormack. Cada um ficou responsável por uma parte do projeto, sendo Hounsfield responsável pelo projeto do sistema eletrônico de detecção do feixe de radiação e processo de formação das imagens, e o matemático Cormack responsável pelos cálculos matemáticos necessários para a reconstrução das imagens tomográficas (MACKEIVICZ, 2017).

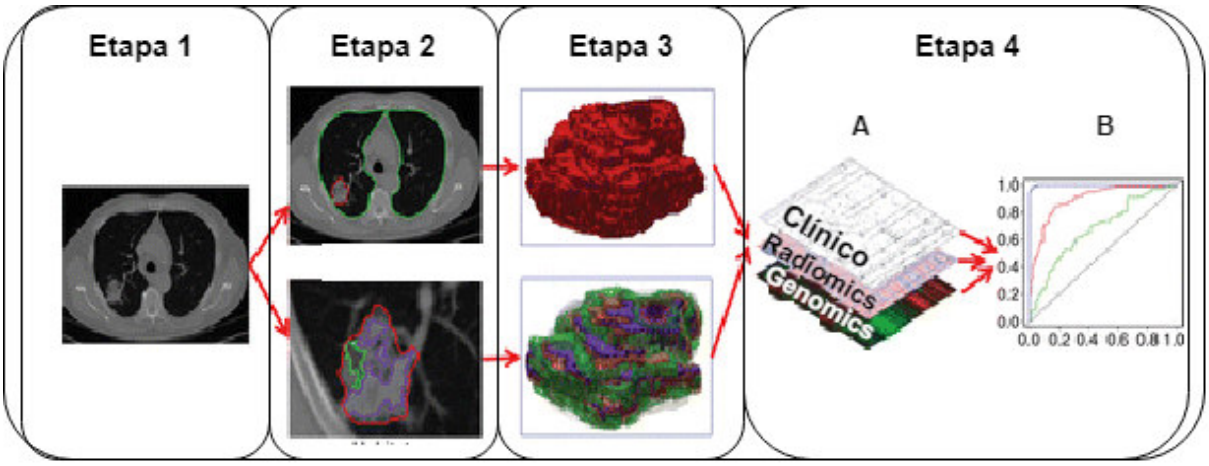

Figura 2 – Associação entre genótipos, Radiomics e o clínico.

Adaptado de (GILLIES; KINAHAN; HRICAK, 2015)

A TC tem origem a partir da atenuação dos feixes de raios X que atravessam o organismo em diversas projeções, e facilitando a reconstrução das estruturas internas do organismo (HOUNSFIELD, 1973).

O aparelho de TC possui um cabeçote que irradia feixes de raios X, que são coletados por uma série de sensores, ou detectores, posicionados no lado diamentralmente oposto, com o objetivo de capturar os feixes que atravessam o organismo estudado. Há alguns outros aparelhos utilizados para ajudar nos exames de TC, como por exemplo os colimadores, um equipamento cujo funcionamento visa restringir o ângulo de um feixe de radiação, que são responsáveis pela exposição do paciente à região a ser analisada no exame (bastante utilizado em forma de leque)(MACKEIVICZ, 2017). Esse aparelho influencia diretamente na sensibilidade de exame, pois a colimação do feixe está relacionada à espessura do corte.

O computador é utilizado para armazenamento dos dados capturados pelos sensores durante o movimento de rotação do conjunto cabeçote-detector (a posição angular, a intensidade de raios X, os valores de atenuação), e também para armazenar o deslocamento da mesa, em alguns décimos de milímetro. As estruturas internas de uma região selecionada do corpo humano são reconstruídas através de uma técnica chamada *filtered back projection*, cujo funcionamento se dá por meio da aplicação de uma série de equações matemáticas sobre os dados adquiridos. Uma matriz de valores de atenuação é consequência dessa aplicação, ou seja, uma matriz de valores de dose absorvida. Esses

valores de atenuação são exibidos em uma série de imagens seccionais em níveis de cinza do objeto varrido, obtendo assim, o diagnóstico.

Essas atenuações dos tecidos podem ser medidas na escala numérica expressa em unidades de Hounsfield (HU). As diferentes substâncias possuem diversos graus de atenuação medidos nessa escala. Nas condições-padrão de temperatura e pressão, a água destilada é definida como zero unidades de HU, enquanto a atenuação do ar é definida como -1000 HU.

3.3.1 Imagem 3D

Uma vez que o computador obtenha uma lista com todas as atenuações medidas pelos sensores, cálculos matemáticos são computados para identificar o valor de atenuação em cada ponto da matriz a ser gerada, de dimensão N x N. Esses pontos são as menores unidades formadoras da imagem digital, o *pixel* (*picture elements*). Durante o exame, cada *pixel* é percorrido por numerosos fótons de raios-X e a intensidade da radiação transmitida é medida pelos sensores. A escala de cinza do *pixel* é proporcional à atenuação sofrida na porção da área do tecido o qual o pixel representa, que corresponde a uma escala de -1.024 (menos atenuante) a 3.071 (mais atenuante) na escala HU. O *pixel* não tem uma dimensão definida, pois depende do tamanho do campo de visão (FOV, *field of view*) e da matriz de imagem. Ambos valores podem ser ajustáveis pelo técnico operador (SOARES et al., 2006).

A TC cria uma sequência ordenada de imagens bidimensionais que são obtidas das seções transversais do corpo humano, também conhecidas como cortes, fatias ou *slices*. Por meio de algoritmos de renderização, as fatias são agrupadas formando uma imagem tridimensional, permitindo a visualização de estruturas internas do corpo humano. A representação tridimensional do *pixel* é denominada *voxel* (*volume element*).

Após a aquisição dos dados volumétricos, a imagem 3D da TC pode ser pós-processada de acordo com as necessidades clínicas, como reformatação multiplanar, técnicas de renderização 3D e segmentação de imagens (CRUZ, 2007).

3.4 Arquivo DICOM

Esse tipo de arquivo teve origem devido o surgimento da TC, seguido pelo aumento do uso de computadores para aplicações clínicas e por outras modalidades digitais de

diagnósticos por imagem na década de 70. Assim, o *National College of Radiography* (NCR) e a *National Eletrical Manufacturers Association* (NEMA) observou que surgiu a necessidade de criar um padrão para a transferência de imagens e informações associadas entre dispositivos por diversos fornecedores (VARMA et al., 2007).

A NCR e a NEMA uniram-se para criar uma comissão conjunta em 1983 para desenvolver um padrão que seguisse algumas diretrizes, como promover a comunicação de informação de imagem digital independentemente do fabricante do dispositivo, para facilitar o desenvolvimento e a expansão de arquivamento de imagens e sistemas de comunicação e por último, permitir a criação de bases de informações de dados de diagnóstico que podem ser acessados por uma grande variadade de dispositivos distribuídos geograficamente.

Dar-se o nome desse padrão de *Digital Imaging Communications in Medicine* (DICOM), sendo uma melhoria às versões antigas do padrão ACR-NEMA (VARMA et al., 2007). As principais características desse padrão referem-se a sua aplicação em um ambiente de rede, aplicável a um ambiente de comunicação *off-line*, é estruturado como um documento de peça múltipla, especifica uma técnica estabelecida para identificar exclusivamente qualquer objeto de informação, especifica também como os dispositivos em conformidade com a norma que regem comandos e dados a serem trocados, além de introduzir objetos de informação explícita não só para imagens e gráficos, mas também para formas de onda, relatórios, impressão, etc.

A estrutura de um arquivo DICOM, é formada por um corpo (onde se encontra a imagem propriamente dita) e um cabeçalho (contendo informações úteis para a interpretação do conteúdo do corpo do arquivo, encontrando informações como tamanho da imagem, data do exame, etc), diferenciando-o da maioria dos padrões para imagens médicas (PAIM; RIGHI; COSTA, 2018).

3.5 Processamento Digital de Imagens

Pelo fato do Processamento Digital de Imagens (PDI) ser a manipulação de imagens a partir de computadores, onde a imagem e a saída do processo são imagens, é possível fornecer técnicas que facilitem a identificação e a extração de características dessas imagens, e posteriormente realizar a classificação. Sendo assim, a partir da utilização de algoritmos de PDI, tornou-se possível a aplicação de várias técnicas para a realização dos

objetivos de PDI.

Segundo mencionado por (MOHAN; POOBAL, 2018), o principal objetivo do PDI almeja a detecção, identificação, aprimoramentos e mensuração de objetos de uma cena, a partir de atividades multi-conceitos (como multiespectral, multitemporal e multiescalar) que abrangem princípios estatísticos para o reconhecimento de padrões, modelagem, sistemas de redes neurais, fotogrametria e interpretação de imagens, formando assim as etapas do PDI.

As etapas de PDI são bem definidas, são elas: aquisição de imagens digitais, pré-processamento, segmentação, representação, descrição e, reconhecimento e interpretação. Cada etapa, segue um *pipeline*, ou seja, a saída de uma etapa, irá servir de entrada para a próxima etapa, podendo essa entrada ser digital, ou não.

Na aquisição de imagens as imagens são obtidas por meio de um dispositivo ou sensor, e em seguida convertidos para uma representação adequada para o processamento subsequente. A imagem resultante desse processo pode apresentar deficiência quanto a suas estruturas, como por exemplo nas condições de iluminação ou características dos dispositivos. A etapa de pré-processamento tende a melhorar a qualidade da imagem a partir das técnicas para a redução de ruído, correção de contraste ou brilho e suavização de determinadas propriedades da imagem.

Na segmentação, visa-se identificar e extrair as áreas de interesse contidas na imagem, ou seja, a segmentação tem a função de dividir a imagem em seus componentes básicos.

Na etapa de representação e descrição é feita a extração de características, com o objetivo extrair propriedades para que possam ser utilizadas na discriminação entre as classes de objetos.

Por fim, na última etapa são realizados o reconhecimento e a interpretação. Define-se como reconhecimento, o processo que atribui um identificador aos objetos da imagem. Já a interpretação, consiste em definir um significado a um conjunto de objetos reconhecidos. Todo o conhecimento sobre o domínio do problema está codificado em uma base de conhecimento.

Para o desenvolvimento dessa dissertação, foram utilizadas as etapas de extração de características e reconhecimento de padrões, com base no objetivo apresentado na Seção 1.2.

3.5.1 Pré-processamento

Nesta seção são apresentadas as técnicas de melhoramento de imagens utilizadas no método proposto, visando aprimorar o desempenho da extração de características e do reconhecimento de padrões. Para o desenvolvimento do método foi utilizado a quantização para a discretização dos valores de amplitude.

3.5.1.1 Quantização

O processo de quantização visa obter a representação de uma imagem com L níveis para cada *pixel* com $L = 2^b$, sendo b o número de *bits* usados para o armazenamento do valor do *pixel*. Dessa forma, se houver a necessidade de quantizar para L' níveis de cinza, utiliza-se a quantização uniforme, que consiste em dividir a escala de cinza da imagem em intervalos iguais, onde cada intervalo é mapeado para um valor de cinza na imagem quantizada (GONZALEZ; WOODS, 2010). O mapeamento é calculado como:

$$q(i,j) = (2^b - 1)\frac{p(i,j) - I_{min}}{I_{max} - I_{min}} \tag{3.1}$$

sendo $q(i,j)$ o nível de cinza do *pixel* (i,j) da nova imagem quantizada, $p(i,j)$ o nível de cinza do *pixel* da imagem original, $[I_{max} - I_{min}]$ os limites inferiores e superiores da escala de cinza da imagem original e b o número de *bits* utilizados para armazenar cada *pixel* da imagem quantizada.

3.6 Análise por Textura

Textura é definida como a característica de uma região relacionada com os coeficientes de uniformidade, densidade, aspereza, intensidade, regularidade, entre outras características das imagens (KANUNGO; HARALICK; PHILLIPS, 1993). Outro conceito dado a textura, é dito por (HATT et al., 2017), onde textura é dado como a descrição intuitiva de medidas que quantificam as propriedades de rugosidade, regularidade e de suavidade. Em outras palavras, textura ainda pode ser considerada como uma dimensão que possui propriedades primitivas da tonalidade e a relação entre essas primitivas.

Segundo (HATT et al., 2017), a análise de textura é capaz de diferenciar regiões da imagem que apresentam as mesmas características intraclasses e extraclasses, possuindo

três abordagens principais usadas na área de processamento de imagens, que são estrutural, espectral e estatística.

A abordagem espectral baseia-se em peculiaridades do espectro de Fourier, sendo bastante utilizada principalmente na detecção de periodicidade global em uma imagem, a partir da localização de picos de alta energia no espectro. A abordagem Estrutural leva em consideração que as texturas são compostas de primitivas organizadas de forma aproximadamente regular e repetitiva. Já a última abordagem, a estatística, caracteriza a textura como um conjunto de medidas locais extraídas do padrão, favorecendo a descrição de imagens através de regras estatísticas que definem tanto a distribuição quanto a relação entre os diferentes níveis de cinza.

Nessa dissertação, foram utilizados os índices de diversidade filogenética e funcional, como medidas estatísticas de caracterização das texturas dos nódulos de NSCLC de TC.

3.6.1 Índices de Diversidade

No contexto da Ecologia de Comunidades e em várias aplicações da biologia da conservação, diversidade indica variedade de espécies, podendo ou não incluir informações sobre a importância relativa de cada espécie. A diversidade é um dos atributos mais fundamentais no estudo de comunidades e, para tal, há uma ampla gama de métodos para sua mensuração. Uma forma de mesurar a diversidade é baseada em índices estatísticos, chamados índices de diversidade.

Segundo Pielou um índice de diversidade é uma medida de dispersão qualitativa de uma população de indivíduos pertencentes a várias categorias qualitativamente diferentes (PIELOU, 1977). Existem basicamente dois tipos de índices de diversidade: os paramétricos e não paramétricos. Esta divisão se deve ao fato de um grupo de índices paramétricos dependerem de parâmetros de uma distribuição, enquanto os não paramétricos não dependem.

No caso dos índices paramétricos, os parâmetros mais utilizados são a riqueza de espécies e equabilidade.

A riqueza de espécies refere-se ao número de espécies em uma determinada área geográfica, região ou comunidade, ou seja, quanto maior o número de espécies, maior a riqueza de uma comunidade. Já a equabilidade é a abundância (riqueza) relativa em uma

determinada comunidade, ou seja, quanto mais próximas as abundâncias das espécies dentro de uma comunidade, maior a equabilidade.

Pode-se relacionar os índices de diversidade filogenética da biologia para o contexto de processamento de imagens, surgindo assim alguns conceitos. Para implementar essa ideia, o primeiro passo é fazer uma correspondência entre os termos usados na biologia e os utilizados neste trabalho. A Tabela 1 mostra um resumo da correlação entre o método proposto e a biologia.

Tabela 1 – Correlação entre os termos da biologia e o método proposto..

Biologia	**Metodologia**
Comunidade	Conjunto de nódulos, ou regiões do nódulo da imagem de TC
Espécie	cada Unidade de Hounsfield (UH) do nódulo
Indivíduo	Cada *voxels* do nódulo
Abundância relativa	Número de *voxels* encontrados no nódulo que possuem o mesmo valor UH
Riqueza de espécies	Número de intensidades encontradas no nódulo

Foram utilizados os índices de diversidade filogenética que criam uma relação evolutiva entre as espécies, e os índices de diversidade funcional que analisam a importância de cada uma das espécie para a comunidade, com o intuito de extrair as características, sendo descritos no decorrer do trabalho. Todos os índices são extraídos da imagem original de 12 *bits*, e das quantizações de 11, 10 e 9 *bits* sobre a imagem original, permitindo novas representações e novos indivíduos de uma mesma espécie e possivelmente também novas espécies para serem agregadas com as espécies e indivíduos já existentes na imagem original. Esses processos serão descritos detalhadamente nas próximas seções.

3.6.1.1 Índices de Diversidade Filogenética

A extração de característica do CADx foi adaptada do trabalho de (FILHO et al., 2017) para o contexto Radiomics, a partir da utilização da base NSCLS-Radiomics, com o intuito de mostrar que os índices de diversidade filogenética conseguem extrair informações que caracterizam a fisiopatologia dos tecidos, sendo desenvolvida a partir da análise de texturas que serão detalhados no decorrer do trabalho.

Segundo (MAGURRAN, 2013), diversidade filogenética é o coeficiente de uma comunidade que contém as relações filogenéticas das espécies. Para (OLIVEIRA et al., 2015), a textura é considerada como uma forma mais simples de representar os índices de uma diversidade em uma imagem, sendo a comunidade ou região representadas pela imagem.

Ainda, pode-se definir como diversidade filogenética a medida de uma comunidade que contém as relações filogenéticas das espécies. Considera-se a forma mais simples de representação dos índices de diversidade em imagens, quando a imagem representa a comunidade ou sua região (LOSOS, 2008).

