

**UNIVERSIDADE FEDERAL DO PIAUÍ - UFPI
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO**

**DESENVOLVIMENTO DE DESCRITORES DE FORMA PARA A REDUÇÃO
DE FALSOS POSITIVOS NA DETECÇÃO AUTOMÁTICA DE NÓDULOS
PULMONARES**

Alexandre Ribeiro Cajazeira Ramos

**PICOS – PIAUÍ
2017**

ALEXANDRE RIBEIRO CAJAZEIRA RAMOS

**DESENVOLVIMENTO DE DESCRITORES DE FORMA PARA A REDUÇÃO
DE FALSOS POSITIVOS NA DETECÇÃO AUTOMÁTICA DE NÓDULOS
PULMONARES**

Monografia submetida ao Curso de Bacharelado de Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Antonio Oseas de Carvalho Filho

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

R175d Ramos, Alexandre Ribeiro Cajazeira.

Desenvolvimento de descritores de forma para a redução de falsos positivos na detecção automática de nódulos pulmonares / Alexandre Ribeiro Cajazeira Ramos.– 2017.

CD-ROM : il.; 4 ¾ pol. (35 f.)

Trabalho de Conclusão de Curso (Curso Bacharelado em Sistemas de Informação) – Universidade Federal do Piauí, Picos, 2017.

Orientador(A): Prof. Dr. Antônio Oseas de Carvalho Filho

1. Câncer de Pulmão-Análise de Forma. 2.Câncer-Tecnologia 3.Nódulos Pulmonares-Falso Positivo. I. Título.

CDD 005

DESENVOLVIMENTO DE DESCRITORES DE FORMA PARA A REDUÇÃO DE
FALSOS POSITIVOS NA DETECÇÃO AUTOMÁTICA DE NÓDULOS PULMONARES

ALEXANDRE RIBEIRO CAJAZEIRA RAMOS


Monografia aprovada como exigência parcial para obtenção do grau de
Bacharel em Sistemas de Informação.

Data de Aprovação

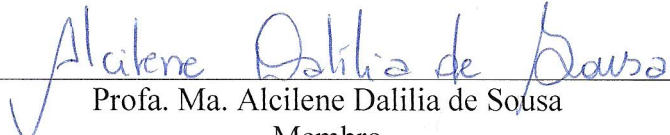
Picos – PI, 11 de Janeiro de 20 17



Prof. Dr. Antonio Oseas de Carvalho Filho
Orientador



Profa. Ma. Patrícia Medyna Lauritzen de Lucena Drumond
Membro



Profa. Ma. Alcilene Dalilia de Sousa
Membro

Aos meus pais, amigos
e amores.

AGRADECIMENTOS

Ao Prof. Antonio Oseas, pelo incentivo, simpatia e presteza no auxílio às atividades e discussões sobre o andamento deste trabalho.

Especialmente às professoras Alcilene, Patricia e Ivenilton, pelos conhecimentos compartilhados e suporte no desenvolvimento das atividades acadêmicas.

À comunidade acadêmica da UFPI campus de Picos, representada pela Prof. Alvenir Barros. Especialmente aos colegas que em todos os fóruns deliberativos pleiteados por mim, confiaram o voto e empenho na luta por uma universidade mais justa e democrática.

A todos os professores do curso de Sistemas de Informação pelo carinho, dedicação e entusiasmo demonstrado ao longo do curso.

“A utopia está no horizonte.

Eu sei muito bem que nunca a alcançarei.

Se eu avançar dez passos, ela se afasta dez passos.

Quanto mais eu a buscar, menos a encontrarei: ela se distancia à medida que eu me aproximo.

E, ora, para que serve a utopia?

A utopia serve para isso, para caminhar.”

Fernando Birri

RESUMO

O impacto causado pela detecção tardia de casos de câncer é devastador. Entre os diagnósticos de tumores malignos mais comuns e letais, destaca-se o câncer de pulmão, que se tornou, também, uma das principais causas evitáveis de morte. O objetivo deste trabalho é desenvolver uma metodologia para a redução de falsos positivos, através da extração de características de forma relacionadas à compacidade, desproporção esférica, diâmetros de Feret, esqueleto e medidas estatísticas. Assim, irá compor uma etapa muito importante em um sistema CAD, auxiliando na distinção consistente entre nódulos pulmonares e outras estruturas presentes em exames de tomografia computadorizada. Foram utilizadas 24156 imagens pertencentes a 833 exames da base de imagens pública LIDC-IDRI e um conjunto de classificadores para testes, com destaque ao algoritmo MLP. Através desta organização, o método proposto alcançou 91% de sensibilidade, 92,8% de especificidade e 92,3% de acurácia. Os resultados mostraram que a análise morfológica é satisfatória e pode ter contribuições significativas para os sistemas CAD e, conseqüentemente, para a vida das pessoas que enfrentam esses problemas.

