Classificação do câncer de pulmão de células não pequenas usando índice de diversidade filogenética e índices de forma em uma abordagem Radiomics

Antonino Calisto dos S. Neto [1], João O. B. Diniz [1], Pedro H. B. Diniz [1], André B. Cavalcante [1], Aristófanes C. Silva [1], Anselmo C. de Paiva [1]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
CEP 65080-805 – São Luís – MA – Brasil

{antuninosantos,joao.obd,phb.diniz,abcborges}@gmail.com,
ari@dee.ufma.br, anselmo.c.paiva@gmail.com

Abstract. *Lung cancer is the most common type of cancer and has the highest mortality rate in the world. The automatic process for the diagnosis by computer vision systems, through medical images, provides an interpretation about the pathology. This work proposes the classification of non-small cell lung cancer using as the texture descriptor the index of phylogenetic diversity based on topology and some indexes of shape based on region and contour, adapting them to the Radiomics approach. Tests showed promising results of 98.83 % accuracy, a Kappa index of 0.993 and an area under the ROC curve of 0.999.*

Resumo. *O câncer de pulmão é o tipo de câncer mais comum e tem a maior taxa de mortalidade no mundo. O processo automático para o diagnóstico por sistemas de visão por computador, através de imagens médicas, fornece uma interpretação sobre a patologia. Este trabalho propõe a classificação do câncer de pulmão de células não pequenas utilizando como descritor de textura o índice de diversidade filogenética baseado em topologia e alguns índices de forma baseados em região e contorno, adaptando-os para a abordagem Radiomics.Os testes mostraram resultados promissores de 98,83% de acurácia, um índice Kappa de 0,993 e uma área sob a curva ROC de 0,999.*

1. Introdução

Denomina-se câncer como um conjunto de mais de cem patologias caracterizadas pelo crescimento desordenado de células que podem invadir tecidos ou órgãos. Suas causas podem ser tanto internas (relacionadas à capacidade de defesa do organismo) ou externas (ligadas ao meio ambiente e aos hábitos e/ou costumes). Segundo [Setio et al. 2016], o câncer de pulmão atualmente é o tipo de carcinoma com a maior taxa de mortalidade entre homens. Em mulheres, o câncer de pulmão perde apenas para o de mama.

O tratamento precoce do câncer de pulmão aumenta as chances de sobrevivência do paciente em até 90% [Ferlay et al. 2015]. Pode-se citar o exame por imagem como um dos métodos mais eficazes para o diagnóstico de patologias. O surgimento da Tomografia Computadorizada (TC) tornou mais eficaz a detecção e diagnóstico de nódulos pulmonares, sendo de grande uso no tratamento do câncer.

Porém, a análise das características das imagens de um exame feita por especilistas, pode ocasionar alguns problemas no diagnóstico por ser realizada de forma

subjetiva, como por exemplo fadiga visual, distração, entre outros. Devido a isto, várias técnicas computacionais vêm sendo desenvolvidas para serem utilizadas em conjunto com sistemas de diagnóstico auxiliados por computador (*Computer Aided Diagnosis* - CADx) com a finalidade de melhorar a acurácia do diagnóstico, servindo como um auxílio ao especialista na tomada de decisões.

Assim, com o grande auxílio de ferramentas CADx, tornou-se mais fácil a extração de características quantitativas relevantes, que resultaram na conversão de imagens em dados, e na utilização destes dados para suporte na decisão. Esta prática é chamada de Radiomics [Aerts et al. 2015].

Um processo que é conhecido como radiomics, é motivado pelo conceito de que as imagens biomédicas contêm informações que refletem a fisiopatologia e que podem revelar algum prognóstico da patologia [Aerts et al. 2015].

Neste contexto, apresenta-se um método de classificação de lesões pulmonares, utilizando o índice diversidade filogenética baseado em topologia, e índices baseados na forma dos nódulos pulmonares, em exames de TC de pacientes com câncer de pulmão de células não pequenas (*Non-Small Cell Lung Cancer* - NSCLC).

Como contribuição do trabalho, pode-se destacar: (a) o uso de índices de diversidade filogenético baseado em topologia, já utilizados em outros contexto, sendo pela primeira vez utilizados juntos no contexto de Radiomics; e (b) a utilização do índice filogenético e índices de forma na aplicação em um banco de dados com uma quantidade de pacientes superior aos trabalhos recentes da literatura Radiomics, mostrando que a metodologia proposta é promissora na tarefa de classificação dos nódulos de NSCLC.