A filogenia é o ramo da biologia onde estuda as relações evolutivas entre as espécies, os relacionamentos entre elas, com o intuito de determinar possíveis ancestrais comuns (BAXEVANIS; OUELLETTE, 2004).

Uma maneira de representar estas relações são as árvores filogenéticas, ou filogenia, onde os organismos são representados pelas folhas e os ancestrais pelos nós internos, como ilustrado na Figura 3B, sendo as folhas do cladograma as intensidades das imagens. Cada intensidade contém uma quantidade de indivíduos associados.

Na Figura 3B pode-se observar um exemplo de um cladograma criado a partir de uma imagem sintética (Figura 3A). Analisando-o, percebe-se que os nós do cladograma representam as espécies da imagem, além de uma associação do histograma com a criação da árvore.

Os índices de diversidade filogenética extraídos a partir das árvores filogenéticas são utilizados na biologia com o objetivo de comparar amostras de comportamento entre espécies que pertencem a diferentes regiões, e pode-se utilizar na área da computação para classificar em carcinoma de célula grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa ou não especificados.

Os índices de diversidade filogenéticos foram escolhidos devido a capacidade de caracterização de uma determinada região. Esses índices são divididos em 4 grupos, são eles: baseados na riqueza de espécies, baseados na distância entre pares de espécies, baseados na topologia e baseados em caminho mínimo. Cada grupo de índice mensura alguma propriedade que outro grupo não consiga, gerando resultados promissores. Vale salientar que todos os grupos de índices utilizam a fórmula da distância exemplificada na Equação 3.2.

$$D_{ij} = \begin{cases} j - i + 1, & \text{se } j \text{ igual a zero e } i \text{ diferente de } 0 \\ i - j + 2, & \text{se } j \text{ diferente de } 0 \text{ e menor que } i \text{ e } i \text{ diferente de } 0 \\ i - j + 2, & \text{se } i \text{ diferente de zero} \end{cases} \tag{3.2}$$

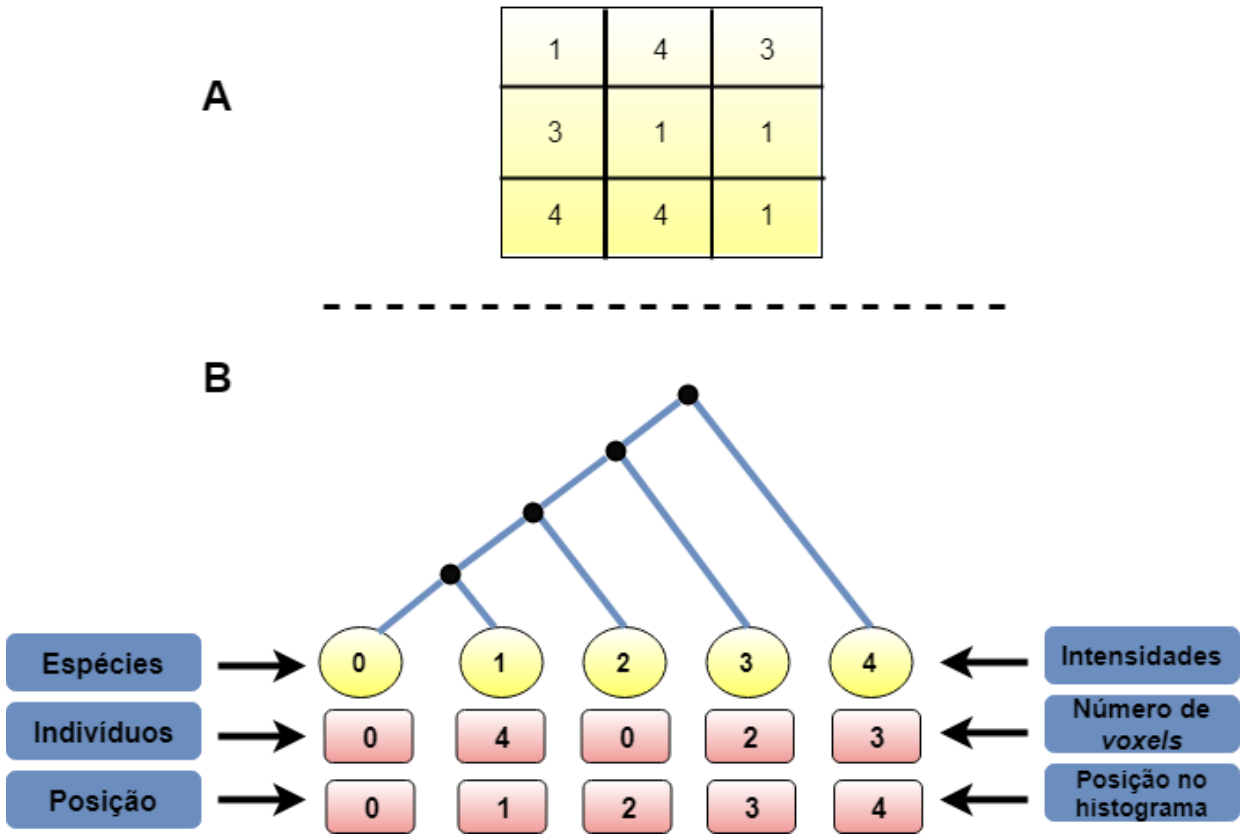

Figura 3 – Árvore filogenética na forma do cladograma a partir de uma imagem sintética.

onde D_{ij} é a distância da espécie i para a espécie j.

3.6.1.1.1 Índice de Diversidade Filogenética Baseado em Riqueza de Espécies

Os índices de diversidade filogenética baseados na riqueza de espécies, permitem medir a relação entre dois organismos escolhidos aleatoriamente em uma filogenia existente em uma comunidade, por meio dos índices de diversidade taxonômica e distinção taxonômica (PIEŃKOWSKI; WESTWALEWICZ-MOGILSKA, 1986).

O índice de diversidade taxonômica (Δ) representa a abundância das espécies e a relação taxonômica entre eles. Este índice é definido pela Equação 3.3.

$$\Delta = \frac{\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} D_{ij} * X_i * X_j}{[T_h * E_m]} \tag{3.3}$$

onde x_i é abundância de individuos (número de *voxels*) da *i-ésima* espécie, x_j representa a quantidade de individuos da *j-ésima* espécie, T_h representa o tamanho do histograma, E_m representa o bin central do histograma e D_{ij} é a distância entre a espécie i para a espécie j na classificação taxonômica.

Já o índice de distinção taxonômica (Δ^*) representa a distância taxonômica média entre dois indivíduos de espécies diferentes. Este índice é definido pela Equação 3.4.

$$\Delta* = \frac{\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} D_{ij} * X_i * X_j}{[X_i * X_j]} \tag{3.4}$$

onde x_i é abundância (número de *voxels*) da *i-ésima* espécie, x_j representa a abundância da *j-ésima* espécie, e D_{ij} é a distância entre as espécies i e j.

3.6.1.1.2 Índices de Diversidade Filogenética Baseados em Pares de Espécies

Diversos estudos ecológicos, dependem da riqueza de espécies como medida de biodiversidade. Porém, a utilização da riqueza de espécies como único reflexo da biodiversidade pode ter valor limitado, uma vez que não leva em relação as relações filogenéticas (VANE-WRIGHT; HUMPHRIES; WILLIAMS, 1991). Para resolver isto, pode-se utilizar as relações de distância entre pares de espécie.

As relações de distância entre pares são baseadas na matriz de distância levando em consideração todas as espécies de uma comunidade. As distâncias podem se basear no comprimento dos ramos das relações filogenéticas compostas em dados moleculares (SOLOW; POLASKY; BROADUS, 1993), em diferenças morfológicas ou funcionais (IZSÁK; PAPP, 2000), ou, caso os cumprimentos dos ramos não sejam conhecidos, sobre a quantidade de nós que separam cada pares de espécies (FAITH, 1994).

Os índices que se baseiam nas distâncias entre pares de espécies são: entropia quadrática intensiva (IZSÁK; PAPP, 2000), entropia quadrática extensiva (IZSÁK; PAPP, 2000), distinção taxonômica média (PIENKOWSKI et al., 1998) e a diversidade pura (WEITZMAN, 1992).

No trabalho de (IZSÁK; PAPP, 2000) é proposto o índice entropia quadrática intensiva (J) com o objetivo de ligar os índices de diversidade e os índices de medidas de biodiversidade. Este índice é definido como a ligação entre o número de espécies e suas

relações taxonômicas, quando tiverem os mesmos valores de abundância de espécies. A Equação 3.5 define este índice.

$$J = [\frac{\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} Dij}{s^2}] \tag{3.5}$$

onde D_{ij} representa a distância entre as espécie i e j, e s representa o número de espécies.

As medidas de diversidade possuem a característica de monotocidade, representando o aumento do valor do índice, sempre que for adicionado uma nova espécie c a um grupo de espécie K, que faz com que o índice J não seja totalmente preciso. Já o índice chamado de entropia quadrática extensiva F (IZSÁK; PAPP, 2000) representa a soma das diferenças das espécies. Esse índice é representado na Equação 3.6.

$$F = \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} Dij \tag{3.6}$$

onde D_{ij} representa a distância entre as espécies i e j.

Outro grupo de índices leva em consideração as relações taxonômicas, porém esses podem ser adaptados para serem usados na filogenia (SCHWEIGER et al., 2008), são eles: índice de distinção taxonômica média ($AvTD$) (CLARKE; WARWICK, 1994) e distinção taxonômica total (TTD) (PIENKOWSKI et al., 1998). A distinção taxonômica entre quaisquer pares de espécies escolhidas de maneira aleatória é representado pela Equação 3.7. Os índices $AvTD$ e TTD são representados pelas Equações 3.7 e 3.8, respectivamente.

$$AvTD = \frac{\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} D_ij}{\frac{s(s-1)}{2}} \tag{3.7}$$

$$TDD = [\frac{\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} D_ij}{(s-1)}] \tag{3.8}$$

onde D_{ij} indica a distância entre as espécies i, j representando o número de espécies e s representa a quantidade de espécies.

O último índice pertencente a esse grupo representa a verificação do valor da distância de uma espécie para o seu vizinho mais próximo (WEITZMAN, 1992). As

Equações 3.9 e 3.10 representam este índice.

$$d_{imin} = MIN[d_{ij} \forall\ j\ ] \tag{3.9}$$

$$Dd = \sum_{i=1}^{n-1} d_{imin} \tag{3.10}$$

onde d_{imin} representa o valor mínimo (MIN) de D_{ij} a cada interação i, e n representa a quantidade de espécies.

3.6.1.1.3 Índices de Diversidade Filogenética Baseados em Topologia

Além dos índices baseados na riqueza de espécies e baseado na distância entre pares de espécies, existem os índices filogenéticos baseados na topologia.

Um dos primeiros estudos a propor a aplicação de métodos baseados em topologia foi o estudo realizado por (VANE-WRIGHT; HUMPHRIES; WILLIAMS, 1991), que leva em consideração a ordem filogenética dentro de um determinado grupo. Nesse tipo de índice, cada espécie de uma comunidade é ponderada pelo número de nós entre a espécie e a raiz da árvore filogenética (cladograma). Assim, atribui-se os maiores pesos às espécies que possuem a maior distância da raiz.

Os índices topológicos utilizados para descrever a textura das regiões de interesse foram a soma básica dos pesos (Q) e soma básica dos pesos normalizdos (W). Esses índices visam o relacionamento da árvore das espécies presentes com toda a comunidade, fazendo com que esses índices mencionem o grau de parentesco (similiaridade) entre as espécies (KEITH et al., 2005).

O índice Q representa o somatório das contribuições de cada espécie para a diversidade, dado pela divisão entre a totalidade de nós para todo o grupo e pelo número de nós entre a raiz e uma determinada espécie. Pode-se observar a definição de Q na Equação 3.11.

$$Q = \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} D_{ij} \tag{3.11}$$

onde Q representa a contribuição de cada espécie para a comunidade.

O índice W representa o peso normalizado para cada espécie, ou seja, a divisão entre os valores de Q de cada espécie pelo valor de D_{min}. Este índice é definido na Equação

3.12.

$$W = \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \frac{D_{ij}}{D_{imin}} \tag{3.12}$$

onde D_{imin} representa o valor da distância mínima utilizada no cálculo do índice da Equação 3.10, a cada cálculo da distância entre as espécies i e j.

3.6.1.1.4 Índices de Diversidade Filogenética Baseados em Caminho Mínimo

Os índices de diversidade filogenética baseados em caminho mínimo representam o último grupo utilizado na metodologia deste trabalho. Esse tipo de índice possui a característica de aferir a distância entre as espécies e a comunidade, ou seja, quanto menor o caminho, menor a quantidade de espécies, e menor a diversidade.

O primeiro índice desse grupo é uma medida quantitativa da diversidade filogenética (PD_{NODE}), que indica o total mínimo do total de ramos filogenéticos, necessários para medir um *taxón* e uma árvore filogenética. Segundo (FAITH, 1992) e (FILHO et al., 2016), quanto maior os valores do PD_{NODE}, maior a diversidade. As Equações 3.13, 3.14, 3.15 e 3.16 representam esse índice.

$$A_i = \frac{\sum_{i=1}^{n-1} I_i}{D_{INDij}} \tag{3.13}$$

$$C_1 = \sum_{i=1}^{n-1} D_{mom} * A_i \tag{3.14}$$

$$C_2 = \sum_{i=1}^{n-1} A_i \tag{3.15}$$

$$PD_{NODE} = \frac{C_1}{C_2} \tag{3.16}$$

onde n indica o número de espécies, D_{INDij} representa a distância individual em cada interação, D_{mom} representa o somatório da distância no momento da interação.

O segundo índice do grupo, denominado de PD_{ROOT}, inclui os ramos de base, representando o número de nós dentro do máximo caminho enraizado (RODRIGUES;

GASTON, 2002). A Equação 3.17 representa esse índice.

$$PD_{ROOT} = \frac{T_h * E_m}{\Delta^*} \tag{3.17}$$

onde PD_{ROOT} indica o número de nós dentro do caminho, *Delta* o índice calculado na Equação 3.4, T_h o tamanho do histograma, e E_m a intensidade do meio do histograma.

O terceiro e último índice deste grupo é a diversidade filogenética média (*AvPD*). A Equação 12 representa este índice.

$$AvPD = \frac{PD_{NODE}}{s} \tag{3.18}$$

onde s indica o total de espécies.

3.6.1.2 Índices de Diversidade Funcional

As ciências biológicas há muito tempo vêm estudando a diversidade funcional, e o interesse dos pesquisadores por este tema está crescendo muito nos últimos anos em diversos campos da ecologia e em estudos com diversos grupos taxonômicos, indicando que o conceito está ganhando grande importância. Em virtude da potente relação entre a diversidade funcional e o funcionamento e manutenção dos processos das comunidades (PETCHEY; GASTON, 2006), é importante definir de maneira precisa, o conceito de diversidade funcional. Segundo (TILMAN, 2001), diversidade funcional é definida como sendo o valor e a variação das espécies e de suas características que influenciam no funcionamento das comunidades.

Dessa maneira, medir a diversidade funcional (DF) implica em medir a diversidade de características funcionais dos indivíduos, que são componentes dos fenótipos dos organismos que influenciam os processos dentro da comunidade. Para facilitar melhor o entendimento, a Figura 4 ilustra a nomenclatura da árvore filogenética de uma comunidade de plantas e suas diversidades. Essa definição é adotada neste trabalho, juntamente com as medidas de diversidade funcional e as associações descritas na Tabela 1.

Medir a DF, consiste na soma dos comprimentos dos braços de um dendograma funcional, ou seja, um dendograma gerado a partir de uma matriz de espécies x características funcionais (CIANCIARUSO; SILVA; BATALHA, 2009). O cálculo da DF é o mais simples e baseia-se em fundamentos da análise de agrupamento e segue os seguintes passos:

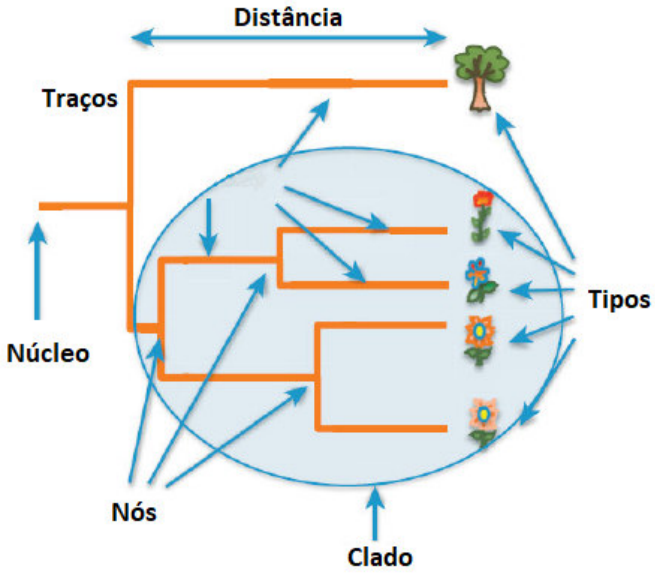

Figura 4 – Nomenclatura da árvore funcional.