Palavras-chave: Câncer de Pulmão; Análise de forma; Redução de falso positivos.

ABSTRACT

The impact of late detection of cancer cases is devastating. Among the diagnoses of malignant tumors most common and lethal lung cancer stands out, which has also become one of the leading avoidable causes of death. The objective of this paper is to develop a methodology to reduce false positives, through the extraction features with shape analysis related to compactness, spherical disproportion, Feret diameters, skeleton and statistical measures. Thus will compose a very important step in a CAD system, support in the consistent distinction between nodes and other structures present in the CT scans. Using 24156 images belonging to 833 exams from LIDC-IDRI base images and performing tests with a set of classifiers, highlights the MLP algorithm. Through this joint, the proposed method achieved 91% sensitivity, 92.8% specificity and 92.3% of accuracy. The results showed that the morphological analysis is satisfactory and may have significant contributions to the CAD systems and consequently the lives of people facing these problems. For performing well in the characterization of lung nodules peculiarities and other tissues found in the CT scans.

Palavras-chave: Lung cancer; Shape analysis; False positive reduction.

LISTA DE FIGURAS E TABELAS

Figura 1 - Etapas de um Sistema PDI.....	16
Figura 2 – Nódulo Pulmonar.....	20
Figura 3 – Fluxograma da metodologia proposta.....	23
Figura 4 - Exemplos de candidatos a nódulo pulmonar.....	24
Figura 5 - Comparação entre objeto (a) Compacto e (b) não compacto.....	25
Figura 6 - Ilustração de diâmetro de Feret.....	26
Figura 7 - Ilustração do esqueleto de um retângulo, onde os pontos “A” e “B” peretencem ao esqueleto e “C” não.....	26
Tabela 1 - Fórmulas obtidas através da relação entre descritores.....	27
Tabela 2 - Resultados utilizando múltiplos classificadores.....	31
Tabela 3 - Resultados após a seleção de atributos.....	32
Tabela 4 - Resultados comparativo entre métodos.....	32

LISTA DE ABREVIATURAS E SIGLAS

INCA	Instituto Nacional de Câncer
CAD	<i>Computer-Aided Detection</i>
TC	Tomografia computadorizada
PDI	Processamento Digital de Imagens
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i>
ITK	<i>Insight Segmentation and Registration Toolkit</i>
VTK	Visualization Toolkit
NCI	<i>National Câncer Institute of EUA</i>
MVS	Máquina de vetor de suporte
LIDC	Lung Image Database Consortium
IDRI	Image Database Resource Initiative
PCA	<i>Principal Component Analysis</i>
WEKA	<i>Waikato Environment for Knowledge Analysis</i>
ROC	<i>Receiver-Operating Characteristics</i>
MLP	<i>Multilayer Perceptrom</i>
RF	<i>Randon Forest</i>
SL	<i>Simple Logistic</i>

SUMÁRIO

1 INTRODUÇÃO	13
1.1 OBJETIVO	13
1.2 ORGANIZAÇÃO DO TRABALHO	14
2 FUNDAMENTAÇÃO TEÓRICA	15
2.1 CONCEITOS BÁSICOS	15
2.1.1 Processamento Digital de Imagens (PDI)	15
2.1.2 Informática Médica	17
2.2.3 <i>Insight Segmentation and Registration Toolkit - ITK</i>	18
2.1.4 Câncer de Pulmão	19
2.2 TRABALHOS RELACIONADOS	21
3 METODOLOGIA	23
3.1 AQUISIÇÃO DE IMAGENS	23
3.2 EXTRAÇÃO DE CARACTERÍSTICAS	24
3.2.1 Desproporção Esférica	24
3.2.2 Compacidade	25
3.2.3 Diâmetros de Feret	25
3.2.4 Esqueleto	26
3.2.5 Medidas Estatísticas	27
3.2.6 Descritores Propostos	27
3.3 SELEÇÃO DE ATRIBUTOS	28
3.4 CLASSIFICAÇÃO	29
3.5 VALIDAÇÃO DE RESULTADOS	29
4 RESULTADOS E DISCUSSÃO	31
5 CONSIDERAÇÕES FINAIS	34
6 REFERÊNCIAS	35

1 INTRODUÇÃO

O impacto causado pela detecção tardia de casos de câncer é devastador. Entre os diagnósticos de tumores malignos mais comuns e letais, destaca-se o câncer de pulmão, que ao final do século XX se tornou uma das principais causas de mortes evitáveis, com sobrevida média cumulativa de cinco anos variando entre 13 e 21% em países desenvolvidos e entre 7 e 10% nos países em desenvolvimento (INCA, 2016).