As seções são divididas da seguinte forma: na Seção 2, apresenta-se os trabalhos relacionados à classificação de nódulos de NSCLC; na Seção 3 descreve-se o método proposta neste artigo, onde é definido as técnicas e sua aplicação como uma alternativa de fornecer um diagnóstico do problema; os resultados são apresentados e discutidos na Seção 4; finalmente, na Seção 5 conclui-se o trabalho, propondo-se futuras melhorias.

2. Trabalhos Relacionados

Na literatura, muitos trabalhos estão relacionados com o desenvolvimento de sistemas automáticos para classificação dos nódulos pulmonares no contexto radiomics.

No trabalho de [Shen et al. 2017] é feita uma comparação das radiografias 2D e 3D, apresentado-se as eficiências das características nesses dois tipos de imagens. São analisados 588 pacientes, sendo extraídas 1014 características Radiomics de forma e textura (507 características para as imagens 2D, e 507 características para as imagens 3D). Para classificação, utilizou-se a ferramenta WEKA [Hall et al. 2009]. Foi utilizado a área sob a curva ROC (*Rceiver Operating Characteristic*) para avaliar o desempenho de previsão dos classificadores treinados (máquina de vetor de suporte e regressão logística). Dois grupos de imagens foram separados para treinamento (463 pacientes para cada tipo de imagens 2D e 3D), e outro para validação (125 pacientes para cada tipo de imagens 2D e 3D). No grupo de treinamento, nas imagens 2D, obteve-se uma área sob a curva ROC de 0,653, e no 3D de 0,671. Na validação, obteve-se uma área sob a curva ROC nas imagens 2D de 0,755 e nas 3D, de 0,663.

Em [Coroller et al. 2016] avalia que as características Radiomics são capazes de prever a resposta patológica em paciente de NSCLC localmente estabelecidas. Foram utilizados exames de TC de 127 pacientes, extraindo 15 características Radiomics de textura selecionadas quanto a relevância e estabilidade quanto ao poder de prever a resposta patológica. A métrica de validação utilizada foi a área sob a curva ROC, alcançando um valor de 0,63 como resultado.

No trabalho de [Huynh et al. 2016] demonstra-se que as características em uma base Radiomics em exames de TC de nódulos de NSCLC em estágios iniciais, tratados com terapia de radiação de corpo estereotáxico, possuem poder de predição. 12 características Radiomics de forma foram selecionadas de acordo com sua relevância e estabilidade. A métrica de validação utilizada para seu valor prognóstico foi o índice de concordância *Kappa*. O *Kappa* teve um valor preditivo de 0,67.

Em [van Timmeren et al. 2017], extraiu-se características Radiomics baseadas em conebeam TC (CBCT) para pacientes com NSCLC. O trabalho divide-se em duas partes. Na primeira parte é separada uma base com 132 pacientes, utilizada para treinamento. A segunda parte, separou-se outras duas bases de imagens referentes ao teste, uma com 62 pacientes, e outra base com 94 pacientes. Inicialmente foram extraídas 1119 características de forma e textura, e em seguida selecionou-se 149 características com um algoritmo de seleção. Como resultado, obteve-se um índice *Kappa* de 0,69.

Estes são exemplos de sistemas que foram desenvolvidos para a extração de características em exames de TC de NSCLC no contexto da abordagem Radiomics. Dois pontos importantes comuns nestes trabalhos são que os índices de validação obtiveram valores baixos, e o número pequeno de casos. Em nosso método, tentar-se-á explorar essas deficiências com o intuito de melhorar as métricas que avaliam o desempenho, como acurácia, área sob a curva ROC e índice *Kappa*, em um maior número de casos, e com uma quantidade menor de características.

3. Materiais e Metodologia Proposta

A metodologia desenvolvida para classificação de NSCLC (carcinoma de célula grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados) seguiu algumas etapas: primeiramente houve a aquisição das imagens; porteriormente, foram extraídos os nódulos pulmonares de acordo com as marcações dos especialistas na base; em seguida, houve a extração das características a partir de índices de diversidade filogenética ([de Carvalho Filho et al. 2017]) e de forma; depois, utilizou-se a ferramenta WEKA [Hall et al. 2009] para classificação. A ilustração da metodologia pode ser observado na Figura 1.