Adaptado de (NETO et al., 2017)

1º Obter uma matriz funcional (espécies x características funcionais);

2º Converter a matriz funcional em uma matriz de distâncias;

3º Realizar o agrupamento da matriz de distâncias para produzir um dendograma;

4º Calcular o comprimento total das ramificações do dendograma.

Para construir o dendograma a partir dos *voxels* pertencentes a uma ROI, utilizou-se o algoritmo do Otsu (OTSU, 1979) para separar os grupos de *voxels* a partir da similiaridade dos tons de cinza, formando as comunidades. Cada nó representa os grupos separados pelo algoritmo de Otsu. Na Figura 5B observa-se o nódulo (na realidade, é o histograma desse nódulo ou região que é dividio pelo Otsu, utilizando a imagem de um nódulo apenas para fins de melhor entendimento) dividido pelo limiar gerado pelo Otsu (representado pela linha vermelha), devido trabalhar melhor com intervalos bimodais para separação de dois objetos. As intensidades (espécies) menores que o limiar, são inseridos na comunidade da esquerda (no dendograma, no nó da esquerda) e as intensidades menores que o limiar, são inseridos na comunidade da direita (no dendograma, no nó da direita). Na Figura 5C, observa-se os primeiros nós formados pelo limiar do Otsu. Na Figura 5D, observa-se novamente os limiares do otsu dividindo sucessivamente as comunidades, e os nós no dendograma. O algoritmo de Otsu finaliza sua divisão naquela comunidade, quando esta tiver no máximo duas intensidades. São criados dendogramas para cada uma das quantizações.

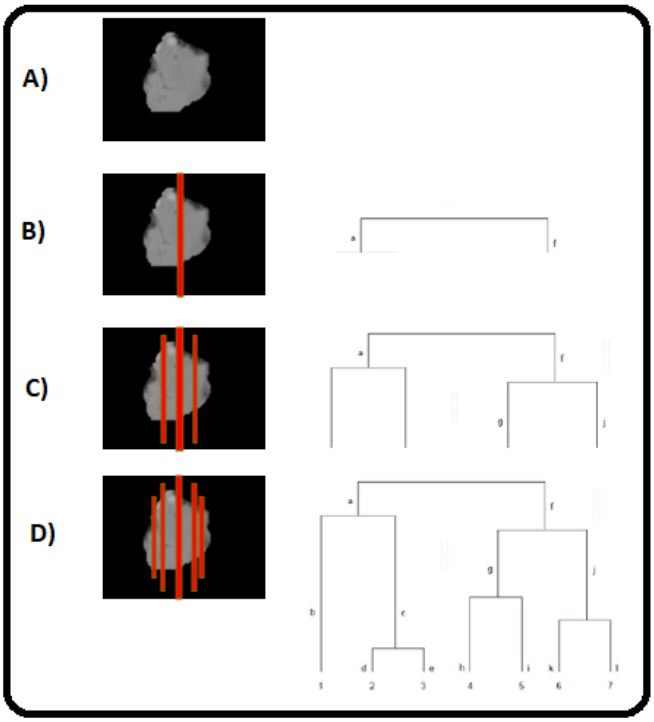

Figura 5 – Árvore funcional na forma de dendograma criado a partir do Otsu.

3.6.1.2.1 Diversidade Funcional Abundante

A diversidade Funcional Abundante (FADa) estima a dispersão pela soma das distâncias pareadas entre as espécies no espaço multidimensional, enquanto que uma medida semelhante o faz pela média dessas distâncias (HEEMSBERGEN et al., 2004). Outra medida proposta baseia-se na entropia quadrática de (RAO, 1982) que permite a inclusão da abundância das espécies. A Equação 3.19 representa o cálculo do índice FADa.

$$FADa = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n-1} D_{ij} * a_i * a_j \tag{3.19}$$

onde D_{ij} é a distância levando em consideração o dendograma e seus valores, e a representa a abudância de espécies.

A equação da distância utilizada nos índices de diversidade funcional, é definida na Equação 3.20, onde D_{ij} representa a distância euclidiana entre as posições das especies P_i e P_j no dendograma criado, e n é a quantidade de espécies do dendograma. A abundância a é dada através da soma das probabilidades de cada espécie daquele volume, sendo o

valor total igual a 1, como mostrado na Equação 3.21.

$$D_{ij} = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} (P_i - P_j)^2 \tag{3.20}$$

$$\sum_{i=1}^{n-1} a_i = 1 \tag{3.21}$$

3.6.1.2.2 Índice de Diversidade Funcional de Abundância de Espécies

A extração utilizando o índice de diversidade funcional abundante da espécie (FADe) é feito através da Equação 3.22, que leva em consideração a abundância da quantidade de *voxels* da mesma espécie no dendograma. A Equação 3.22 representa esse índice.

$$FADe = \sum_{i=1}^{n-1} \sum_{j=i+1}^{j=n-1} D_{ij} * a_i * a_j \tag{3.22}$$

onde D_{ij} é a distância levando em consideração o dendograma e seus valores, e a, representa a abundância de espécies. O cálculo da distância é observado na Equação 3.23, onde I_i e I_j, representa os valores das intensidades i e j, respectivamente. I_{indi} e I_{indj} representa a quantidade de indivíduos nas intensidades i e j, respectivamente. A Equação 3.21 é utilizada para calcular o valor de a, representando a abundância daquela espécie de indivíduos.

$$D_{ij} = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=1}^{j=n-1} (I_i - I_j)^2 \tag{3.23}$$

3.6.1.2.3 Índice de Diversidade Funcional Abundante de Pixels

Esse índice foi adaptado do trabalho de (NETO et al., 2017) para ser utilizado para *voxels*. O índice de diversidade funcional de diversidade de *pixels* (FADp) é definido pela Equação 3.24.

$$FADp = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n-1} D_{ij} * ap_i * ap_j \tag{3.24}$$

onde D_{ij} é a distância levando em consideração o dendograma e seus valores, e ap representa a abundância de espécies. Nesse índice, será aplicado o cálculo da distância de acordo com a Equação 3.20.

O cálculo da soma das abundâncias individuais de cada espécie deve ser 1, como obervado na Equação 3.25.

$$\sum_{i=1}^{n-1} ap_i = 1 \tag{3.25}$$

3.6.1.2.4 Índice de Diversidade Funcional

O índice de diversidade funcional (DF) representa a quantidade de braços do dendograma.

3.6.1.2.5 Índice de Diversidade Funcional Abundante Relacionando Valores de Voxels e Distâncias entre Espécies

O índice de diversidade funcional abundante relacionando valores de *voxels* com a distância das espécies (DFaPe) é definido pela Equação 3.26.

$$DFaPe = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n-1} D_{ij} * ape_i * ape_j \tag{3.26}$$

onde D_{ij} é a distância entre pares de espécie, e a é abundância individual das espécies.

A abundância individual ape_i deve ser calculado através da abundância de cada espécie, dividas pela soma das multiplicações dos valores de cada espécie pela sua quantidade de indivíduos, como definido na Equação 3.27 .

$$ape_i = \frac{a}{e * Ind} \tag{3.27}$$

onde a é abundância individual de cada espécie, e é o valor da intensidade daquela espécie e Ind é a quantidade de indivíduos daquela espécie.

3.6.1.2.6 Entropia Quadrática Funcional

O índice de entropia quadrática funcional (EQF) leva em consideração os braços do dendograma (FADq) e é representado pela Equação 3.28.

$$EQF = \sum_{i=1}^{E-1} \frac{FADa_i}{(E)^2} \tag{3.28}$$

onde *FADa* representa o valor do índice calculado na Equação 3.19, e E representa a quantidade de espécies na comunidade.

3.6.1.2.7 Índice de Diversidade Funcional Abundante Considerando Espécies e Voxels

O índice de diversidade funcional abundante que leva em consideração as espécies e os *voxels* (FADs) é definido na Equação 3.29.

$$FADs = \sum_{i=1}^{E-1} \frac{FADe_i}{(E)^2} \tag{3.29}$$

onde *FADs* representa o índice calculado na Equação 3.22 e E representa a quantidade de espécies na comunidade.

3.6.1.2.8 Soma das Distâncias Funcionais

No índice de soma das distâncias funcionais (FAD) soma-se todas as distâncias entre as posições na árvore nos pares de espécies. A Equação 3.30 representa esse índice.

$$SDF = \sum_{i=1}^{E-1} \sum_{j=i+1}^{E-1} D_{ij} \tag{3.30}$$

onde, D_{ij} representa a distância entre as posições de cada espécie na árvore. Utiliza-se a Equação 3.20 para calcular a distância.

3.6.1.2.9 Soma de Intensidades Funcionais

O índice de soma das distâncias entre pares de intensidades (SIF) funcionais da árvore funcional. A Equação 3.31 define este índice.

$$SIF = \sum_{i=1}^{E-1} \sum_{j=i+1}^{E-1} D_{ij} \tag{3.31}$$

onde D_{ij} representa a distância entre pares de intensidades funcionais da árvore funcional. Utiliza-se a Equação 3.23 para calcular a distância.

3.7 Reconhecimento de Padrões

O reconhecimento de padrões é o campo da ciência que tem como objetivo a classificação de objetos em um determinado número de categorias ou classes, a partir da observação de suas características (THEODORIDIS; KOUTROUMBAS, 2008). Dessa forma, visa construir uma representação mais simples de um conjunto de dados através de suas características mais relevantes, possibilitando sua partição em classes (DUDA; HART; STORK, 2000).

O reconhecimento de padrões tem uma ampla aplicação em astronomia, medicina, robótica, reconhecimento de fala, sensoriamento remoto por satélites e etc. O processo de reconhecimento de padrões é composto por duas etapas: classificação e reconhecimento. Durante a etapa de classificação uma amostra de uma população qualquer é particionada em grupos chamados classes. Na etapa de reconhecimento uma amostra desconhecida, porém pertencente à mesma população, é reconhecida como integrante de uma das classes criadas anteriormente (LOONEY et al., 1997).

As técnicas de classificação são compostas por dois grupos: supervisionada e não-supervisionada. A classificação supervisionada consiste em um processo prévio de treinamento de um classificador para o conhecimento dos padrões desejados. Posteriormente, o classificador é capaz de identificar a classe de qualquer objeto desconhecido da mesma população de treinamento. Já na classificação não supervisionada não há informações prévia sobre as classes às quais os padrões de amostras pertencem (PEDRINI; SCHWARTZ, 2008).

Neste trabalho foi utilizado a técnica de classificação supervisionada Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) e *Random Forest* (RF), visando

classificar informações, ou padrões, baseado em um conhecimento prévio ou em informações estatísticas extraídas dos padrões encontrados,das classes de nódulos de NSCLC. Mais detalhes são apresentados nas subseções seguintes.

3.7.1 Classificadores

O método utiliza o SVM e RF (BIAU, 2012) para o reconhecimento dos padrões das classes de nódulos de NSCLC. A seguir são apresentados conceitos desses classificadores.

O RF é uma técnica baseada na construção de um conjunto de árvores que são classificadores fracos e posteriormente são usados para construir um classificador estável e forte, melhor que a árvore média criada. O RF é um algoritmo flexível que mesmo sem ajuste de hiperparâmetros apresenta um ótimo resultado na maioria das vezes. Além disso, é possível utilizá-lo tanto para tarefas de classificação como de regressão (BREIMAN, 2001).

O SVM é uma técnica baseada na teoria do aprendizado estatístico, usado para encontrar hiperplanos ideais para as classes linearmente separáveis e não separáveis. Em seguida, os dados de entrada são mapeados para um espaço dimensional superior (usando função do *kernel*) e definido um hiperplano de separação (VAPNIK; VAPNIK, 1998). O SMV utiliza o princípio de minimização do risco estrutural, que se baseia no fato de que a taxa de erro de uma máquina de aprendizado nos dados de teste é limitada pela soma da taxa de erro de treinamento e por um termo que depende da dimensão Vapnik-Chervonenkis (dimensão VC) (ZHUANG; DAI, 2006).

3.8 Seleção de Características

Em aplicações de visão computacional é bastante recorrente a utilização de técnicas que geram um grande número de características, que muitas vezes são desnecessárias para realizar a separação dos indivíduos em suas classes. Portanto, a escolha do conjunto de características relevantes é importante para simplificar o modelo e aumentar seu poder de generalização. Neste contexto, realiza-se a seleção das características que é uma etapa opcional no processo de reconhecimento de padrões.

A seleção de características visa reduzir a dimensionalidade do espaço de características, selecionando as características mais relevantes. As principais razões para tal redução são: os custos de medição, pois quanto menos características, mais rápido será o

classificador, e a precisão do classificador. Por este motivo, faz-se necessário a utilização de algoritmos de seleção de características que propiciem a obtenção de representações dos padrões de forma robusta. Sendo assim, o algoritmo *Greedy Stepwise* (AMBELU; LOCK; GOETHALS, 2010) foi o escolhido para efetuar este procedimento.

3.8.1 Algoritmo *Greedy Stepwise*

O *Greedy Stepwise* é um algoritmo guloso de paradigma algorítmico, que segue a heurística de resolução de problemas fazendo a escolha localmente ótima em cada estágio (GUTIN; YEO; ZVEROVICH, 2002), com a intenção de encontrar um ótimo global. Em muitos problemas, uma estratégia gulosa geralmente não produz uma solução ótima, mas mesmo assim uma heurística gananciosa pode produzir soluções ótimas localmente que se aproximam de uma solução globalmente ótima em um período de tempo razoável, utilizando uma análise discriminante.

A análise discriminante é uma técnica comum no âmbito de aplicações envolvendo métodos de reconhecimento de padrões. Essa técnica utiliza informações das classes associadas a cada padrão para extrair linearmente os atributos mais discriminantes através da computação de uma combinação linear de n variáveis quantitativas que mais eficientemente separam grupos de amostras em um espaço m-dimensional, fazendo com que maximixe as diferenças inter-classes e minimize as diferenças intra-classes (NASCIMENTO, 2012).

O problema é então reduzido a achar um vetor que melhor represente esse conjunto de dados. A ideia básica que envolve a análise discriminante é determinar o quão as classes são distintas em relação à média de um grupo de variáveis, e em seguida usar a média desse grupo para separá-los entre si, e também para associar ao grupo mais próximo.

Dois métodos computacionais podem ser utilizados para determinar uma função discriminante: o método simultâneo (direto) e o método *stepwise*.

O método simultâneo utiliza uma função discriminante que é calculada baseada em todo conjunto de variáveis independentes, sem consideração do poder discriminatório de cada variável (HAIR, ANDERSON e BLACK, 2005).

Assim, esse algoritmo executa uma busca gulosa, no sentido *forward* ou *backward*, através do espaço de busca dos subconjuntos de atributos. A busca pode iniciar com nenhum/todos os atributos ou de um ponto arbitrário no espaço, e encerra quando todas

as inclusões ou exclusões de atributos resultarem em uma diminuição na avaliação. Para tanto, este algoritmo encontra, utilizando uma análise discriminante, as características que discriminam melhor as cinco classes de nódulos de NSCLC para o conjunto de características geradas, que contém menos redundâncias que poderiam prejudicar os classificadores durante a etapa de validação (AMBELU; LOCK; GOETHALS, 2010).

3.9 Métricas de Validação

O resultado da etapa de classificação é a identificação das classes de nódulos de NSCLC. Após o processo de reconhecimento de padrões existe a necessidade de validar os resultados produzidos, através das métricas de desempenho. Essa atividade tem por finalidade avaliar o desempenho do método desenvolvido por meio de uma análise estatística dos resultados.

Para avaliar o desempenho do método proposto foram utilizadas estatísticas tipicamente aplicadas em sistemas CADx para análise de imagens médicas: acurácia (Acc) (DUDA, 1973) e Desvio Padrão (StdDev) em relação a acurácia (VIERA; GARRETT et al., 2005).

Outra medida utilizada é o índice Kappa, que é um coeficiente de concordância utilizado em escalas nominais. O índice Kappa mede a relação entre concordância e causalidade, e também o desacordo esperado, indicando quão legítimas são as interpretações (ROSENFIELD; FITZPATRICK-LINS, 1986). Esse índice é recomendado como uma medida que representa integralmente uma matriz de confusão. A categorização dos níveis de acurácia de classificação, pelo índice Kappa, pode ser visualizada na Tabela 2, conforme definido por Landis e Koch (LANDIS; KOCH, 1977).