A previsão de aumento anual em sua incidência mundial é de 2%, sendo que a última estimativa apontou para 1,92 milhão de novos casos de câncer de pulmão para o ano de 2012. No Brasil, a previsão do Instituto Nacional de Câncer - INCA é de 28.220 novos casos, sendo 17.330 homens e 10.890 mulheres, com um número de mortes de 24.490, sendo 14.811 homens e 9.675 mulheres (INCA, 2016).

A partir da problemática gerada por estas doenças em nível mundial, uma gama de técnicas que auxiliem a detecção e diagnóstico vem sendo estudadas e implementadas nas mais diversas áreas do conhecimento. Na literatura podem-se encontrar os *Computer-Aided Detection* (CAD), ferramentas que associam as informações peculiares de patologias em soluções computacionais, utilizando técnicas relacionadas ao processamento digital de imagens, inteligência artificial, aprendizado de máquina, dentre outras, com respostas significativas às demandas em questão.

A detecção antecipada de lesões que podem se tornar câncer tem uma contribuição substancial na sobrevivência dos pacientes, os sistemas CAD, neste contexto, associam informações de forma e textura destas lesões para que se possam diferenciar nódulos de outros tecidos encontrados nos exames de tomografia computadorizada (TC), como vasos sanguíneos por exemplo.

1.1 OBJETIVOS

O objetivo deste trabalho consiste em apresentar o desenvolvimento de uma metodologia para redução de falsos positivos, através da extração de características de forma relacionadas a compacidade, desproporção esférica, diâmetros de Feret, esqueleto e medidas estatísticas. Dessa maneira apresenta a capacidade de diferenciação entre classes exclusiva da extração de características de forma e irá

compor uma etapa de grande relevância em um sistema CAD, auxiliando na distinção consistente entre nódulos e outras estruturas presentes nos exames de TC.

1.2 ORGANIZAÇÃO DO TRABALHO

O trabalho está dividido da seguinte maneira: A Seção 2 apresenta a fundamentação teórica e trabalhos relacionados ao projeto; na Seção 3, é descrita a metodologia proposta; na Seção 4, são apresentados os resultados finais da execução da metodologia proposta; e por fim, são elencadas as conclusões e trabalhos finais na Seção 5.

2 FUNDAMENTAÇÃO TEÓRICA

Este trabalho constitui uma etapa de um sistema de auxílio ao diagnóstico, ou sistema CAD, para a detecção automática de nódulos pulmonares em exames de TC, neste sentido se faz necessária uma breve abordagem dos conceitos básicos relacionados à doença e das técnicas de processamento digital de imagens empregadas. Essa seção apresenta a fundamentação teórica deste trabalho e, para tanto, explora os conceitos de câncer, nódulo pulmonar, sistemas de auxílio à detecção, processamento digital de imagens e descritores de forma.

2.1 CONCEITOS BÁSICOS

2.1.1 Processamento Digital de Imagens (PDI)

Entende-se por Processamento Digital de Imagens (PDI) como a manipulação de uma imagem por computador, de modo que a entrada e a saída desse processo seja uma imagem. O objetivo de se usar PDI é melhorar o aspecto visual de certas feições estruturais para o analista humano e fornecer outros subsídios para a sua interpretação, inclusive gerando produtos que possam ser posteriormente submetidos a outros processamentos (SPRING, 1996).

De acordo com Silva (2001), a principal função do processamento digital de imagens é fornecer ferramentas para facilitar a identificação e a extração de informações contidas nas imagens, para posterior interpretação. Nesse sentido, sistemas computacionais são utilizados para atividades interativas de análise e manipulação das imagens. O resultado desse processo é a produção de outras imagens, estas já contendo informações específicas, extraídas e realçadas a partir das imagens originais.

O processamento da imagem não é uma tarefa simples, pois é feito em várias etapas interconectadas. Algumas etapas do processamento digital de imagens são Aquisição, Pré-processamento, Segmentação, Extração de Características e, Reconhecimento e Interpretação. A Figura 1 ilustra um diagrama para um sistema de PDI.