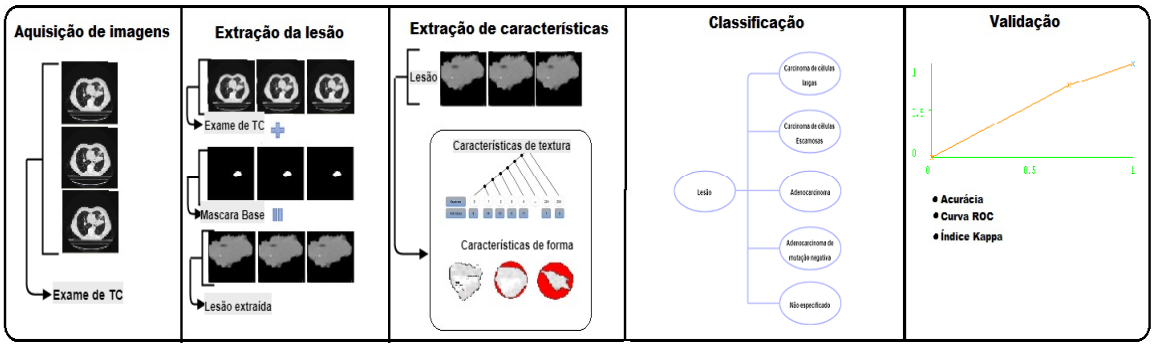

Figura 1. Ilustração da metodologia proposta.

3.1. Aquisição de imagens

Uma base de dados Radiomics é aquela base que possui imagens dos tecidos dos pacientes, tais que os tecidos já foram analisados em laboratório, e comprovou-se que as imagens podem fornecer predições a respeito da patologia, que condizem com o diagnóstico dos tecidos e as opiniões dos especialistas [Oliveira et al. 2017].

O banco de imagens utilizado neste trabalho foi a base de imagens pública NSCLC-Radiomics[Aerts et al. 2015], disponibilizada pelo *Cancer Imaging Archive*. Este banco de imagens contém imagens de 422 pacientes com NSCLC, sendo que cada paciente pode pertencer a umas das 5 classes: carcinoma de célula grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados. Esta base foi utilizada para extração das caracteristicas pelos algoritmos desenvolvidos [Aerts et al. 2015].

A criação da NSCLC-Radiomics resultou em uma base composta de exames de pacientes que variam entre 110 e 180 fatias, sendo que cada exame contém um arquivo RT-STRUCT (*Radiotherapy Structure Set*) com marcações da base (classe que o paciente está incluído, localização dos nódulos, idade, entre outros) [Aerts et al. 2015].

3.2. Descritores de Características de Textura

A extração de características de textura do sistema CADx foi desenvolvida com a aplicação do índice de diversidade filogenética baseado na topologia das espécies, adaptados para o contexto Radiomics de NSCLC, que serão detalhados no decorrer do trabalho.

Uma maneira de representar estas relações são as árvores filogenéticas, ou filogenia [de Oliveira et al. 2015], representando os organismos pelas folhas e os ancestrais pelos nós internos. Uma ilustração da árvore filogenética pode ser observado na Figura 2, representando um cladograma, uma das formas representativas das relações dos ancestrais com os organismos.

A diversidade filogenética é a medida de uma comunidade que contém as relações filogenéticas das espécies [de Oliveira et al. 2015]. Considera-se a forma mais simples de representação dos índices de diversidade em imagens, sendo a imagem a representação da comunidade ou região [de Oliveira et al. 2015].

A extração dos índices de diversidade filogenética através das árvores filogenéticas são utilizados na biologia para comparar amostras de comportamento entre espécies

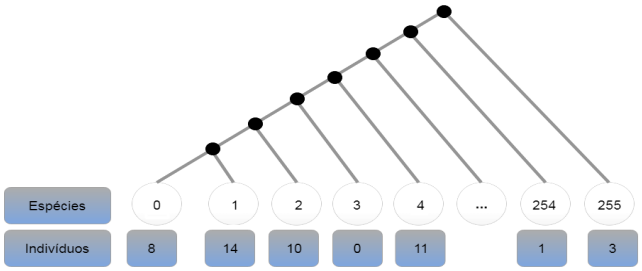

Figura 2. Árvore filogenética sob a forma de um cladograma. Adaptada de [de Carvalho Filho et al. 2017]

pertencentes a diferentes regiões, e na área da computação pode-se utililizá-los para classificar o padrão de nódulos de NSCLC, devido os índices de diversidade filogenética possuirem um alto grau de discriminação. Portanto, deve-se fazer uma correlação entre a biologia e a metodologia proposta, como ilustrado na Tabela 1.