Tabela 2 – Níveis de precisão de classificação, segundo o índice Kappa.

Índice Kappa (K)	**Qualidade**
$K < 0.2$	Ruim
$0.2 \leq K < 0.4$	Razoável
$0.4 \leq K < 0.6$	Bom
$0.6 \leq K < 0.8$	Muito bom
$K \geq 0.8$	Excelente

O desempenho do nosso método também é avaliado usando a curva ROC, que indicam a taxa positiva verdadeira (sensibilidade) como uma função da taxa de falsos positivos (1 - especificidade) (FAWCETT, 2006). A curva é construída variando o limiar

do classificador e avaliando os resultados gerados. A principal informação extraída de uma curva ROC é a área sob a curva (*Area Under Curve* - AUC) e quanto maior a área, ou seja, mais próximo de 1 (equivalente a 100%), melhor é o desempenho do classificador.

4 MATERIAIS E MÉTODOS

Este capítulo apresenta as etapas usadas na metodologia proposta para a classificação de nódulos de NSCLS, por meio de imagens de TC. Ela está dividida em 6 etapas, sendo as etapas de 1 a 5, a metodologia de classificação, e a 6º etapa a avaliação, como detalhada na Figura 7. Em síntese, a primeira etapa às imagens de TC de NSCLC. Na segunda etapa é realizada a extração da lesão. O pré-processamento é feito na terceira etapa. Na quarta etapa é realizada a extração de características utilizando índices de diversidade filogenética e funcional. Após esta etapa, é feita a classificação. Na sexta etapa a metodologia é avaliada.

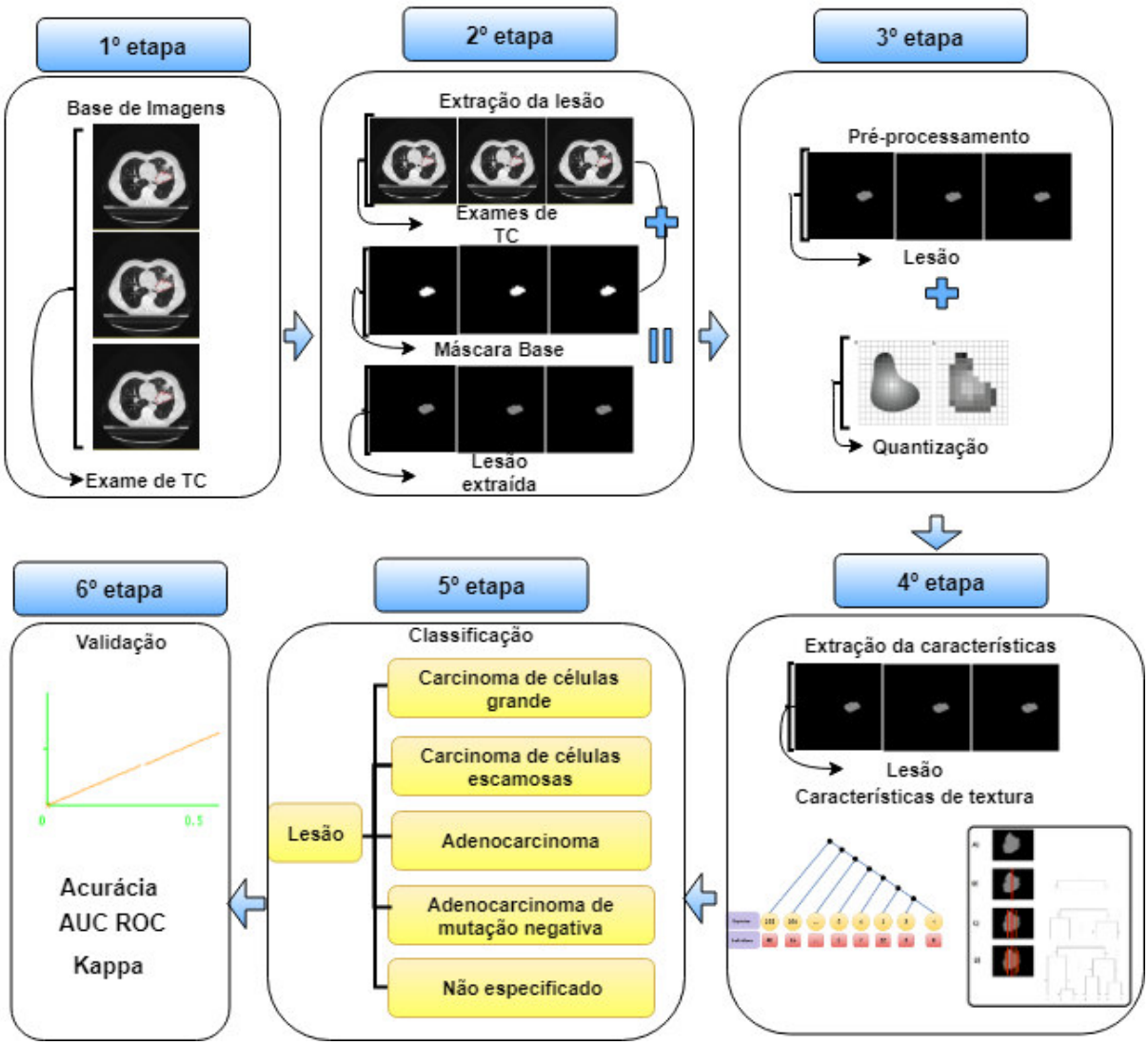

Figura 6 – Fluxograma da metodologia proposta.

4.1 Base de imagens

Nesse trabalho, foram utilizadas as imagens da base pública NSCLC-Radiomics disponibilizadas pelo *Cancer Imaging Archive*, criadas a partir de um repositório de imagens de câncer (AERTS et al., 2015) com o intuito de pesquisas e treinamentos para predição do prognóstico (aqui usada para classificação) de nódulos de NSCLC. Essa base de imagens contém 422 exames de nódulos de NSCLC (cada exame representa um paciente e possui de 120 a 180 fatias), sendo que cada exame pode pertencer a uma das 5 classes: carcinoma de célula grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados (AERTS et al., 2015).

Essa base contém um arquivo individual para cada exame com informações sobre cada nódulo presente no exame, dispostas em um arquivo denominado de *Radiotherapy Structure Set* (RT-STRUCT) com as marcações da base (localização dos nódulos, classe que o paciente está incluso, entre outros) e utilizadas para extração das características pelos algoritmos desenvolvidos no método proposto (AERTS et al., 2015). Esta base foi utilizada para extração das características pelos algoritmos desenvolvidos no método proposto.

4.2 Extração da lesão

As marcações são feitas baseadas em protocolos diferentes. Primeiramente cada especialista analisa os exames de forma independente. Em seguida, os resultados das análises são apresentados juntos para cada um dos especialistas. Durante essa fase, eles analisam e refazem livremente as marcações (AERTS et al., 2015).

Os nódulos foram extraídos a partir das marcações dos especialistas, totalizando 319 dos 422 exames de TC existentes, devido os demais estarem sem marcações.

4.3 Extração de características

As características extraídas são baseadas em textura descritas na Subseção 3.6. Com as medidas de textura, podem-se avaliar os casos de estruturas que possuem propriedades semelhantes e definir a qual classe cada caso pertence.

Nessa seção, são exemplificados os cálculos dos índices extraídos. Na Figura 3A é ilustrado uma imagem sintética que foi utilizada para criar um cladograma (Figura 3B),

utilizado como exemplo para os cálculos dos índices de diversidade filogenética.

Observa-se na Figura 7 um exemplo de uma imagem 3x3 com seus valores de *voxels* e o seu respectivo dendograma criado a partir dessa imagem. Nas folhas do referido dendograma, estão as espécies representadas pelos valores de *voxels*. Os grupos são representados pelas espécies que pertencem ao mesmo nó. Essa técnica é utilizada para criar os dendogramas das imagens dos nódulos de NSCLC, com o objetivo de extrair os índices de diversidade funcional. Vale salientar que todos os índices, utilizam a fórmula da distância exemplificada na Equação 3.2.

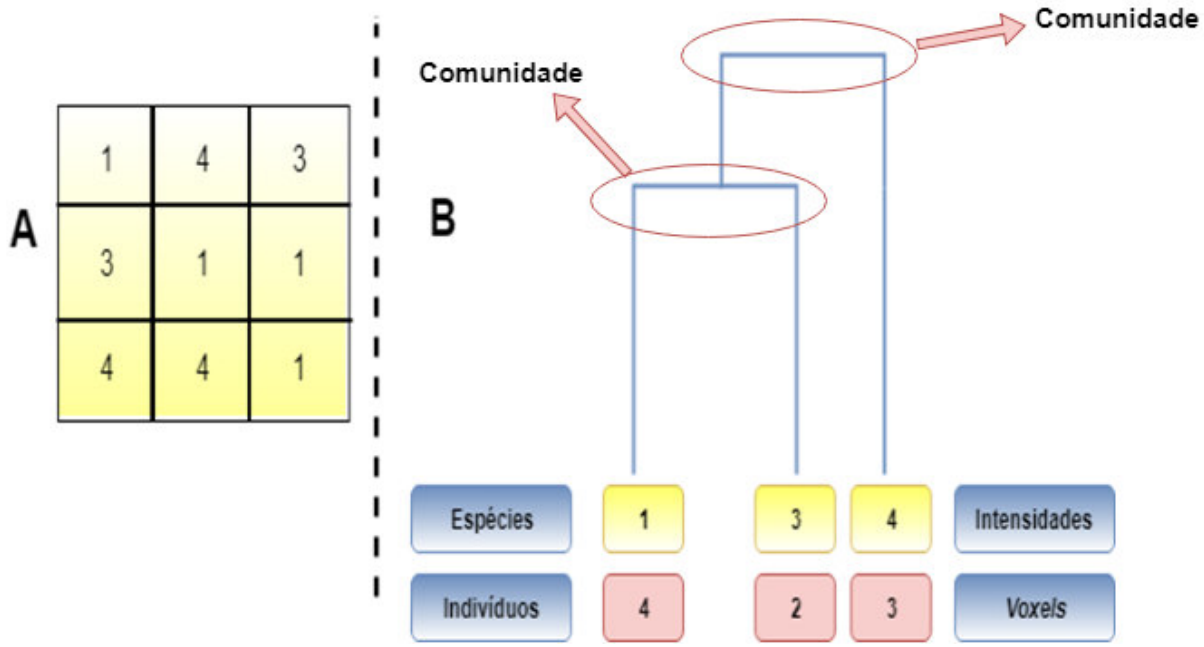

Figura 7 – Imagem sintética para exemplificar o dendograma

A Tabela 3 exemplifica o cálculo da distância para facilitar possíveis reproduções (nível de implementação) e com o intuito de melhorar o entendimento, sendo a mesma equação de distâncias tanto para o cladograma (Figura 3B) como para o dendograma (Figura 7B) a partir da imagem sintética (Figura 3A e/ou Figura 7A). Vale salientar que os cálculos das distâncias na Tabela 3 foram reproduzidos dos trabalhos de (NETO et al., 2017) e (FILHO et al., 2016), referências base para as adaptações dos índices de diversidade filogenética e funcional para a abordagem Radiomics, logo, algumas condições expostas nessa tabela foram definidas nessas referências.

Na Tabela 3 pode-se observar três colunas, sendo a primeira coluna, o ID referente a aquela linha; a segunda representa os valores assumidos pelas variáveis i e j no momento

do cálculo da distância; e a terceira coluna representa a justificativa para o cálculo da distância. A variável X indica quando a distância não é calculada para aquela interação.

Tabela 3 – Exemplo do cálculo da distância filogenética e funcional.

ID	D_{ij}	**Valores de i e j**	**Justificativa**
1	X	i=0,j=0	Não calcula porque a espécie i não tem individuos
2	X	i=1,j=0	A espécie i possui indivíduos, porém a j não
3	X	i=1,j=1	Não calcula D_{ij}, porque i=j
4	X	i=1,j=2	Não calcula por j não tem indivíduos
5	D_{ij}=3-1+2 = 4, soma=0+4=4	i=1,j=3	Mesma justificativa do ID 4
6	D_{ij}=4-1+2 = 5, soma=4+5=9	i=1,j=4	Mesma justificativa do ID 2
7	X	i=2,j=0	Mesma justificativa do ID 1
8	X	i=3,j=0	Mesma justificativa do ID 2
9	D_{ij}=3-1+2 = 4, soma=9+4=13	i=3,j=1	Calcula-se de acordo com a condição 2 do cálculo da distância
10	X	i=3,j=2	Mesma justificativa do ID 2
11	X	i=3,j=3	Mesma justificativa do ID 3
12	D_{ij}=4-3+2 = 3, soma=13+3=16	i=3,j=4	Mesma justificativa do ID 9
13	X	i=4,j=0	Mesma justificativa do ID 2
14	D_{ij}=4-1+2 = 5, soma=16+5=21	i=4,j=1	Mesma justificativa do ID 9
15	X	i=4,j=2	Mesma justificativa do ID 2
16	D_{ij}=4-3+2 = 3, soma=21+3=24	i=4,j=3	Mesma justificativa do ID 9
17	X	i=4,j=4	Mesma justificativa do ID 3

4.3.1 Índice de diversidade taxonômica

Esse índice é definido na Equação 3.3. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$\Delta = \frac{(4*4*2)+(5*4*3)+(4*2*4)+(3*2*3)+(5*3*4)+(3*3*2)}{(5*2.5)} = 17,959183673469386$$

4.3.2 Índice de distinção taxonômica

Esse índice é definido na Equação 3.4. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$\Delta^* = \frac{(4*4*2)+(5*4*3)+(4*2*4)+(3*2*3)+(5*3*4)+(3*3*2)}{((4*2)+(4*3)+(2*4)(2*3)+(3*4)+(3*2)} = \frac{220}{52} = 4,230769230769231$$

4.3.3 Entropia quadrática intensiva

Esse índice é definido na Equação 3.5. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$J = \frac{24}{(3)^2} = 2,66666667$$

4.3.4 Entropia quadrática extensiva

Esse índice é definido na Equação 3.6. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$F = 24$$

4.3.5 Distinção taxonômica média

Esse índice é definido na Equação 3.7. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$AvTD = \frac{24}{[\frac{(3-1)}{2}]} = 24$$

4.3.6 Distinção taxonômica total

Esse índice é definido na Equação 3.8. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$TTD = \frac{24}{[\frac{3(3-1)}{2}]} = 8$$

4.3.7 Valor da distância de uma espécie para o seu vizinho mais próximo

Esse índice é definido na Equação 3.9 e na Equação 3.10. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$Dd = 3 + 4 + 3 = 10$$

4.3.8 Soma básica dos pesos

Esse índice é definido na Equação 3.11. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$Q = 24 * 6 = 144$$

No cálculo da Equação 3.11 há uma multiplicação por 6, porque esse número representa a quantidade de vezes que a distância foi calculada de acordo com a Tabela 3.