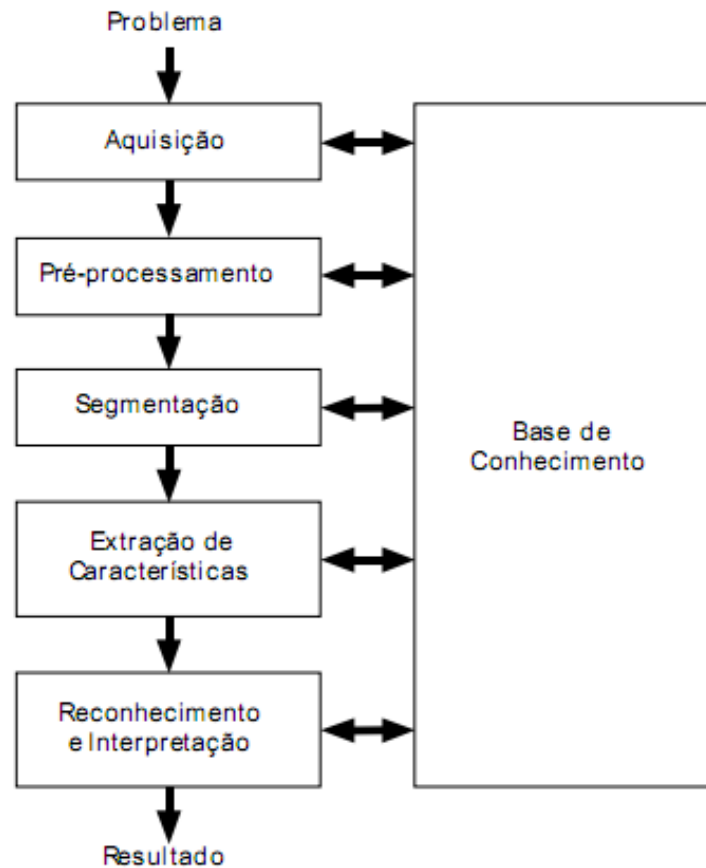


Figura 1 - Etapas de um Sistema PDI. Fonte: (FILHO; NETO, 1999)

Na primeira etapa, tem-se a aquisição de imagens. É nessa etapa que as imagens são capturadas e representadas de forma computacional para serem interpretadas na etapa seguinte.

No pré-processamento, são realizados procedimentos capazes de proporcionar um melhoramento nos aspectos visuais e estruturais da imagem. Dessa maneira, consegue-se aumentar contraste, eliminar ruídos, etc.

A etapa de segmentação consiste em dividir a imagem em objeto (s) e fundo. Em outras palavras, essa etapa consiste em técnicas que de alguma maneira consigam formar padrões de agrupamento, gerando sub-regiões que possuem entre si alguma similaridade.

A representação e descrição, conhecida também por extração de características, tem como objetivo representar, através de valores, uma imagem ou partes dela. Estes valores são características fundamentais que representam propriedades contidas nas imagens.

Entende-se por base de conhecimento, toda e qualquer informação útil para o processamento como um todo, sendo esta composta em geral por imagens, informações de especialistas, dentre outros. Essas informações são úteis pois são usadas em todas as etapas, com destaque para a etapa do reconhecimento e interpretação.

E por fim, tem-se a etapa de reconhecimento de padrões. Nessa etapa, os valores obtidos na etapa de extração de características são os insumos para que uma técnica de aprendizado de máquina possa, então, discernir entre possíveis padrões contidos em um grupo de imagens.

2.1.2 Informática Médica

A partir da problemática gerada por diversas doenças em nível mundial, uma gama de técnicas que auxiliam profissionais de saúde tem sido estudadas e implementadas nas mais diversas áreas do conhecimento. Na computação, podemos pontuar os sistemas de auxílio ao especialista, que fornecem informações complementares de exames médicos e são capazes de contribuir significativamente às interpretações e tomadas de decisão de alto risco. Para construção destes sistemas são associadas técnicas relacionadas a Processamento Digital de Imagens, Inteligência Artificial, Aprendizado de Máquina, dentre outras áreas de pesquisa, apresentando respostas significativas às demandas apresentadas.

A Informática médica compreende um conjunto de soluções computacionais capazes de auxiliar no trabalho dos especialistas de saúde e na relação dos pacientes com diversas doenças. Esta área de conhecimento perpassa desde os sistemas de auxílio ao diagnóstico à aplicativos que realizam a prevenção ou acompanhamento de diversas doenças.