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

Biologia	**Metodologia**
Comunidade	Regiões do nódulo da imagem de TC
Espécie	Cada Unidade de Hounsfield (UH) da região
Indivíduos	Quantidade de voxels de cada UH
Abundância relativa	Número de voxels encontrados no nódulo que possuem o mesmo valor UH
Riqueza de espécies	Número de voxels encontrados no nódulo

Para o método proposto, será utilizado um dos índices filogenéticos baseados na topologia das espécies. O índice é denominado: soma básica dos pesos (Q). Este índice se baseia na topologia que visa o relacionamento da árvore das espécies presentes com toda a comunidade, fazendo com que mencione o grau de parentesco (similiaridade) entre as espécies [Keith et al. 2005].

O trabalho de [Vane-Wright et al. 1991] foi um dos primeiros estudos a propor a aplicação de métodos baseados em topologia, cuja ideia leva em consideração a ordem filogenética dentro de um determinado grupo. Neste tipo de índice, cada espécie de uma comunidade é ponderada pelo número de nós entre a espécie e a raiz da árvore fologenética (cladograma). Assim, atribui-se os maiores pesos às espécies que possuem a maior distância da raiz.

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 nas Equações 1, 2 e 3.

$$I = \sum_{i=0}^{N-1} I_i \tag{1}$$

$$Q_i = \frac{I}{I_i} \tag{2}$$

$$Q = \sum_{i=0}^{N-1} Q_i \tag{3}$$

onde Q_i representa o quociente do total de nós dos caminhos, partindo da raiz até todas as espécies da árvore filogenética por I_i, representando o número de nós entre a raiz e a espécie *i*.

3.3. Descritores de Características de Forma

Foi utilizado a extração de características de forma ados nódulos pulmonares, devido estas serem capazes de fornecer informações capazes de distinguir a classe a qual o nódulo pertencente. A Figura 3 ilustra alguns exemplos de classes de nódulos e suas variações de forma, em que alguns nódulos possuem formas mais esféricas, e outros são mais espiculados ou pontiagudos.

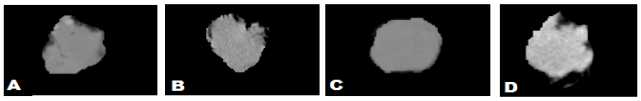

Figura 3. Exemplos de formas de nódulos da base.

A seguir serão descritos os índices de forma utilizados para desenvolver a metodologia proposta.

3.3.1. Área

A área (*A*) de um nódulo é dada pela quantidade de *voxels* que o nódulo possui.

3.3.2. Volume

O volume (*V*) de um nódulo representa a quantidade de *voxels* que a borda do nódulo possui.

3.3.3. Desproporção Esférica

Segundo [da Silva Sousa et al. 2007] o cálculo da desproporção esférica é feito por meio da comparação entre a área de um objeto com a área que este objeto viria a ocupar caso fosse totalmente esférico, podendo assim, obter uma diferenciação quantitativa de sua estrutura morfológica. A densidade esférica é representada pelas Equações 4 e 5 e um exemplo de sua ilustração pode ser observada na Figura 4.

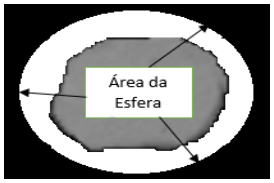

Figura 4. Imagem representativa da Desproporção Esférica.

$$R = \frac{\sqrt[3]{3V}}{4\pi} \tag{4}$$

$$D_e = \frac{A}{4\pi R^2} \tag{5}$$

onde D_e é o valor da desproporção espérica, *V* é o volume do nódulo, *R* representa o raio do nódulo e *A* é a área da superficie do nódulo.

3.3.4. Compacidade

O cálculo da compacidade é feito a partir da medição da densidade do nódulo em relação a uma figura geométrica perfeitamente densa, ou seja, uma esfera. Na Figura 3C pode-se observar um nódulo com uma forma mais arredondada, ou compacto, já na Figura 3B observa-se um nódulo com maiores espiculações, ou menos compacto. O cálculo da compacidade é representado pela Equação 6.

$$C = \frac{V}{\frac{4\pi R^3}{3}} \tag{6}$$

onde *V* é o volume do nódulo e *R* é o raio do nódulo.