4.3.9 soma básica dos pesos normalizados

Esse índice é definido na Equação 3.12. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$Q = \frac{24}{3} + \frac{24}{4} + \frac{24}{3} = 22$$

4.3.10 Diversidade filogenética

Esse índice é definido nas Equações 3.13, 3.14, 3.15 e 3.16. Para exemplificar o cálculo, são utilizadas apenas as espécies i= 1 e j=3, e i= 3 e j=1 . Aplicando as equações para o primeiro calculo da distância (i= 1 e j=3), têm-se:

$$A_i = \frac{6}{4} = 1,5$$

$$C_1 = 4 * 1,5 = 6$$

$$C_2 = 6$$

Considerando o segundo cálculo da distância (i= 3 e j=1), têm-se:

$$A_i = \frac{6}{4} = 1,5$$

$$C_1 = 6 + (4 + 4) * 1,5 = 15,5$$

$$C_2 = (1,5 + 1,5)$$

$$PD_{NODE} = \frac{15,5}{3} = 5,16666667$$

4.3.11 Número de nós dentro do máximo caminho enraizado

Esse índice é definido na Equação 3.17. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$PD_{ROOT} = \frac{5 * 2,5}{20,3076923} = 0,615530303$$

4.3.12 Diversidade filogenética média

Esse índice é definido na Equação 3.18. Para exemplificar o cálculo, consideram-se as distâncias calculadas na Tabela 3, sendo assim, têm-se:

$$AvPD = \frac{5,16666667}{3} = 1,72222222$$

4.3.13 Diversidade funcional abundante

Esse índice é definido nas Equações 3.19 e 3.20. Para exemplificar o cálculo do índice FADa, foram analisados os dados da Figura 7. Podem-se observar que há 3 espécies de *voxels* (1,3 e 4), cada uma com sua quantidade de indivíduos, ou seja, espécie 1 com 4 indivíduos, espécie 3 com 2 indivíduos, e espécie 4 com 3 indivíduos. Após somar os índivíduos, foi obtido um valor de 9 indivíduos, sendo a_i a abundância individual de cada espécie e a soma da abundância de todas as espécies igual a 1, ou seja:

$$\sum_{i=1}^{i=n-1} a_i = \frac{1}{8} + \frac{3}{8} + \frac{4}{8} = 1$$

Levando em consideração as espécies 1 e 4 do dendograma, as distâncias entre elas são de $D_{ij} = 2$, pois essas duas espécies correspondem à primeira e à terceira posição no dendograma da Figura 7, respectivamente. Assim, têm-se:

$$FADa = 2 * \frac{1}{8} * \frac{4}{8} = 0,125$$

4.3.14 Diversidade funcional abundante das espécies

Esse índice é definido nas Equações 3.22 e 3.23. Levando em consideração as espécies 1 e 4 do dendograma, observam-se que as distâncias entre elas são de $D_{ij} =$ 3,5, pois diferente do cálculo do FADa que utiliza os valores das posições dos *voxels* no

dendograma, esse índice utiliza os valores das intensidades dos *voxels* nos cálculos das distâncias. Portanto, aplicando na Equação 3.22, têm-se:

$$FADe = 3,5 * \frac{4}{9} * \frac{3}{9} = 0,518518519$$

4.3.15 Diversidade funcional abundante dos pixels

Esse índice é definido nas Equações 3.24 e 3.25. Para exemplificar a Equação 3.25, utilizam-se as espécies do dendograma ilustrado na Figura 7. Notam-se que há 3 espécies de *voxels* (1,3 e 4), cada um com seus indivíduos (espécie 1 com 4 indivíduos, a espécie 3 com 2 indivíduos e a espécie 4 com 3 indivíduos). A abundância individual deve ser calculada multiplicando o valor da espécie, pela sua quantidade de indivíduos, ou seja, para a espécie 1 têm-se 1 * 4 = 4, para a espécie 3 têm-se 3 * 2 = 6 e para a espécie 4 têm-se 4 * 3 = 12. Assim, a soma da abundância de espécie dessa comunidade é igual a 22, ou seja:

$$\sum_{i=1}^{i=n-1} ap_i = \frac{4}{22} + \frac{6}{22} + \frac{12}{22} = 1$$

Feito isso, o índice FADe é calculado utilizando a distância D_{ij} e os pares de espécies com suas abundânci as, ou seja:

$$FADp = 3.5 * \frac{4}{22} * \frac{12}{22} = 0.347107438.$$

4.3.16 Diversidade funcional

Esse índice é descrito na Subseção 3.6.1.2.4. Levando em consideração essa subseção, a quantidade de braços do dendongrama ilustrado na Figura 7 é igual a 3.

4.3.17 Diversidade funcional abundante relacionando valores dos voxels e a distância entre as espécies

Esse índice é definido na **??**. Para exemplificar o cálculo do FDaPe foram analisados os mesmos exemplos dos dados da Figura 7. A distância D_{ij} é calculada utilizando a Equação 3.23. Já ape_j é calculado combinando cada intensidade com a abundância de indivíduos utilizada no cálculo do índice FADp, ou seja:

$$\sum_{i=1}^{n-1} ape_i = \frac{1}{22} + \frac{3}{22} + \frac{4}{22}$$

Considerando apenas as espécies 1 e 4 do dendograma, pode-se calcular a distância $D_{ij} = 3{,}5$. Fazendo isso, o índice FDaPe é exemplificado como:

$$FDaPe = 3,5 * \frac{1}{22} * \frac{4}{22} = 0,0289256198$$

4.3.18 Entropia quadrática funcional

Esse índice é definido na Equação 3.28. Para exemplificar o cálculo do índice, consideram-se as espécies 1 e 4 do dendograma. Utilizando a Equação 3.19, obteve-se um valor do índice $FADa = 0{,}125$. Segundo a Figura 7, têm-se 3 espécies (1,3 e 4). Aplicando isso na Equação 3.28:

$$EQF = \frac{0.125}{(3)^2} = \frac{0.125}{9} = 0,0138888889$$

4.3.19 Diversidade funcional abundante considerando espécies e voxels

Esse índice é definido na Equação 3.29. Para exemplificar o cálculo do índice, consideram-se as espécies 1 e 4 do dendograma. Utilizando a Equação 3.22, obteve-se um valor do índice $FADe = 0{,}518518519$. Segundo a Figura 7, têm-se 3 espécies (1,3 e 4). Aplicando isso na Equação 3.28:

$$FADs = \frac{0.518518519}{(3)^2} = \frac{0,518518519}{9} = 0,0576131688$$

4.3.20 Soma das distâncias Funcionais

Esse índice é definido na Equação 3.30. Para exemplificar o cálculo do índice, devem-se considerar as espécies do dendograma (1, 3 e 4). Utilizando a Equação 3.20 para calcular a distância de cada uma das espécies para as demais, têm-se um valor de: $D_{13} = 1$; $D_{14} = 2$; $D_{31} = 1$; $D_{34} = 1$; $D_{41} = 2$; e $D_{34} = 1$. Assim, o valor de $SDF = 8$.

4.3.21 Soma das intensidades funcionais

Esse índice é definido na Equação 3.31. Para exemplificar o cálculo do índice, devem-se considerar as espécies do dendograma (1, 3 e 4). Utilizando a Equação 3.20 para

calcular a distância de cada uma das espécies para as demais, têm-se um valor de: $D_{13} = 2$; $D_{14} = 3$; de $D_{31} = 2$; $D_{34} = 1$; $D_{41} = 3$; e $D_{34} = 1$. Assim, o valor de $SIF = 12$.

4.4 Seleção de Características e Classificação

Esta etapa é muito utilizada no contexto de treinamento de algoritmos em processamento de imagens, com a utilização do algoritmo *Greedy Stepwize* para selecionar as melhores características dos indivíduos (GIGER et al., 2000). Essas características selecionadas foram submetidas aos classificadores para realizarem o processo de disciplinação das classes.

Diversas experimentos foram feitos, são eles: com os índices de diversidade filogenética (12 características); com os índices de diversidade funcional (9 características); com índices de diversidade filogenética sob a imagem original (12 *bits*) e as quantizações (11, 10 e 9 *bits*) totalizando 48 características; com índices de diversidade funcional sob a imagem original (12 *bits*) e as quantizações (11, 10 e 9 *bits*) totalizando 36 características; depois extrairam-se as características dos índices de diversidade filogenética e funcional sob a imagem original (12 *bits*) e as quantizações (11, 10 e 9 *bits*) totalizando 84 características; por fim, selecionaram-se 8 características (7 características filogenéticas e 1 funcional da imagem quantizada) utilizando o algoritmo *Greedy Stepwise*. Cada experimento desse foram submetidos aos classifcadores.

Os 2 classificadores utilizados para o trabalho foram o *Random Forest*, por ser um classificador utilizado na maioria dos trabalhos relacionados e por ser melhor para classes desbalanceadas (LEE; KOUZANI; HU, 2010), e o *Support Vector Machine* (SVM)(CHANG; LIN, 2011), utilizado para provar a eficiência das características, devido não ser considerado eficaz para multiclasses com os parâmetros estimados pelo algoritmo *grid search* (**??**) e mostrar que mesmo em classificadores não apropriados para bases desbalanceadas, as características conseguem discriminar as classes. Foram testados nas proporções treino/teste da base total de 20%/80%,40%/60%,60%/40% e 80%/20%, respectivamente.

@articlebecsey1968nonlinear, title=Nonlinear least squares methods: A direct grid search approach, author=Becsey, JC and Berke, Laszlo and Callan, James R, journal=Journal of Chemical Education, volume=45, number=11, pages=728, year=1968, publisher=ACS Publications

4.5 Validação

Após a etapa de classificação dos índivíduos em carcinoma de célula grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa ou não especificados, faz-se necessário a validação dos resultados. Comumente, os sistemas CADx utilizam métricas aceitas pela literatura para analisar a eficiência de sistemas que se baseiam em processamento de imagens e reconhecimento de padrões. Para este trabalho, utilizaram-se o índice Kappa, a acurácia e a área sob a curva ROC (Seção 3.9).

A análise das métricas têm diversas finalidades subjetivas, como identificar onde o modelo está errando, onde melhorar o modelo, entre outras, sendo que o principal objetivo é medir o desempenho do modelo proposto como promissor ou não, além de fornecer auxílio na identificação de aspectos positivos e negativos para futuros trabalhos, tanto nas fases de treinamentos, como nas fases de testes.

5 RESULTADOS E DISCUSSÃO

Para validar o método proposto foram realizados testes de eficiência com a base NSCLC-Radiomics utilizando 319 dos 422 exames, devido 103 dos exames disponíveis na base estarem sem marcações. A classe carcinoma de células grandes é representada por 101 exames, 92 exames são classificados como carcinoma de células escamosas, 31 exames são classificados como adenocarcinoma, 56 exames são classificados como adenocarcinoma de mutação negativa e 39 exames não tiveram as classes especificadas (classe neste trabalho chamada de "não especificados").

Nas seções que seguem, serão apresentados os resultados e é discutida a utilização das caracteristicas e dos classificadores utilizados na classificação automática de nódulos de NSCLC em carcinoma de célula grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados, nas proporções para treino/teste de 20/40, 40/60, 60/40 e 80/20, respectivamente, para os descritores de diversidade filogenética e funcional. Por fim, será comparado o melhor resultado com a literatura. Os resultados aqui apresentados são provenientes de uma média de 5 execuções para cada experimento nos classificadores.

5.1 Resultados para os Descritores de Diversidade Filogenética

Os resultados aqui apresentados referem-se aos descritores de diversidade filogenética descritos na Subseção 3.6.1.1. As métricas de desvio padrão da curva ROC, índice Kappa e acurácia, são representados nas tabelas por DPR, DPK e DPA, respectivamente.

Analisando os resultados na Tabela 4 observam-se resultados promissores quando comparados com a literatura e seus respectivos desvios-padrões utilizando o classificador *Random Forest*, com uma AUC ROC de 0,991, para uma proporção 80/20. Vale salientar que a diferença de valores entre as proporções 20/80 (AUC ROC de 0,970) e a 80/20 é pequena, o que mostra também que para o classificador *Random Forest* as características são eficazes tanto em um conjunto pequeno de dados de treino, como em um conjunto grande. Os devios-padrões foram baixos, mostrando concordância e pouca variação independentemente dos indivíduos utilizados no treino/teste.

Tabela 4 – Resultados para a classificação com análise de textura baseada em diversidade filogenetica para o classificador *Random Forest*.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,997	ROC: 0,997	ROC: 0,999	ROC: 0,999
Kappa: 0,932	Kappa: 0,917	Kappa: 0,960	Kappa: 0,981
Acurácia: 94,86%	Acurácia: 93,68%	Acurácia: 96,98%	Acurácia: 98,80%
DPR: 0,00152	DPR: 0,00114	DPR: 0,00055	DPR: 0,00045
DPK: 0,02156	DPK: 0,02536	DPK: 0,02817	DPK: 0,01834
DPA: 1,644	DPA: 1,9481	DPA: 2,20231	DPA: 1,41972

Analisando os resultados na Tabela 5 observam-se resultados aceitáveis quando comparados com a literatura e seus respectivos desvios-padrões, porém um pouco mais baixo que os classificadores da Tabela 4, com uma AUC ROC de 0,961 e uma acurácia de 93,96% para uma proporção 80/20. Vale salientar que a diferença de valores entre as proporções 20/80 (AUC ROC de 0,956 e uma acurácia de 92,96%) e a 80/20 é pequena, mostrando mais uma vez que as características são eficazes tanto em um conjunto pequeno de dados de treino, como em um conjunto grande, para o classificador SVM. O melhor resultado foi obtido pela proporção 40/60 com uma AUC ROC de 0,964 e uma acurácia de 94,31%. Acredita-se que o resultado inferior aos demais classificadores, é devido o classificador SVM ser projetado para classificações binárias, sendo feito uma adaptação para classificação multiclasses (CASTRO et al., 2007).

Tabela 5 – Resultados para a classificação com análise de textura baseada em diversidade filogenetica para o classificador SVM.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,956	ROC: 0,964	ROC: 0,958	ROC: 0,961
Kappa: 0,907	Kappa: 0,925	Kappa: 0,917	Kappa: 0,922
Acurácia: 92,96%	Acurácia: 94,31%	Acurácia: 93,65%	Acurácia: 93,96%
DPR: 0,0165	DPR: 0,01106	DPR: 0,01394	DPR: 0.03071
DPK: 0,03507	DPK: 0,2727	DPK: 0,02755	DPK: 0,06029
DPA: 2,70109	DPA: 2,15084	DPA: 2,17336	DPA: 4,68186

Com o intuito de criar uma nova representação de diversidade, para uma possível visualização do comportamento dos índices de diversidade filogenética e funcional, as imagens originais (que são de 12 *bits*) foram quantizadas para 11, 10 e 9 *bits*, pois quantiazações com valores de *bits* menores que esses causariam perda de informações.

Observando os resultados da Tabela 6 verificam-se que não foram melhores que os resultados utilizando somente as imagens originais, obtendo uma AUC ROC de 0,999 e uma acurácia de 98,79% para uma proporção 80/20. Vale salientar que a diferença de valores entre as proporções 20/80 (AUC ROC de 0,996 e uma acurácia de 95,90%) e a

80/20 é pequena, mostrando mais uma vez que as características são eficazes tanto em um conjunto pequeno de dados de treino, como em um conjunto grande para o classificador *Random Forest*.

Tabela 6 – Resultados para a classificação com análise de textura baseada em diversidade filogenetica para o classificador *Random Forest* com Quantizações 11, 10 e 9 bits.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,996	ROC: 0,999	ROC: 0,999	ROC: 0,999
Kappa: 0,946	Kappa: 0,978	Kappa: 0,977	Kappa: 0,983
Acurácia: 95,90%	Acurácia: 98,32%	Acurácia: 98,26%	Acurácia: 98,79%
DPR: 0,00451	DPR: 0,00045	DPR: 0,00045	DPR: 0,00045
DPK: 0,02581	DPK: 0,01225	DPK: 0,01126	DPK: 0,00903
DPA: 1,98028	DPA: 0,93657	DPA: 0,86257	DPA: 0.69877

Na Tabela 7, os resultados também são melhores que os resultados somente com as imagens originais, obtendo uma AUC ROC de 0,999 e uma acurácia de 97,81%, para uma proporção 80/20. Observa-se uma diferença considerável nos valores entre as proporções 20/80 (AUC ROC de 0,996 e uma acurácia de 91,65%) e a 80/20, mostrando que o classificador SVM necessita de uma quantidade maior de dados para treino, porém os resultados ainda são eficazes tanto em um conjunto pequeno de dados de treino, como em um conjunto grande para o classificador SVM.

Tabela 7 – Resultados para a classificação com análise de textura baseada em diversidade filogenetica para o classificador SVM com Quantizações 11, 10 e 9 bits.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,945	ROC: 0,973	ROC: 0,974	ROC: 0,986
Kappa: 0,890	Kappa: 0,943	Kappa: 0,950	Kappa: 0,971
Acurácia: 91,65%	Acurácia: 95,70%	Acurácia: 96,22%	Acurácia: 97,81%
DPR: 0,00966	DPR: 0,0041	DPR: 0,00823	DPR: 0,01354
DPK: 0,01818	DPK: 0,0078	DPK: 0,01353	DPK: 0,0281
DPA: 1,37513	DPA: 0,57352	DPA: 1,02664	DPA: 2,09631

Podem-se observar os resultados e a eficiência das características para os índices de diversidade filogenéticos nas Tabelas 4, 5, 6 e 7, mostrando o poder discriminativo das características tanto nas imagens originais, como nas imagens quantizadas. Para o classificador SVM, os resultados foram melhores quando extraíram-se características tanto das imagens originais como das imagens quantizadas, supondo-se que os índices de diversidade funcional necessitam um pouco mais de diversidade para uma melhor discriminação. Para o classificador *Random Forest* os resultados foram melhores utilizando apenas a extração sob as imagens originais se utilizada a métrica acurácia para comparação, com aumento em todas as demais métricas e redução dos desvios-padrões. Acredita-se

que o *Random Forest* obteve melhores resultados, também devido trabalhar melhor com bases desbalanceadas (LENTO, 2017).