Os sistemas especialistas, responsáveis pelo auxílio à detecção e/ou diagnóstico, associam informações peculiares de doenças em soluções computacionais capazes de enriquecer e auxiliar o trabalho de especialistas na interpretação de exames e emissão de parecer. Para tanto, esses sistemas aglomeram técnicas relacionadas ao processamento digital de imagens, inteligência artificial e aprendizado de máquina.

Através destes sistemas podemos observar uma contribuição significativa, dentre elas, podemos pontuar:

Em seu trabalho, Sampaio (2015) propôs uma metodologia para o diagnóstico automático de câncer de mama utilizando técnicas de melhoramento, segmentação, extração de característica de textura e forma, seleção de características e classificação. Para tanto, ele explorou as técnicas de algoritmos genéticos, máquina de vetor de suporte, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), dentre outras. Os melhores resultados obtidos produziram uma sensibilidade de 94,02%, especificidade de 82,28% e acurácia de 84,08%, com uma taxa de 0,85 falsos positivos por imagem com uma área sob a *Free-Response ROC Curve* (curva FROC) de 1,13 nas análises de mamas não densas. Para mamas densas, obteve-se uma sensibilidade de 89,13%, especificidade de 88,61% e acurácia de 88,69%, com uma taxa de 0,71 falsos positivos por imagem com uma área sob a curva ROC de 1,47.

Já Claro *et al.* (2015), desenvolveram um método para detecção automática de Glaucoma, que é a segunda principal causa de cegueira no mundo e não possui cura. A metodologia utilizada no estudo foi: aquisição de imagem, pré-processamento nas imagens da retina, extração de características de cor e entropia na área alvo e logo após a seleção de atributos. Os melhores resultados produziram uma sensibilidade de 93,7%, especificidade de 93,6%, acurácia de 93,67%, *F-measure* de 93.6 e 0.83 no índice *Kappa*.

Os exemplos de aplicações da informática médica apresentadas utilizaram, dentre os recursos, a biblioteca pública *Insight Segmentation and Registration Toolkit* - ITK, que compreende um conjunto de técnicas e ferramentas de manipulação e análise de imagens médicas. O ITK pode contribuir na construção de sistemas completos de auxílio ao especialista, esmiuçar e compreender técnicas específicas de manipulação de imagens, ou construir novas abordagens e metodologias que contribuam para o contexto da informática médica.

2.1.3 *Insight Segmentation and Registration Toolkit* – ITK

A biblioteca livre *Insight Segmentation and Registration Toolkit* (ITK) é um sistema multi-plataforma de código aberto que fornece aos desenvolvedores um

extenso conjunto de ferramentas de *software* para processamento de imagens. Desenvolvido através de metodologias ágeis, o ITK emprega algoritmos de ponta para registro e segmentação de dados multidimensionais (ITK, 2016).

O ITK possui algoritmos e estruturas para executar segmentação e registro de imagens. Segmentação é definido como o processo de identificar e classificar os elementos encontrados em uma imagem digital. Já o registro é a tarefa de alinhar a imagem, por meio de algum tipo de critério de correspondência entre dados. Essa biblioteca foca em imagens médicas, embora não haja restrições quanto ao processamento de outros tipos de dados. O ITK não realiza visualização de imagens, deixando a cargo de outras bibliotecas, como o VTK (*Visualization Toolkit*).

A biblioteca ITK é multi-plataforma implementada na linguagem C++, mas possui uma interface que permite a utilização através de outras linguagens como Java e Python. Por possuir programação genérica, o ITK apresenta flexibilidade em suas aplicações, muitos *templates* em C++ já são disponíveis.

Outra característica importante do ITK é ser um *software* livre, desta forma, ele é continuamente desenvolvido por colaboradores de todo o mundo, através do modelo de desenvolvimento *extreme programming*, o que consiste em um rápido, dinâmico e constante ciclo de projeto, programação, teste e lançamento. As atualizações são realizadas diariamente, assim essa biblioteca se mantém em constante evolução.

2.1.4 Câncer de Pulmão

Entende-se por câncer um conjunto de doenças com um aspecto em comum: o crescimento desordenado de células que invadem tecidos e órgãos causando inúmeros danos. Trata-se de um processo agressivo e incontrolável, determinando a formação de tumores malignos que poderão espalhar-se para outras regiões (NCI, 2016).

Os fatores que levam a essa condição são variados, estando comumente associados a variáveis ambientais, hábitos de vida e contato com substâncias com potencial cancerígeno, a exemplo do que ocorre por ocasião de contaminações, hábito do tabagismo ou ingestão de agrotóxicos (SILVA *et al.*, 2009).