3.3.5. Variância da Borda

Este descritor foi utilizado com a intenção de calcular a variação da borda de um nódulo, com o intuito de diferenciar as diferentes classes da base. As Equações 7 e 8 representam os cálculos da variância da borda.

$$D_i = \sqrt{(Cx_2 - X_1)^2 + (Cy_2 - Y_1)^2} \tag{7}$$

$$V_B = \frac{\sum_{i=0}^{N-1} D_i}{N} \tag{8}$$

onde D_i é a distância do *voxel* que está mais no centro do nódulo (centróide do nódulo) para um *voxel* da borda, sendo Cx_2 e Cy_2 as coordenadas *x* e *y* do centróide, respectivamente e X_1 e Y_1 as coordenadas do *voxel* da borda. V_B representa a distância média das distâncias dos *voxels* do nódulo, onde *N* é a quantidade de *voxels* existentes no nódulo.

3.4. Classificação

A classificação é a etapa onde um grupo de dados são divididos obedecendo alguma métrica, formando classes ou grupos. Esta técnica é muito utilizada no contexto de treinamento de algoritmos em processamento de imagens [Giger et al. 2000].

A ferramenta do WEKA utilizada para a realização dos testes foi o Auto-WEKA [Thornton et al. 2013], devido permitir a automatização dos testes e da estimação dos parâmetros de vários algoritmos de classificação, sendo selecionados quatro classificadores: o *Random Forest* [Dean 2014], *J48* [Dean 2014], *Bagging* [Dean 2014] e o *JRip* [Dean 2014] e utilizou-se o método *K-fold cross-validation*, usando 10 conjuntos de características (k=10) onde 9 são para treinamento e 1 para testes.

Para validar o método proposto, utilizou-se algumas métricas de validação: acurácia, índice *Kappa* [Cook 1998] e área sob a curva ROC [Hanley and McNeil 1982].

4. Resultados e Discussão

Para validação da metodologia desenvolvida, foram realizados testes com a base NSCLC-Radiomics, utilizando 319 dos 422 exames, devido 103 dos exames disponíveis na base pública estarem sem marcações.

Nas seções que seguem, serão apresentados os resultados e 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. Por fim, será comparado o melhor resultado com a literatura.

4.1. Descritor de Textura

Os resultados aqui apresentados referem-se ao descritor de diversidade filogenética baseado em topologia descrito na Subseção 3.2.

Analisando os resultados na Tabela 2 pode-se observar que o classificador *JRip* obteve melhor resultado com uma acurácia de 98,10%, onde o classificador *RandomForest* obteve o resultado mais baixo dos classificadores com uma acurácia de 97,46%, porém ainda considerado um resultado promissor.

Tabela 2. Resultados para a classificação com análise de textura, utilizando índice de diversidade filogenética com base em topologia "soma básica dos pesos".

Classificador	Acurácia	ROC	Kappa
Random Forest	97,46%	0,976	0,966
Bagging	97,78%	0,979	0,970
JRip	**98,10%**	**0 981**	**0,974**
J48	97,78%	0,986	0,970

Pode-se observar uma baixa variação entre os resultados dos classificadores, provando que este descritor conseguiu descrever de forma eficaz as texturas do nódulos, independentemente do tipo de classificador utilizado. A possível razão deste índice ter obtido bons resultados, pode ter sido devido os índices de diversidade filogenética baseados em topologia, evidenciarem as características extraídas que levam em consideração os nós entre as espécies e a raiz da árvore filogenética, generalizando as características diferentes entre as classes. A possível razão pelo pequeno erro de aproximadamente 2% tenha sido devido a semelhança de textura entre as classes, confundindo o classificador, como pode ser observado na Figura 5A, 5B, 5C, 5D, e 5E, correspondentes aos carcinoma de células grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados, respectivamente.

Figura 5. Exemplos de nódulos com texturas semelhantes na base.

4.2. Descritores de Forma

Os resultados aqui apresentados referem-se aos descritores que analisam as formas dos nódulos descritos na Subseção 3.3.

Como pode ser observado na Tabela 3, o classificador *JRip* foi o que obteve o melhor resultado, com uma acurácia de 34,17%, e o classificador *RandomFOrest* obteve o pior resultado com uma acurácia de 27,53%, mostrando que apenas os descritores de formas aqui utilizados não foram eficazes para descrever as propriedades dos nódulos.