A Tabela 8 mostra dois melhores resultados entre as classificações utilizando somente os índices de diversidade filogenética, indicando que os dois melhores resultados foram feitos pelo classificador *Random Forest*.

Tabela 8 – Melhores resultados para os classificadores utilizando somente índices de diversidade filogenética.

Random Forest	***Random Forest* COM QUANTIZAÇÃO**
Proporção: 80/20	**Proporção: 80/20**
ROC: 0,999	ROC: 0,999
Kappa: 0,980	Kappa: 0,983
Acurácia: 98,81%	Acurácia: 98,79%
DPR: 0,00045	DPR: 0,00045
DPK: 0,01834	DPK: 0,00903
DPA: 1,41972	DPA: 0,69877

5.2 Resultados para os Descritores de Diversidade Funcional

Os resultados aqui apresentados referem-se aos descritores de diversidade funcional descritos na Subseção 3.6.1.2. As métricas de desvio padrão da AUC ROC, do índice Kappa e da acurácia, são representados nas tabelas por DPR, DPK e DPA, respectivamente.

Analisando os resultados na Tabela 9 pode-se observar que o classificador *Random Forest*, também não obteve um resultado expressivo, alcançando uma melhor AUC ROC de 0,905 e uma acurácia de 73,75% para uma proporção 80/20. É notória a diferença de valores entre as proporções 20/80 (AUC ROC de 0,626 e uma acurácia de 36,86%) e a 80/20, o que mostra que as características necessitam de um treino maior.

Tabela 9 – Resultados para a classificação com análise de textura baseada em diversidade funcional para o classificador *Random Forest*.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,626	ROC: 0,729	ROC: 0,805	ROC: 0,905
Kappa: 0,149	Kappa: 0,291	Kappa: 0,435	Kappa: 0,649
Acurácia: 36,86%	Acurácia: 46,80%	Acurácia: 57,81%	Acurácia: 73,75%
DPR: 0,01971	DPR: 0,03573	DPR: 0,04027	DPR: 0.05556
DPK: 0,02418	DPK: 0,03278	DPK: 0,05355	DPK: 0,01225
DPA: 3,36925	DPA: 2,47237	DPA: 4,24329	DPA: 8,07058

Na Tabela 10 pode-se observar, assim como no classificador SVM, um resultado ainda não muito eficiente, com uma AUC ROC de 0,777 e uma acurácia de 71,25% para

uma proporção 80/20. Comparando as proporções 20/80 e 80/20, nota-se uma grande diferença em seus resultados, mostrando novamente que as características necessitam de um treino com maior proporção.

Tabela 10 – Resultados para a classificação com análise de textura baseada em diversidade funcional para o classificador SVM.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,575	ROC: 0,675	ROC: 0,734	ROC: 0,777
Kappa: 0,149	Kappa: 0,354	Kappa: 0,472	Kappa: 0,620
Acurácia: 36,70%	Acurácia: 51,93%	Acurácia: 60,46%	Acurácia: 71,25%
DPR: 0,0457	DPR: 0,01967	DPR: 0,03546	DPR: 0,07907
DPK: 0,09199	DPK: 0,04016	DPK: 0,07244	DPK: 0,10777
DPA: 6,07274	DPA: 3,0167	DPA: 5,8097	DPA: 7,08886

Quando os índices de validação nos classificadores são analisados, observa-se que o *Random Forest*, foi o melhor classificador. Acredita-se que esse teve melhor desempenho, por ser melhor com bases desbalanceadas (LENTO, 2017), características da base de imagens NSCLC-Radiomics, utilizada neste trabalho. O SVM teve um desempenho menor que o classificador *Random Forest*, acreditando-se que diferente dos dois classificadores que utilizam árvores, o SVM utiliza um hiperplano criado (inicialmente) para separar duas classes, aumentando ainda mais o grau de dificuldade desse trabalho, já que a base de dados utilizada para testes das características possui cinco classes. Além disso, a discriminação das classes torna-se mais difícil devido a semelhança na textura, como pode ser observada na Figura 8.

Figura 8 – Nódulos de texturas semelhantes.

utilizaram-se as quantizações 11,10 e 9 *bits* para ter uma outras representações, com o intuito de aumentar a diversidade de indivíduos e possivelmente a quantidade de espécies com o surgimento de novas, e para analisar as diferenças no comportamento dos índices de diversidade funcional.

Na Tabela 11 pode-se observar um aumento de quase 10% na acurácia em relação a extração somente com os índices de diversidade funcional, obtendo uma AUC ROC de 0,992 e uma acurácia de 80,37%, para uma proporção 80/20, porém, melhor que os resultados mostrados com os índices de diversidade funcional somente sobre as imagens

originais. Nota-se que há diferença de valores entre as proporções 20/80 (AUC de 0,675 e uma acurácia de 40%) e a 80/20, porém obteve uma melhora nos resultados dessas proporções, sugerindo resultados promissores.

Tabela 11 – Resultados para a classificação com análise de textura baseada em diversidade funcional para o classificador *Random Forest*.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,675	ROC: 0,792	ROC: 0,881	ROC: 0,992
Kappa: 0,184	Kappa: 0,344	Kappa: 0,527	Kappa: 0,733
Acurácia: 40%	Acurácia: 52,14%	Acurácia: 65,15%	Acurácia: 80,37%
DPR: 0,03483	DPR: 0,01442	DPR: 0,0319	DPR: 0,00843
DPK: 0,03138	DPK: 0,0599	DPK: 0,05807	DPK: 0,10235
DPA: 1,7556	DPA: 4,51296	DPA: 4,4401	DPA: 4,66585

Na Tabela 12 pode-se observar, assim como no classificador SVM, que aumentando a quantidade de imagens (com diferentes representações), obtiveram-se pequenos, mas consideráveis, aumentos nas métricas em relação à extração utilizando os índices de diversidade funcional somente sobre as imagens originais, conseguindo uma AUC ROC de 0,822 e uma acurácia de 73,12%, para uma proporção 80/20, Notam-se diferenças de valores entre as proporções 20/80 (AUC ROC de 0,593 e uma acurácia de 41,56%) e a 80/20, porém obteve uma melhora nos resultados dessas proporções.

Tabela 12 – Resultados para a classificação com análise de textura baseada em diversidade funcional para o classificador SVM.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,593	ROC: 0,684	ROC: 0,740	ROC: 0,822
Kappa: 0,19	Kappa: 0,371	Kappa: 0,470	Kappa: 0,642
Acurácia: 41,56%	Acurácia: 53,08%	Acurácia: 57,33%	Acurácia: 73,12%
DPR: 0,01281	DPR: 0,00924	DPR: 0,00751	DPR: 0,04242
DPK: 0,02666	DPK: 0,02044	DPK: 0,00387	DPK: 0,64232
DPA: 2,14795	DPA: 1,72056	DPA: 0,28868	DPA: 5,09175

A Tabela 13 mostra os melhores resultados entre as classificações utilizando somente os índices de diversidade funcional, indicando que os dois melhores resultados foram obtidos pelo classificador *Random Forest*.

Tabela 13 – Os dois melhores resultados para os classificadores utilizando somente índices de diversidade funcional.

Random Forest	***Random Forest* COM QUANTIZAÇÃO**
Proporção: 80/20	**Proporção: 80/20**
ROC: 0,905	ROC: 0,992
Kappa: 0,649	Kappa: 0,733
Acurácia: 73,75%	Acurácia: 80,37%
DPR: 0,05556	DPR: 0,00843
DPK: 0,01225	DPK: 0,10235
DPA: 8,07058	DPA: 4,66585

5.3 Resultados para a União dos Descritores de Diversidade Filogenética e Funcional

Os resultados aqui apresentados referem-se aos descritores de diversidade filogenética e funcional descritos na Subseção 3.6.1. As métricas de desvio padrão da curva ROC, índice Kappa e AUC ROC são representados nas tabelas por DPR,DPK e DPA, respectivamente.

Com o intuito de obter um diagnóstico mais eficiente, os índices de diversidade filogenética e funcional foram testados juntos, acreditando que poderiam fornecer uma discriminação mais eficaz. Além disso, com uma maior quantidade de representações se mostrou mais eficiente, novamente utilizou-se as quantizações 11, 10 e 9 bits.

Analisando a Tabela 14 pode-se observar que após as imagens serem quantizadas, as métricas de validação aumentaram para o classificador *Random Forest* com a união dos índices de diversidade filogenética e funcional, sugerindo melhor discriminação das classes, quando comparada com a extração individual de características dos índices de diversidade filogenética, ou com a extração individual de características dos índices de diversidade funcional. O melhor resultado para o classificador *Random Forest* dentre as proporções após a união dos índices foi uma AUC ROC de 0,999 e uma acurácia de 98,12% para uma proporção 80/20. Nota-se que ainda existe uma diferença de valores entre as proporções 20/80 (uma AUC ROC de 0,999 e uma acurácia de 90,74%) e a 80/20, porém essa diferença foi reduzida de forma significativa, quando comparada às extrações de características dos grupos filogenéticos de forma separada.

Tabela 14 – Resultados para a classificação com análise de textura baseada em diversidade filogenética e funcional para o classificador *Random Forest*.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,999	ROC: 0,999	ROC: 0,999	ROC: 0,999
Kappa: 0,944	Kappa: 0,979	Kappa: 0,975	Kappa: 0,975
Acurácia: 95,76%	Acurácia: 98,424%	Acurácia: 98,12%	Acurácia: 98,12%
DPR: 0,00055	DPR: 0,00045	DPR: 0,00045	DPR: 0,00045
DPK: 0,05114	DPK: 0,00491	DPK: 0,01186	DPK: 0,0172
DPA: 3,91569	DPA: 0,37021	DPA: 0,89077	DPA: 1,30728

Analisando a Tabela 15 observam-se resultados satisfatórios para o classificador SVM, porém houve uma redução nas métricas quando comparadas as métricas dos experimentos utilizando somente os índices de diversidade filogenética. Porém, foi melhor do que a extração de características separada dos índices de diversidade funcional. Como melhor resultado para esse classificador, obtiveram-se uma AUC ROC de 0,964 e uma acurácia de 94,06% para uma proporção 80/20.

Tabela 15 – Resultados para a classificação com análise de textura baseada em diversidade filogenética e funcional para o classificador SVM.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,898	ROC: 0,939	ROC: 0,948	ROC: 0,964
Kappa: 0,785	Kappa: 0,871	Kappa: 0,880	Kappa: 0,922
Acurácia: 83,68%	Acurácia: 90,26%	Acurácia: 91,71%	Acurácia: 94,06%
DPR: 0,03054	DPR: 0,01291	DPR: 0,01639	DPR: 0,02257
DPK: 0,06224	DPK: 0,02922	DPK: 0,03292	DPK: 0,05284
DPA: 4,71076	DPA: 2,23975	DPA: 2,62623	DPA: 4,04443

Analisando os resultados da união dos índices de diversidade filogenética e funcional, observou-se que para os experimentos utilizando o *Random Forest*, obteve-se uma melhora considerável, mostrando que um tipo de índice, complementou o outro, buscando propriedades que somente um tipo, não conseguia analisar sozinho. Já para o SVM, não obteve melhores resultados, quando comparados a experimentos utilizando os índices de diversidade filogenética individualmente.

Com o intuito de melhorar a eficiência da metodologia, foi utilizado o seletor de características *Greedy Stepwise*, para selecionar as características mais relevantes. Essas características selecionadas foram os índices de diversidade taxonômica, de distinção taxonômica, de entropia quadrática intensiva, de entropia quadrática extensiva, o *AvTD*, o *TTD* e o *Dd*, sendo estes características das imagens originais dos índices de diversidade filogenética. O índice funcional *FADa* foi o único selecionado das características quantizadas representando os grupo de índices de diversidade funcional.

Analisando a Tabela 16 pode-se observar que após a união dos dois tipos de índices de diversidade e com a seleção de características, o classificador *Random Forest* melhorou seus resultados de forma significativa comparados aos resultados sem a seleção de características. Obtiveram-se aumentos nas métricas em relação a todos os testes sem a seleção de característica, alcançando uma AUC ROC de 0,999 e uma acurácia de 99,44%, para uma proporção 80/20. Nota-se que há apenas uma pequena diferença de valores entre as proporções 20/80 (AUC ROC de 0,997 e uma acurácia de 95,68%) e a 80/20, mostrando que as características são eficientes também para poucos dados.

Tabela 16 – Resultados para a classificação com análise de textura baseada em diversidade filogenética e funcional para o classificador *Random Forest* com a seleção pelo *Greedy Stepwise*.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,997	ROC: 0,999	ROC: 0,999	ROC: 0,999
Kappa: 0,942	Kappa: 0,969	Kappa: 0,977	Kappa: 0,990
Acurácia: 95,68%	Acurácia: 97,69%	Acurácia: 98,28%	Acurácia: 99,44%
DPR: 0,00096	DPR: 0,00045	DPR: 0,00055	DPR: 0,00045
DPK: 0,02919	DPK: 0,01158	DPK: 0,00864	DPK: 0,01213
DPA: 2,19517	DPA: 0,87607	DPA: 0,65363	DPA: 0,6436

De acordo com a Tabela 17 podem-se observar melhoras nos resultados do classificador SVM, garantindo que união dos dois índices com a seleção de características também foi eficiente para classificadores que originalmente foram feitos para problemas de classes binárias. Obtiveram-se aumentos nas métricas em relação a todos os testes anteriores feitos sem a seleção de características utilizando o SVM, alcançando uma AUC ROC de 0,985 e uma acurácia de 97,5%, para uma proporção 80/20. Nota-se que há diferença de valores entre as proporções 20/80 (AUC ROC de 0,917 e uma acurácia de 87,52%) e a 80/20, mostrando que o classificador SVM necessita de uma quantidade maior de indivíduos para treinamentos.

Tabela 17 – Resultados para a classificação com análise de textura baseada em diversidade filogenética e funcional para o classificador SVM com a seleção pelo *Greedy Stepwise*.

Proporção: 20/80	**Proporção: 40/60**	**Proporção: 60/40**	**Proporção: 80/20**
ROC: 0,917	ROC: 0,943	ROC: 0,973	ROC: 0,985
Kappa: 0,836	Kappa: 0,892	Kappa: 0,952	Kappa: 0,967
Acurácia: 87,52%	Acurácia: 91,83%	Acurácia: 96,00%	Acurácia: 97,5%
DPR: 0,02973	DPR: 0,02131	DPR: 0,00611	DPR: 0,00817
DPK: 0,05892	DPK: 0,04209	DPK: 0,01157	DPK: 0,01842
DPA: 4,46612	DPA: 3,21894	DPA: 0,89077	DPA: 1,39754

A Tabela 18 mostra dois melhores resultados entre as classificações utilizando os índices de diversidade filogenética e funcional, indicando que os dois melhores resultados

foram obtidos pelo classificador *Random Forest*.

Tabela 18 – Os dois melhores resultados para os classificadores utilizando os índices de diversidade filogenética e funcional.

***Random Forest* COM QUANTIZAÇÃO**	***Random Forest* COM SELEÇÃO**
Proporção: 80/20	**Proporção: 80/20**
ROC: 0,999	ROC: 0,999
Kappa: 0,975	Kappa: 0,990
Acurácia: 98,79%	Acurácia: 99,44%
DPR: 0,00045	DPR: 0,00045
DPK: 0,0172	DPK: 0,01213
DPA: 1,30728	DPA: 0,6436

Visto os resultados apresentados após a seleção de características, é notório que os índices se complementam, cobrindo propriedades específicas de cada indivíduo. Além disso, através das imagens foram extraídas diversas características, e que apenas com as características extraídas das imagens, é possível prover um diagnóstico, comprovando a teoria do Radiomics. Vale salientar, que além da extração de uma grande quantidade de dados, é possível também selecionar as características mais relevantes, mostrando que não somente é possivel extrair uma grande quantidade de características, mas também com grande qualidade.

5.4 Discussão

Partindo dos resultados apresentados nas Seções 5.1, 5.2 e 5.3, a metodologia proposta para classificação de NSCLC baseada em análise de textura, demonstram-se bastante eficazes. As análises de textura através dos índices de diversidade filogenética combinados com os índices de diversidade funcional conseguiram resultados expressivos em todas as análises com os diferentes grupos de índices.

Dentre todos os testes, o melhor resultado foi a partir da combinação dos dois grupos de índices sobre as imagens originais e quantizadas, logo após ser feita uma seleção de características. Acredita-se que isso ocorreu, devido a avaliação da diversidade presente na comunidade analisar precisamente a relação existente entre as espécies. Além disso, os índices conseguiram utilizar a importância de cada espécie em diferentes representações, agregando a quantidade de informações a respeito dos índividuos por meio da análise da diversidade.