“O câncer de pulmão é o mais comum de todos os tumores malignos, apresentando aumento de 2% por ano na sua incidência mundial. A última

estimativa mundial apontou incidência de 1,82 milhão de casos novos de câncer de pulmão para o ano de 2012, sendo 1,24 milhão em homens e 583 mil em mulheres. Em 90% dos casos diagnosticados, o câncer de pulmão está associado ao consumo de derivados de tabaco. No Brasil, foi responsável por 22.424 mortes em 2011” (INCA, 2016).

O câncer de pulmão é caracterizado pelo crescimento desordenado de células que constituem nódulos malignos na região do parênquima pulmonar. Um nódulo é uma pequena massa de tecido que se forma no corpo, normalmente em resposta a lesões, mas que geralmente é benigno e não requer atenção médica especial. Todavia quando estes tecidos podem interferir em funções básicas do corpo, ou se tornarem malignos e constituírem câncer, seu impacto pode ser devastador e significativamente grave para a vida dos pacientes (SILVA *et al.*, 2009). A Figura 2 ilustra um nódulo encontrado na região do parênquima pulmonar.

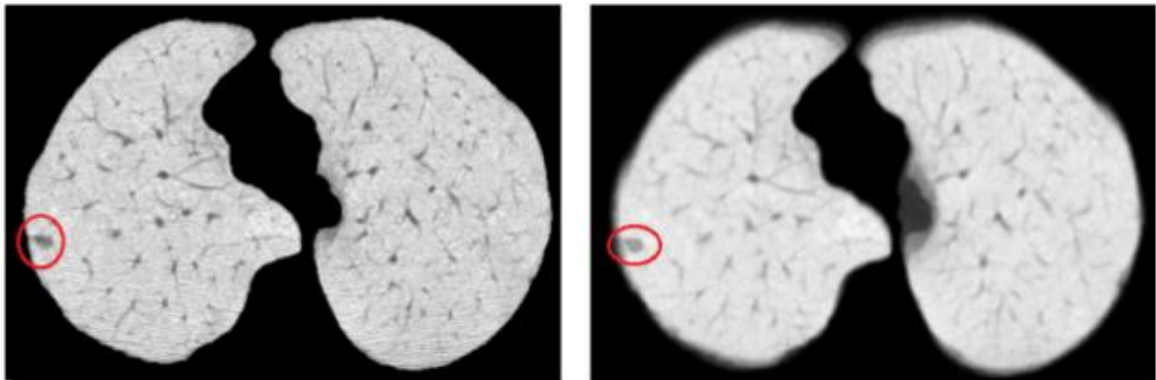


Figura 2 – Nódulo Pulmonar (CARVALHO FILHO *et al.*, 2013).

O diagnóstico deste tipo de câncer é feito por especialistas através da interpretação de uma TC do tórax. Entretanto, como a análise dos componentes destes exames deve ser exaustiva e as implicações que os diagnósticos podem gerar são de alto risco, cresce o número de estudos que objetivam contribuir com o trabalho destes especialistas, reduzindo erros que neste contexto são inaceitáveis.

Estes sistemas especialistas, quando aplicados aos nódulos pulmonares, englobam as etapas de segmentação dos candidatos a nódulo e das regiões de interesse, extração de características, redução de falsos positivos, classificação dos candidatos a nódulo e classificação entre nódulos malignos e benignos.

Para a etapa de redução de falsos positivos em um sistema de auxílio à detecção, podem-se utilizar recursos do processamento digital de imagens, como a análise morfológica, ou extração de características de forma, dos candidatos a nódulo pulmonary, pois um dos critérios utilizados pelos especialistas na

diferenciação entre estas estruturas é a forma que eles assumem, como vasos sanguíneos que são alongados, enquanto os nódulos são arredondados.

2.2 TRABALHOS RELACIONADOS

Na literatura encontram-se trabalhos relacionados a esta pesquisa, pode-se destacar:

Carvalho Filho *et al.* (2016) propuseram uma metodologia para a classificação de nódulos pulmonares usando a imagem LIDC-IDRI (*Lung Image Database Consortium - Image Database Resource Initiative*). Utilizaram os índices de diversidade taxonômica e taxonômica distinção da ecologia para descrever a textura do nódulo e não-nódulo. O cálculo destes índices é baseado em árvores filogenética e são aplicados na caracterização de candidatos a nódulo. Para classificação foi utilizada a máquina de vetor de suporte - MVS. Para aplicar a metodologia, dividiram o banco de dados completo em dois grupos para treinamento e testes. Na fase de testes, foram utilizados 833 exames, obtendo uma precisão média de 98,11%.