Tabela 3. Resultados para a classificação com análise de forma, utilizando os índices de forma.

Classificador	Acurácia	ROC	Kappa
Random Forest	27,53%	0,523	0,024
Bagging	32,91%	0,560	0,076
JRip	**34,17%**	**0,524**	**0,040**
J48	30,37%	0,504	0,055

Ainda analisando a Tabela 3, pode-se observar que a forma dos nódulos não foram eficazes para diferenciar as classes da base. Uma possível razão para que isso ocorra, deve-se ao fato de que as formas das classes trabalhadas são bastante semelhantes, como pode ser observado na Figura 6A, 6B, 6C, 6D, e 6E, correspondentes aos carcinoma de células grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados, respectivamente.

Figura 6. Exemplos de nódulos com formas semelhantes na base.

4.3. Descritores de textura e Forma

Os resultados apresentados na Tabela 4, referem-se à combinação dos descritores de características de textura e de forma, com a seleção de característica utilizando o algoritmo *Greedy Stepwise* [Azuaje 2006].

Na metodologia, o algoritmo *Greedy Stepwise* selecionou o índice de diversidade filogenética baseado em topologia e os descritores de forma desproporção esférica, compacidade e variância da borda, aumentando a eficiência e fazendo a metodologia se destacar ainda mais quando comparada aos trabalhos relevantes da literatura Radiomics.

Na Tabela 4, observa-se que os resultados são promissores, destacando-se o classificador *J48* com 98,83% de acurácia, sendo os classificadores *Bagging*, *JRip* e *RandomForest* obtiveram os resultados mais baixos com 98,42% de acurácia, porém ainda é considerado um resultado promissor na literatura.

Tabela 4. Resultados para a classificação com os índices de forma e textura.

Classificador	Acurácia	ROC	Kappa
Random Forest	98,51%	0,999	0,979
Bagging	98,41%	0,999	0,980
JRip	98,41%	0,990	0,979
J48	**98,83%**	**0,999**	**0,993**

Analisando a Tabela 4, observa-se que a união do índice de diversidade filogenética com os índices de forma selecionados pelo algoritmo *Greedy Stepwise*

obtiveram resultados promissores. Uma das possíveis razões pelo resultado promissor, foi devido os descritores de forma evidenciarem propriedades que as características de textura não conseguiram destacar, tornando possível distinguir bem as classes das bases de teste. O erro de aproximadamente 1,2% nas acurácias, foi devido muitas vezes as texturas dos nódulos das classes envolvidas, estarem próximas uma das outras, além das formas das classes dos nódulos malignos desta base, possuirem formatos semelhantes. Outra razão pelo erro ocorrido, pode ter sido por conta de possíveis excessos devido as divergências das marcações quanto a localização precisa dos nódulos.

4.4. Comparação de trabalhos da literatura no contexto da abordagem Radiomics

Não é possível fazer uma comparação fidedigna com os trabalhos relacionados, já que não utilizam uma base pública ou disponível para análise de nódulos de NSCLC e também devido o conjunto de casos ser diferente. Porém, utiliza-se as métricas de validação dos trabalhos relacionados para uma comparação, devido classificarem em algum dos cinco tipos de nódulos de NSCLC dessa metodologia proposta.

Na Tabela 5 tem-se uma visão resumida (acurácia, ROC, número de características, base de imagens utilizada e número de casos na base) dos resultados encontrados nos trabalhos relacionados e no método proposto. Observando-a, pode-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, área sob a curva ROC e índice *Kappa*), além do método ter mostrado eficiente em uma maior quantidade de casos, quando comparado com os trabalhos relacionados.

Tabela 5. Resumo comparativo de trabalhos da literatura no contexto da abordagem Radiomics.

Trabalho	Kappa	ROC	Nº de características	Base de imagens	Nº de casos
[Shen et al. 2017]	-	0,663	1014	Base privada	125
[Coroller et al. 2016]	-	0,630	15	Base privada	127
[Huynh et al. 2016]	0,670	-	12	Base privada	113
[van Timmeren et al. 2017]	0,690	-	149	Base privada	156
Metodologia proposta	**0,993**	**0,999**	**4**	**NSCLC-Radiomics**	**319**

Analisando a Tabela 5, observa-se um resultado do método proposto considerado promissor por parte do método proposto, uma vez que apresenta valores superiores em todas às 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 considera-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.