Acreditando no potencial dos descritores, é necessário descrever sobre o experimento realizado com todos os índices em conjunto. A premissa básica para a realização dos testes desses índices se dá pelo fato de que um índice possa identificar alguma propriedade da textura do nódulo que outro não consiga. Pode-se observar isso, a partir da Tabela 18, onde a combinação dos índices tornaram-se mais eficientes, isto é, houve uma melhora nos resultados.

O melhor resultado alcançado pelos testes da Tabela 16 foi um valor expressivo de 0,999 de AUC ROC, um índice Kappa de 0,990 e 99,44% de acurácia, com desvios-padrões baixos quando comparados com a literatura para a classificação de nódulos de NSCLC, sugerindo uma concordância entre as classificações, independentemente da mudança de semente inicial a cada execução demonstrando a efetividade do método proposto, e mostrando a eficiência em conjunto da combinação das características mais relevantes dos índices de diversidade filogenética e funcional em diferentes representações.

Acredita-se que a melhor eficiência do classificador *Random Forest* em relação ao SVM se dá por estes serem *ensembles* de árvores que utilizam limiares, ou seja, irão combinar as melhores saídas dentre várias outras saídas, baseadas em uma decisão utilizando diferentes limiares que dividem os grupos de características, e os melhores grupos formados por aqueles *clusters* de valores de características, irão se referir a aquela classe.

Na maioria dos testes apresenados, o classificador SVM obteve o menor resultado (porém satisfatório) na classificação dos nódulos de NSCLC. Acredita-se que isso se deve ao fato de que esse classificador foi inicialmente desenvolvido para classes binárias, sendo feitos arranjos técnicos para uma discriminação multiclasse.

Além disso a base NSCLC-Radiomics é uma base complexa e de grande diversidade de classes, sendo uma base que conta apenas com exames convencionais de TC, ou seja, não usam elementos químicos (como sulfato de bário, iodo, gadolínio, entre outros) para elevação do contraste das estruturas para facilitar a discriminação de classes, isto é importante para mostrar a eficiência dos métodos propostos, pois alcançam bons resultados sem a exposição a uma maior radiação, o que causaria risco a saúde do paciente e do especialista.

Por fim, observa-se na Figura 9, um gráfico de análise Radiomics com as características de textura (índices de diversidade filogenética e funcional) sobre as representações (12, 11, 10 e 9 bits) dos nódulos pulmonares. Os valores dos descritores foram normalizados

utilizando o *Z-score*. As lesões pulmonares dos pacientes estão distribuidos sobre o eixo x (lesões adquiridas da base NSCLC-Radiomics). Os descritores extraídos das lesões estão distribuídos sobre o eixo y (descritores filogenéticos em rosa, e descritores funcionais em laranja). Analisando o mapa Radiomics, verificam-se pequenos *clusters* criados com os indivíduos de cada classe, separados pelas linhas de cor vermelha, sugerindo que as características foram efetivas na caracterização das modificações das diversidades intratumorais, representando assim, os genótipos dos tecidos de maneira equivalente às análises clínicas. Além disso, observam-se que algumas classes são muito semelhantes, como as classes 3 e 4, o que dificultaria a análise visual por especialistas.

A intenção de ilustrar todas as características extraídas através da Figura 9 é de mostrar graficamente que mesmo sem uma seleção de características, os descritores utilizados nesse trabalho conseguem caracterizar os indivíduos, comprovando graficamente que há características que os correlacionam. A seleção de característica foi utilizada, para mostrar que dentre todo um conjunto de características que podem ser extraídas das imagens digitais, existem características que melhor correlacionam os pacientes com seu tipo de lesão, sugerindo que através das imagens digitais, não somente é possivel extrair uma grande quantidade de características, como também pode existir qualidade de características.

5.5 Comparação com os Trabalhos que Utilizam uma Abordagem Radiomics

Utilizam-se as métricas de validação dos trabalhos relacionados para uma comparação mais fidedígna, devido classificarem em 2 ou 3 classes dentre as cinco classes de nódulos de NSCLC da metodologia proposta.

Na Tabela 19 têm-se uma visão resumida (área sobre a curva ROC, número de características, base de imagens utilizada e número de casos na base) dos resultados encontrados nos trabalhos relacionados e na metodologia proposta. Assim, pretende-se mostrar que a metodologia proposta é promissora, uma vez que, em comparação com outras obras, alcançou os resultados da literatura, que estão em torno de uma AUC ROC de 0,600 e 0,700 para a tarefa de classificação de nódulos de NSCLC.

Observando a Tabela 19, podem-se analisar que os trabalhos são recentes no contexto da abordagem Radiomics e que o método proposto obteve melhores resultados nos índices de validação (acurácia - ACC, AUC ROC e índice *Kappa*), mesmo quando com-

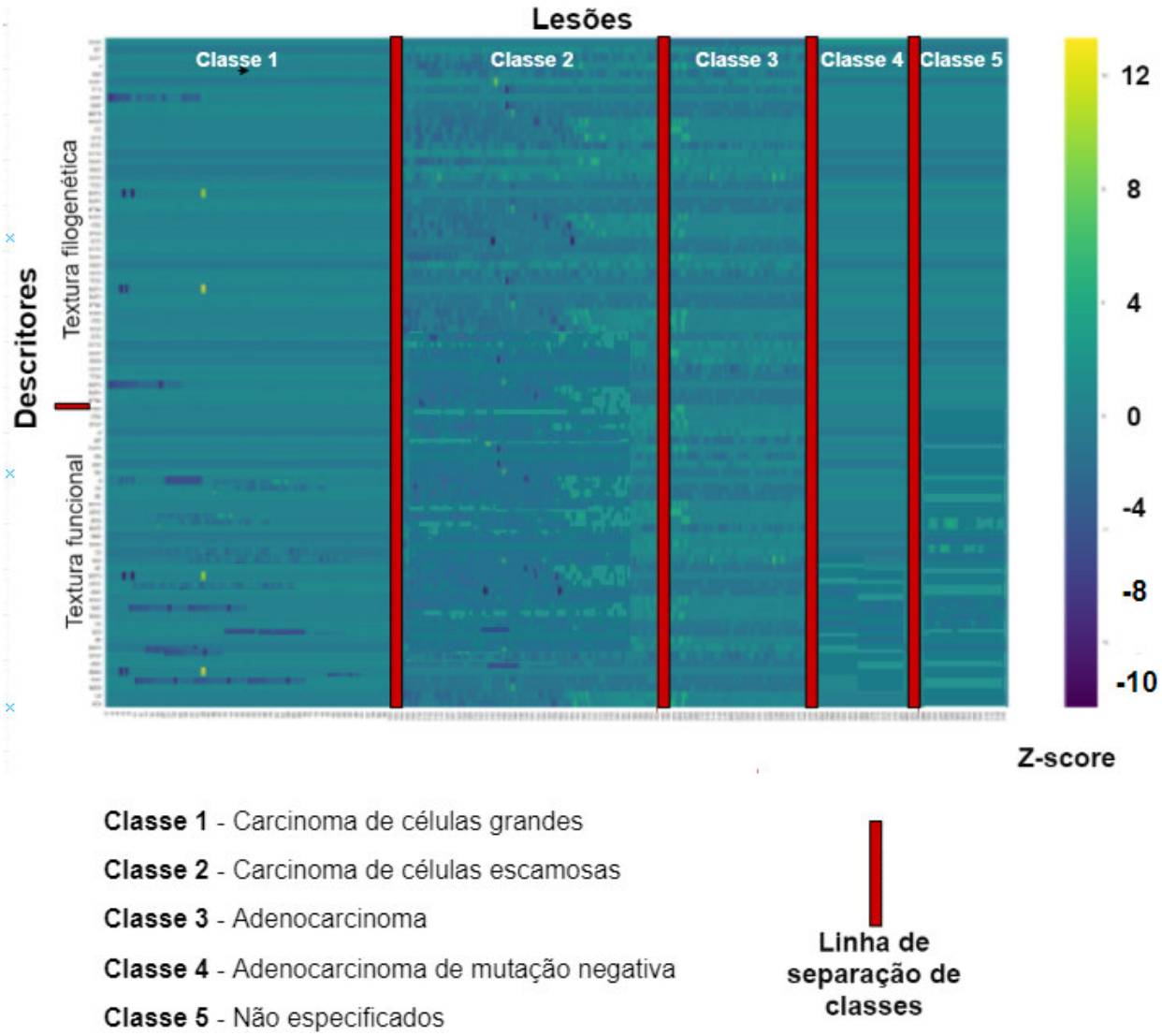

Figura 9 – Gráfico Radiomics de correlação entre as características, baseadas nos indivíduos.

parado com os piores resultados deste trabalho, além do método ter se mostrado eficiente em uma maior quantidade de casos, quando comparado com os trabalhos relacionados.

Tabela 19 – Comparação de trabalhos da literatura em uma abordagem Radiomics.

Trabalho	ACC	Kappa	ROC	Nº de características	Base de imagens	Amostras
(LUKE et al., 2014)	-	0,900	-	40	Base privada de NSCLC	56
(LI et al., 2014)	86,05%	-	-	1000	Base privada de NSCLC	208
(AERTS et al., 2014)	-	-	0,663	440	Base privada de NSCLC	125
(COROLLER et al., 2016)	-	-	0,630	15	Base privada de NSCLC	127
(HUYNH et al., 2016)	-	0,670	-	12	Base privada de NSCLC	113
(PATIL; MAHADEVAIAH; DEKKER, 2016)	88%	-	-	1000	NSCLC-Radiomics	317
(TIMMEREN et al., 2017)	-	0,690	-	149	Base privada de NSCLC	229
(SHEN et al., 2017)	-	-	0,671	507	Base privada de NSCLC	508
(LU et al., 2019)	-	-	0,857	1160	NSCLC-Radiomics	156
Pior resultado do método	97,5%	**0,967**	**0,985**	**8**	**NSCLC-Radiomics**	**319**
Melhor resultado do método	99,77%	**0,990**	**0,999**	**8**	**NSCLC-Radiomics**	**319**

Observam-se na Tabela 19 que os trabalhos selecionados são recentes no contexto da abordagem Radiomics e que o método proposto obteve melhores resultados, quando

observado os índices área sob a curva ROC e *Kappa*. Vale destacar que o método proposto obteve um melhor comportamento em uma maior quantidade de amostras, levando em consideração os trabalhos promissores da literatura no contexto de Radiomics.

O resultado do método proposto pode ser considerado promissor quando comparados à literatura, uma vez que apresenta valores superiores em todas as métricas de validação em relação a todos os trabalhos relacionados. Comparando os números de casos, o método proposto utilizou uma quantidade superior de casos do que a maioria dos trabalhos da literatura, pois consideram-se de extrema importância que os testes sejam realizados com o maior número de casos, garantindo uma generalização. Vale salientar, que a base de teste utilizada é uma base pública, o que garante que outros trabalhos possam fazer comparações.

6 CONCLUSÃO

A classificação de nódulos de NSCLC é uma tarefa de complexidade elevada e de grande relevância, dado pelo fato de que as ROIs de nódulo pulmonar, muitas vezes possuem textura bastante semelhante, mesmo sendo de classes distintas.

Em suma, este trabalho apresentou um método automático para classificação de nódulos de NSCLC em carcinoma de células grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados, explorando as propriedades de textura com 12 índices de diversidades filogenética e 9 índices de diversidade funcional nas representações de imagens de TC de 12, 11, 10 e 9 *bits*.

Os métodos propostos foram avaliados através da base NSCLC-Radiomics. Em resumo, alguns descritores merecem destaque nos métodos propostos, considerados também como uma das principais contribuições dessa dissertação. A primeira é a adaptação e o uso de índices de diversidade funcional já utilizados em mama, sendo alguns a primeira vez utilizados em pulmão, além de novos índices de diversidade funcional adaptados para o contexto de processamento de imagens e a primeira vez utilizados no contexto de Radiomics.

Os resultados produzidos pelos índices de diversidade filogenética em conjunto com os índices de diversidade funcional demostraram resultados promissores das técnicas de textura no contexto Radiomics, com uma taxa de acerto de 99,44%, um índice Kappa de 0,990 e uma área sob a curva ROC de 0,999.

Assim, os resultados mostraram que mesmo com uma proporção baixa de casos para treino, descreveram bem as classes da base, confirmando uma parte da teoria de Radiomics, de que as imagens podem fornecer características eficazes para classificação de nódulos de NSCLC, podendo proporcionar um tratamento precoce e com maiores chances de um prognóstico mais favorável ao paciente utilizando somente imagens digitais sem a necessidade de uma biópsia.

Para trabalhos futuros, pode-se fazer a combinação entre índices de diferentes grupos de diversidade filogenética e funcional, criando novas representações dos nódulos, e obtendo mais características para descrever cada nódulo. Além disso, pode-se combinar os índices com descritores de forma e outros descritores de textura. Como ideia para classificação, pode-se utilizar uma abordagem *Deep Learning* para verificar o desempenho

de classificação e comparar com os classificadores utilizados.

6.1 Produções científicas

A Tabela 20 lista os artigos científicos publicados e submetidos que possuem relação com a metodologia proposta neste trabalho. Todos os trabalhos foram produzidos como autor principal.

Tabela 20 – Artigos publicados e submetidos que possuem relação com a metodologia proposta.

Tipo	Artigo	Qualis	Status
Congresso	Antonino Calisto dos S. Neto , Joao O. B. Diniz, Pedro H. B. Diniz, Andre B. Cavalcante, Aristofanes C. Silva, Anselmo C. de Paiva (2018). Classificação do câncer de pulmão de células não pequenas usando ındice de diversidade filogenética e ındices de forma em uma abordagem Radiomics.Em CSBC 2018 - 18º SBCAS.	B4	Publicado
Capítulo de livro/ Congresso	Antonino C. dos S. Neto, Pedro H. B. Diniz, João O. B. Diniz, André B. Cavalcante, Aristófanes C. Silva, Anselmo C. de Paiva, and João D. S. de Almeida (2018). Diagnosis of Non-Small Cell Lung Cancer Using Phylogenetic Diversity in Radiomics Context. 15th International Conference on Image Analysis and Recognition.	B1	publicado

REFERÊNCIAS

AERTS, H.; VELAZQUEZ, E. R.; LEIJENAAR, R. T.; PARMAR, C.; GROSSMANN, P.; CARVALHO, S.; LAMBIN, P. **Data From NSCLC-Radiomics. The Cancer Imaging Archive**. 2015.

AERTS, H. J.; VELAZQUEZ, E. R.; LEIJENAAR, R. T.; PARMAR, C.; GROSSMANN, P.; CARVALHO, S.; BUSSINK, J.; MONSHOUWER, R.; HAIBE-KAINS, B.; RIETVELD, D. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. **Nature communications**, Nature Publishing Group, v. 5, p. 4006, 2014.

AMBELU, A.; LOCK, K.; GOETHALS, P. Comparison of modelling techniques to predict macroinvertebrate community composition in rivers of ethiopia. **Ecological Informatics**, v. 5, n. 2, p. 147 – 152, 2010. ISSN 1574-9541. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1574954109001083>.

AZEVEDO, M. J. C. de; SIQUEIRA, A. C. Estudo comparativo do nível da ansiedade de mulheres com câncer participantes e não participantes de grupos de apoio. **Revista FAROL**, v. 2, n. 2, p. 65–80, 2017.

BAILEY, B.; KLEIN, R.; LEEF, S. Hardware/software co-simulation strategies for the future. **Mentor Graphics Co., http://www. mentor. com**, 2000.

BAXEVANIS, A. D.; OUELLETTE, B. F. **Bioinformatics: a practical guide to the analysis of genes and proteins**. [S.l.]: John Wiley & Sons, 2004. v. 43.

BIAU, G. Analysis of a random forests model. **J. Mach. Learn. Res.**, JMLR.org, v. 13, n. 1, p. 1063–1095, abr. 2012. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=2503308.2343682>.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

CASTRO, J. L.; FLORES-HIDALGO, L.; MANTAS, C. J.; PUCHE, J. M. Extraction of fuzzy rules from support vector machines. **Fuzzy Sets and Systems**, Citeseer, v. 158, n. 18, p. 2057–2077, 2007.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. **ACM transactions on intelligent systems and technology (TIST)**, Acm, v. 2, n. 3, p. 27, 2011.

CIANCIARUSO, M. V.; SILVA, I. A.; BATALHA, M. A. Diversidades filogenética e funcional: novas abordagens para a ecologia de comunidades. **Biota Neotropica**, SciELO Brasil, v. 9, n. 3, 2009.

CLARKE, K.; WARWICK, R. An approach to statistical analysis and interpretation. **Change in marine communities**, v. 2, 1994.