Santos *et al.* (2014) desenvolveram uma metodologia para detecção automática de pequenos nódulos solitários (de tamanhos entre 2 e 10 mm). Utilizaram métricas relacionadas à entropia e Shannon Tsallis para seus descritores de textura e MVS para a classificação de regiões de interesse - nódulos ou não-nódulos. Os resultados obtidos por esta metodologia, aplicada em uma amostra com 28 exames da base de imagens LIDC-IDRI, mostrou que pequenos nódulos foram detectados com uma sensibilidade de 90,6%, uma especificidade de 85%, precisão de 88,4% e taxa de 1,17 falsos positivos por exame.

Fernandes *et al.* (2016) apresentam uma aplicação de distribuições de forma para diagnosticar nódulos pulmonares em imagens de tomografia computadorizada. O estudo utiliza a superfície do nódulo e superfícies internas geradas por atributos comuns do seu voxel e distribuição. As superfícies são capturadas usando o algoritmo 3D Alpha Shapes. Todas as superfícies são caracterizadas utilizando distribuições forma D1, D2, D3, D4 e A3, a fim de representar o comportamento da forma com base na distribuição estatística dos seus pontos de contorno 3D. Utilizando MVS para classificação e a combinação de todos os descritores, demonstrou resultados superiores a 90% de precisão, indicando um recurso

promissor para distinção entre nódulos malignos e benignos, contribuindo para o diagnóstico do câncer de pulmão.

Carvalho Filho *et al.* (2013) propuseram uma metodologia para detecção automática de nódulos pulmonares, compreendendo sistema CAD completo, composto por três etapas: 1) extração e classificação de nódulos solitários com base na tomografia computadorizada e a reconstrução do parênquima pulmonar, que destaca as estruturas dos nódulos; 2) Segmentação de candidatos de nódulos; e 3) extração de características de forma e textura para classificação através da MVS. A metodologia obteve resultados com uma sensibilidade de 85,91%, uma especificidade de 97,70% e uma precisão de 97,55%.

A metodologia proposta por Fernandes *et al.* (2016) obteve menores resultados nas métricas de avaliação de sensibilidade, especificidade e acurácia. Tendo em vista que dentre os trabalhos analisados este foi o único a explorar apenas características de forma dos nódulos pulmonares solitários, este trabalho objetiva aprimorar seus resultados para sustentar a utilização e contribuição dos descritores. Para tanto, propõe-se a utilização de outras características relacionadas a análise de forma, como obtenção do esqueleto e os cálculos da desproporção esférica, diâmetros de Feret, compacidade e medidas estatísticas.

3 METODOLOGIA

A metodologia empregada para o desenvolvimento deste trabalho engloba a aquisição da base de imagens a ser utilizada, extração de características através da implementação dos descritores de forma para nódulos pulmonares, seleção de características, classificação entre nódulos ou não nódulos e validação dos resultados. A Figura 3 ilustra os processos sequenciais da metodologia.

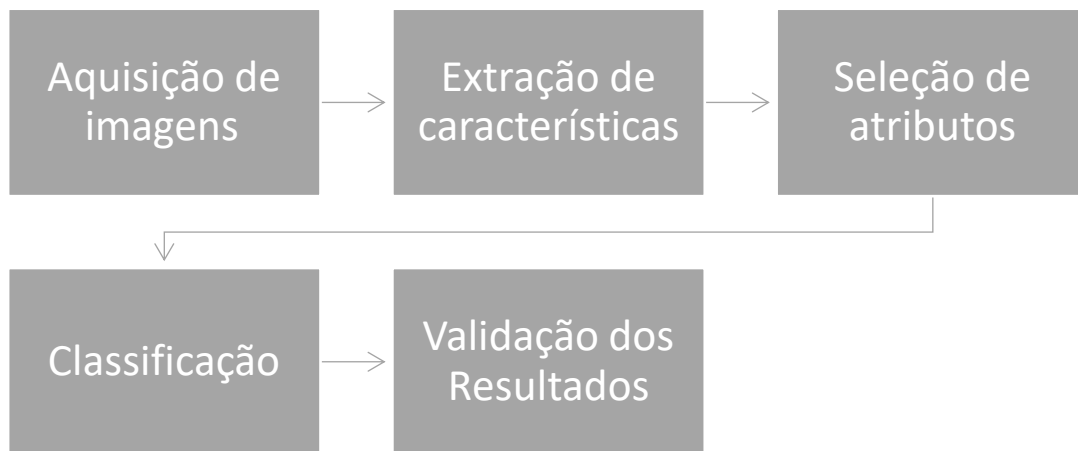


Figura 3 – Fluxograma da metodologia proposta.