5. Conclusão

Este trabalho apresentou um método automático para classificação de nódulos de NSCLC em carcinoma de célula grande, carcinoma de células escamosas, adenocarciona, adenocarcinoma de mutação negativa e não especificados, servindo como uma segunda opinião para o especialista.

Os resultados obtidos confirmam um resultado considerado promissor das técnicas de extração de textura e forma, sugerindo um diagnóstico promissor no contexto

Radiomics, com uma taxa de acerto de 98,83% e uma curva ROC de 0,999. Assim, conclui-se que as imagens podem fornecer características eficazes para a classificação de nódulos de NSCLC, proporcionando um tratamento precoce e com maiores chances de um prognóstico favorável ao paciente.

Como trabalhos futuros, pretende-se utilizar métodos de classificação usando abordagens *Deep Learning*, com o intuito de comparação com os classificadores utilizados no método proposto. Além disso, pode-se combinar os índices proposto com outros índices de textura e/ou índices de forma, para observar melhores resultados.

Agradecimentos

Os autores agradecem a Coordenaçãoo de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico, Tecnológico do Maranhão (FAPEMA) e a Universidade Federal do Maranhão (UFMA) pela ajuda financeira.

Referências

Aerts, H., Rios Velazquez, E., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., and Lambin, P. (2015). Data from nsclc-radiomics. the cancer imaging archive.

Azuaje, F. (2006). Witten ih, frank e: Data mining: Practical machine learning tools and techniques 2nd edition.

Cook, R. J. (1998). Kappa. *Encyclopedia of biostatistics*.

Coroller, T. P., Agrawal, V., Narayan, V., Hou, Y., Grossmann, P., Lee, S. W., Mak, R. H., and Aerts, H. J. (2016). Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiotherapy and Oncology*, 119(3):480–486.

da Silva Sousa, J. R. F., Silva, A. C., and De Paiva, A. C. (2007). Lung structure classification using 3d geometric measurements and svm. In *Iberoamerican Congress on Pattern Recognition*, pages 783–792. Springer.

de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., and Gattass, M. (2017). Computer-aided diagnosis of lung nodules in computed tomography by using phylogenetic diversity, genetic algorithm, and svm. *Journal of digital imaging*, 30(6):812–822.

de Oliveira, F. S. S., de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., and Gattass, M. (2015). Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and svm. *Computers in biology and medicine*, 57:42–53.

Dean, J. (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. John Wiley & Sons.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5).

Giger, M. L., Huo, Z., Kupinski, M. A., and Vyborny, C. J. (2000). Computer-aided diagnosis in mammography. *Handbook of medical imaging*, 2:915–1004.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Huynh, E., Coroller, T. P., Narayan, V., Agrawal, V., Hou, Y., Romano, J., Franco, I., Mak, R. H., and Aerts, H. J. (2016). Ct-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiotherapy and Oncology*, 120(2):258–266.

Keith, M., Chimimba, C., Reyers, B., and Van Jaarsveld, A. (2005). Taxonomic and phylogenetic distinctiveness in regional conservation assessments: a case study based on extant south african chiroptera and carnivora. In *Animal Conservation forum*, volume 8, pages 279–288. Cambridge University Press.

Oliveira, M. C., de Lucena, D. J. F., and Felix, A. (2017). Recuperação de nódulos pulmonares por conteúdo: uma abordagem radiomics em pesquisa reprodutível.

Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S. J., Wille, M. M. W., Naqibullah, M., Sánchez, C. I., and van Ginneken, B. (2016). Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169.

Shen, C., Liu, Z., Guan, M., Song, J., Lian, Y., Wang, S., Tang, Z., Dong, D., Kong, L., Wang, M., et al. (2017). 2d and 3d ct radiomics features prognostic performance comparison in non-small cell lung cancer. *Translational oncology*, 10(6):886–894.

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855. ACM.

van Timmeren, J. E., Leijenaar, R. T., van Elmpt, W., Reymen, B., Oberije, C., Monshouwer, R., Bussink, J., Brink, C., Hansen, O., and Lambin, P. (2017). Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam ct images. *Radiotherapy and Oncology*, 123(3):363–369.

Vane-Wright, R. I., Humphries, C. J., and Williams, P. H. (1991). What to protect?—systematics and the agony of choice. *Biological conservation*, 55(3):235–254.