COLELLA, S. R. L. et al. Segmentação dos pulmões e de suas lesões em imagens de tomografia computadorizada. [sn], 2017.

COROLLER, T. P.; AGRAWAL, V.; NARAYAN, V.; HOU, Y.; GROSSMANN, P.; LEE, S. W.; MAK, R. H.; AERTS, H. J. Radiomic phenotype features predict pathological response in non-small cell lung cancer. **Radiotherapy and Oncology**, Elsevier, v. 119, n. 3, p. 480–486, 2016.

CRUZ, A. A. **Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach**. [S.l.]: World Health Organization, 2007.

DOMENICO, M. D.; POZZI, D.; PALCHETTI, S.; DIGIACOMO, L.; IORIO, R.; ASTARITA, C.; FIORELLI, A.; PIERDILUCA, M.; SANTINI, M.; BARBARINO, M. et al. Nanoparticle-biomolecular corona: A new approach for the early detection of non-small-cell lung cancer. **Journal of cellular physiology**, Wiley Online Library, 2018.

DUDA, R. Pattern classification and scene analysis. **New-York, London, Sydny, Tronto A Wiley-Interscience Publication**, 1973.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification (2Nd Edition)**. New York, NY, USA: Wiley-Interscience, 2000. ISBN 0471056693.

FAIAL, M. Enfermagem e tuberculose em são vicente: o perfil dos utentes com tuberculose na delegacia de saúde em são vicente. 2017.

FAITH, D. P. Conservation evaluation and phylogenetic diversity. **Biological conservation**, Elsevier, v. 61, n. 1, p. 1–10, 1992.

_____. Phylogenetic pattern and the quantification of organismal biodiversity. **Phil. Trans. R. Soc. Lond. B**, The Royal Society, v. 345, n. 1311, p. 45–58, 1994.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861 – 874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016786550500303X>.

FILHO, A. O. d. C. et al. **Métodos para sistemas CAD e CADx de nódulo pulmonar baseada em tomografia computadorizada usando análise de forma e textura**. [S.l.]: Universidade Federal do Maranhão, 2016.

FILHO, A. O. de C.; SILVA, A. C.; PAIVA, A. C. de; NUNES, R. A.; GATTASS, M. Computer-aided diagnosis of lung nodules in computed tomography by using phylogenetic diversity, genetic algorithm, and svm. **Journal of digital imaging**, Springer, v. 30, n. 6, p. 812–822, 2017.

FINAK, G.; LANGWEILER, M.; JAIMES, M.; MALEK, M.; TAGHIYAR, J.; KORIN, Y.; RADDASSI, K.; DEVINE, L.; OBERMOSER, G.; PEKALSKI, M. L. et al. Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. **Scientific reports**, Nature Publishing Group, v. 6, p. 20686, 2016.

GIGER, M. L.; HUO, Z.; KUPINSKI, M. A.; VYBORNY, C. J. Computer-aided diagnosis in mammography. **Handbook of medical imaging**, The Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, v. 2, p. 915–1004, 2000.

GILLIES, R. J.; KINAHAN, P. E.; HRICAK, H. Radiomics: images are more than pictures, they are data. **Radiology**, Radiological Society of North America, v. 278, n. 2, p. 563–577, 2015.

GONZALEZ, R. C.; WOODS, R. C. **Processamento Digital de Imagens**. [S.l.]: Pearson Prentice Hall, 2010.

GUTIN, G.; YEO, A.; ZVEROVICH, A. Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the tsp. **Discrete Applied Mathematics**, Elsevier, v. 117, n. 1-3, p. 81–86, 2002.

HASSAN, M. Shafiq-ul; LATIFI, K.; ZHANG, G.; ULLAH, G.; GILLIES, R.; MOROS, E. Voxel size and gray level normalization of ct radiomic features in lung cancer. **Scientific reports**, Nature Publishing Group, v. 8, n. 1, p. 10545, 2018.

HATT, M.; TIXIER, F.; PIERCE, L.; KINAHAN, P. E.; REST, C. C. L.; VISVIKIS, D. Characterization of pet/ct images using texture analysis: the past, the present... any future? **European journal of nuclear medicine and molecular imaging**, Springer, v. 44, n. 1, p. 151–165, 2017.

HEEMSBERGEN, D.; BERG, M.; LOREAU, M.; HAL, J. V.; FABER, J.; VERHOEF, H. Biodiversity effects on soil processes explained by interspecific functional dissimilarity. **Science**, American Association for the Advancement of Science, v. 306, n. 5698, p. 1019–1020, 2004.

HOUNSFIELD, G. N. Computerized transverse axial scanning (tomography): Part 1. description of system. **The British journal of radiology**, The British Institute of Radiology, v. 46, n. 552, p. 1016–1022, 1973.

HUYNH, E.; COROLLER, T. P.; NARAYAN, V.; AGRAWAL, V.; HOU, Y.; ROMANO, J.; FRANCO, I.; MAK, R. H.; AERTS, H. J. Ct-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. **Radiotherapy and Oncology**, Elsevier, v. 120, n. 2, p. 258–266, 2016.

IZSÁK, J.; PAPP, L. A link between ecological diversity indices and measures of biodiversity. **Ecological Modelling**, Elsevier, v. 130, n. 1-3, p. 151–156, 2000.

JÚNIOR, P. R. B. d. S.; SZWARCWALD, C. L.; JÚNIOR, A. B.; CARVALHO, M. F. d.; CASTILHO, E. A. d. Infecção pelo hiv durante a gestação: estudo-sentinela parturiente, brasil, 2002. **Revista de Saúde Pública**, SciELO Public Health, v. 38, p. 764–772, 2004.

KANUNGO, T.; HARALICK, R. M.; PHILLIPS, I. Global and local document degradation models. In: IEEE. **Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on**. [S.l.], 1993. p. 730–734.

KEITH, M.; CHIMIMBA, C.; REYERS, B.; JAARSVELD, A. V. Taxonomic and phylogenetic distinctiveness in regional conservation assessments: a case study based on extant south african chiroptera and carnivora. In: CAMBRIDGE UNIVERSITY PRESS. **Animal Conservation forum**. [S.l.], 2005. v. 8, n. 3, p. 279–288.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, [Wiley, International Biometric Society], v. 33, n. 1, p. 159–174, 1977. ISSN 0006341X, 15410420. Disponível em: <http://www.jstor.org/stable/2529310>.

LEE, S. L. A.; KOUZANI, A. Z.; HU, E. J. Random forest based lung nodule classification aided by clustering. **Computerized medical imaging and graphics**, Elsevier, v. 34, n. 7, p. 535–542, 2010.

LENTO, G. C. **Random forest em dados desbalanceados: uma aplicação na modelagem de churn em seguro saúde**. Tese (Doutorado), 2017.

LI, B.-Q.; YOU, J.; HUANG, T.; CAI, Y.-D. Classification of non-small cell lung cancer based on copy number alterations. **PLoS One**, Public Library of Science, v. 9, n. 2, p. e88300, 2014.

LIU, Y.; WANG, H.; LI, Q.; MCGETTIGAN, M. J.; BALAGURUNATHAN, Y.; GARCIA, A. L.; THOMPSON, Z. J.; HEINE, J. J.; YE, Z.; GILLIES, R. J. et al. Radiologic features of small pulmonary nodules and lung cancer risk in the national lung screening trial: A nested case-control study. **Radiology**, Radiological Society of North America, v. 286, n. 1, p. 298–306, 2017.

LOONEY, C. G. et al. **Pattern recognition using neural networks: theory and algorithms for engineers and scientists**. [S.l.]: Oxford University Press New York, 1997.

LOSOS, J. B. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. **Ecology letters**, Wiley Online Library, v. 11, n. 10, p. 995–1003, 2008.

LU, L.; LI, L.; YANG, H.; SCHWARTZ, L.; ZHAO, B. et al. Radiomics for classifying histological subtypes of lung cancer based on multiphasic contrast-enhanced computed tomography. **Journal of computer assisted tomography**, 2019.

LUKE, J. J.; OXNARD, G. R.; PAWELETZ, C. P.; CAMIDGE, D. R.; HEYMACH, J. V.; SOLIT, D. B.; JOHNSON, B. E. Realizing the potential of plasma genotyping in an age of genotype-directed therapies. **JNCI: Journal of the National Cancer Institute**, Oxford University Press, v. 106, n. 8, 2014.

MACKEIVICZ, G. A. O. Incidências de infecções urinárias: uma análise a partir de um laboratório privado na cidade de carambeí, estado do paraná. 2017.

MACMAHON, H.; NAIDICH, D. P.; GOO, J. M.; LEE, K. S.; LEUNG, A. N.; MAYO, J. R.; MEHTA, A. C.; OHNO, Y.; POWELL, C. A.; PROKOP, M. et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. **Radiology**, Radiological Society of North America, v. 284, n. 1, p. 228–243, 2017.

MAGURRAN, A. E. **Measuring biological diversity**. [S.l.]: John Wiley & Sons, 2013.

MOHAN, A.; POOBAL, S. Crack detection using image processing: A critical review and analysis. **Alexandria Engineering Journal**, Elsevier, v. 57, n. 2, p. 787–798, 2018.

NETO, A. C. d. S.; DINIZ, J. O.; DINIZ, P. H.; CAVALCANTE, A. B.; SILVA, A. C.; PAIVA, A. C. de. Classificação do câncer de pulmão de células não pequenas usando índice de diversidade filogenética e índices de forma em uma abordagem radiomics. In: SBC. **18º Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2018)**. [S.l.], 2018. v. 18, n. 1/2018.

NETO, O. P. S.; SILVA, A. C.; PAIVA, A. C.; GATTASS, M. Automatic mass detection in mammography images using particle swarm optimization and functional diversity indexes. **Multimedia Tools and Applications**, Springer, v. 76, n. 18, p. 19263–19289, 2017.

OLIVEIRA, F. S. S. de; FILHO, A. O. de C.; SILVA, A. C.; PAIVA, A. C. de; GATTASS, M. Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and svm. **Computers in biology and medicine**, Elsevier, v. 57, p. 42–53, 2015.

OTSU, N. A threshold selection method from gray-level histograms. **IEEE transactions on systems, man, and cybernetics**, IEEE, v. 9, n. 1, p. 62–66, 1979.

PAIM, E. P.; RIGHI, R. da R.; COSTA, C. A. da. Explorando o paradigma publish/subscribe e a elasticidade em níveis aplicados ao procedimento de telemedicina. **Revista Brasileira de Computação Aplicada**, v. 10, n. 1, p. 11–22, 2018.

PATIL, R.; MAHADEVAIAH, G.; DEKKER, A. An approach toward automatic classification of tumor histopathology of non–small cell lung cancer based on radiomic features. **Tomography**, Grapho Publications, v. 2, n. 4, p. 374, 2016.

PEDRINI, H.; SCHWARTZ, W. R. **Análise de imagens digitais: princípios, algoritmos e aplicações**. [S.l.]: Thomson Learning, 2008.

PETCHEY, O. L.; GASTON, K. J. Functional diversity: back to basics and looking forward. **Ecology letters**, Wiley Online Library, v. 9, n. 6, p. 741–758, 2006.

PIEŃKOWSKI, G.; WESTWALEWICZ-MOGILSKA, E. Trace fossils from the podhale flysch basin, poland-an example of ecologically-based lithocorrelation. **Lethaia**, Wiley Online Library, v. 19, n. 1, p. 53–65, 1986.

PIENKOWSKI, M.; WATKINSON, A.; KERBY, G.; CLARKE, K.; WARWICK, R. A taxonomic distinctness index and its statistical properties. **Journal of applied ecology**, Wiley Online Library, v. 35, n. 4, p. 523–531, 1998.

RAO, C. R. Diversity and dissimilarity coefficients: a unified approach. **Theoretical population biology**, Elsevier, v. 21, n. 1, p. 24–43, 1982.

ROCHA, M. A.; ISHIKAWA, W. Y.; SZARF, G. O que o cardiologista precisa saber sobre achados torácicos na tomografia. **Rev. Soc. Cardiol. Estado de Säo Paulo**, v. 27, n. 2, p. f–109, 2017.

RODRIGUES, A. S.; GASTON, K. J. Maximising phylogenetic diversity in the selection of networks of conservation areas. **Biological Conservation**, Elsevier, v. 105, n. 1, p. 103–111, 2002.

ROSENFIELD, G. H.; FITZPATRICK-LINS, K. A coefficient of agreement as a measure of thematic classification accuracy. **Photogrammetric engineering and remote sensing**, v. 52, n. 2, p. 223–227, 1986.

SCHMID, U.; LIESENFELD, K.-H.; FLEURY, A.; DALLINGER, C.; FREIWALD, M. Population pharmacokinetics of nintedanib, an inhibitor of tyrosine kinases, in patients with non-small cell lung cancer or idiopathic pulmonary fibrosis. **Cancer chemotherapy and pharmacology**, Springer, v. 81, n. 1, p. 89–101, 2018.

SCHWEIGER, O.; KLOTZ, S.; DURKA, W.; KÜHN, I. A comparative test of phylogenetic diversity indices. **Oecologia**, Springer, v. 157, n. 3, p. 485–495, 2008.

SHEN, C.; LIU, Z.; GUAN, M.; SONG, J.; LIAN, Y.; WANG, S.; TANG, Z.; DONG, D.; KONG, L.; WANG, M. et al. 2d and 3d ct radiomics features prognostic performance comparison in non-small cell lung cancer. **Translational oncology**, Elsevier, v. 10, n. 6, p. 886–894, 2017.

SOARES, J. V.; LEANDRO, J. J.; CESAR, R. M.; JELINEK, H. F.; CREE, M. J. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. **IEEE Transactions on medical Imaging**, IEEE, v. 25, n. 9, p. 1214–1222, 2006.

SOLOW, A.; POLASKY, S.; BROADUS, J. On the measurement of biological diversity. **Journal of Environmental Economics and Management**, Elsevier, v. 24, n. 1, p. 60–68, 1993.

TAN, W.; DESAI, T. A. Microfluidic patterning of cellular biopolymer matrices for biomimetic 3-d structures. **Biomedical Microdevices**, Springer, v. 5, n. 3, p. 235–244, 2003.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition, Fourth Edition**. 4th. ed. Orlando, FL, USA: Academic Press, Inc., 2008. ISBN 1597492728, 9781597492720.

TILMAN, D. Functional diversity. **Encyclopedia of biodiversity**, v. 3, n. 1, p. 109–120, 2001.

TIMMEREN, J. E. van; LEIJENAAR, R. T.; ELMPT, W. van; REYMEN, B.; OBERIJE, C.; MONSHOUWER, R.; BUSSINK, J.; BRINK, C.; HANSEN, O.; LAMBIN, P. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam ct images. **Radiotherapy and Oncology**, Elsevier, v. 123, n. 3, p. 363–369, 2017.

TSUKAZAN, M. T. R.; VIGO, Á.; SILVA, V. D. da; BARRIOS, C. H.; RIOS, J. de O.; PINTO, J. A. de F. Câncer de pulmão: mudanças na histologia, sexo e idade nos últimos 30 anos no brasil. **Jornal Brasileiro de Pneumologia**, Jornal Brasileiro de Pneumologia, v. 43, n. 5, p. 363–367, 2017.

VANE-WRIGHT, R. I.; HUMPHRIES, C. J.; WILLIAMS, P. H. What to protect?—systematics and the agony of choice. **Biological conservation**, Elsevier, v. 55, n. 3, p. 235–254, 1991.

VAPNIK, V. N.; VAPNIK, V. **Statistical learning theory**. [S.l.]: Wiley New York, 1998. v. 1.

VARMA, M. J. O.; BREULS, R. G.; SCHOUTEN, T. E.; JURGENS, W. J.; BONTKES, H. J.; SCHUURHUIS, G. J.; HAM, S. M. V.; MILLIGEN, F. J. V. Phenotypical and functional characterization of freshly isolated adipose tissue-derived stem cells. **Stem cells and development**, Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA, v. 16, n. 1, p. 91–104, 2007.

VIERA, A. J.; GARRETT, J. M. et al. Understanding interobserver agreement: the kappa statistic. **Fam Med**, v. 37, n. 5, p. 360–363, 2005.

WEITZMAN, M. L. On diversity. **The Quarterly Journal of Economics**, MIT Press, v. 107, n. 2, p. 363–405, 1992.

ZHANG, Y.; OIKONOMOU, A.; WONG, A.; HAIDER, M. A.; KHALVATI, F. Radiomics-based prognosis analysis for non-small cell lung cancer. **Scientific reports**, Nature Publishing Group, v. 7, p. 46349, 2017.

ZHUANG, L.; DAI, H. Parameter optimization of kernel-based one-class classifier on imbalance learning. **Journal of Computers**, Citeseer, v. 1, n. 7, p. 32–40, 2006.