3.1 AQUISIÇÃO DE IMAGENS

Foi utilizada neste trabalho a Base de imagens pública LIDC-IDRI, disponibilizada pelo *National Câncer Institute* dos EUA (NCI), resultado da junção das bases *Lung Image Database Consortium* (LIDC) e o *Image Database Resource Initiative* (IDRI). Estas imagens foram utilizadas para o treinamento e validação do método proposto (LIDC - IDRI, 2015).

Esta base é composta por 1012 casos, a partir das informações e marcações da base LIDC. Cada exame da base possui informações e avaliações de 4 especialistas, que são disponibilizadas em um arquivo XML (LIDC - IDRI, 2015).

Para entrada na etapa de extração de características foram utilizados os resultados da etapa de segmentação de candidatos a nódulo, com esta base de imagens, proposto por Carvalho Filho *et al.* (2016). Tendo em vista que: 1 – A base de imagens LIDC-IDRI é composta por um conjunto de TC de todo o parênquima pulmonar; 2 – Que se faz necessário uma grande quantidade de exemplos de

nódulos e não nódulos solitários para testes e validação da metodologia proposta; 3 – Que o processo de segmentação, responsável por separar tecidos do parênquima pulmonar, é uma etapa complexa e não se pode fazer manualmente; 4 – Que os resultados da segmentação de candidatos a nódulos proposto pelo trabalho de Carvalho Filho *et al.* (2016) é consistente quando comparado às anotações dos 4 especialistas da LIDC-IDRI. A Figura 4 apresenta exemplos de candidatos a nódulo segmentados por Carvalho Filho *et al.* (2016) demonstrando como essas estruturas são de difícil diferenciação.

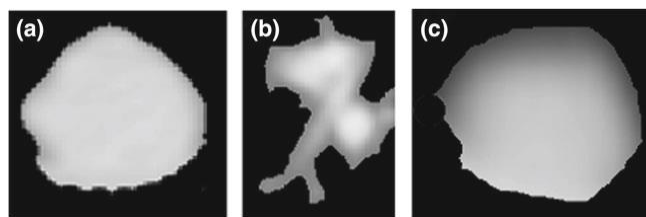


Figura 4 - Exemplos de candidatos a nódulo pulmonar. (a) representa um nódulo, (b) e (c) outro tecido encontrado. Fonte: Adaptado de (CARVALHO FILHO, 2016).

3.2 EXTRAÇÃO DE CARACTERÍSTICAS

A análise de forma utilizada neste trabalho foi motivada pela análise das peculiaridades morfológicas dos nódulos, estas técnicas auxiliam a diferenciação entre nódulos e outras estruturas, bem como entre nódulos malignos e benignos.

Dentre os descritores estudados destacam-se: 1) Desproporção Esférica; 2) Compacidade; 3) Diâmetros de Feret; 4) Esqueleto; 5) Medidas estatísticas relacionadas às distâncias da borda ao centro de Massa do Objeto; Constituindo assim: 6) O conjunto de índices calculados, de modo que através deles possamos submeter informações morfológicas consistentes aos classificadores.

3.2.1 Desproporção Esférica

A desproporção esférica mede o quão irregular é uma superfície em relação a uma superfície perfeitamente esférica, de modo que se possa diferenciar os nódulos pulmonares dos vasos sanguíneos ou outros tecidos, tendo em vista que nódulos tendem a ser esféricos, enquanto vasos tendem a ser alongados (CARVALHO FILHO *et al.*, 2013).

Pode-se mensurar a desproporção esférica de objetos a partir das Equações 1 e 2.

$$DespEsf = \frac{A}{4\pi R^2} \quad (1)$$

$$R = \sqrt[3]{\frac{3V}{4\pi}} \quad (2)$$

onde “A” corresponde à área do objeto e “R” é o raio estimado, medido através da equação 2, em que “V” corresponde ao volume do objeto.

3.2.2 Compacidade

A compacidade representa a densidade de objeto em relação a um objeto perfeitamente denso, como um círculo. Podemos calcular através da Equação 3. A Figura 5 demonstra comparação entre objetos através da compacidade (SAMPAIO, 2015).

$$Comp = \frac{p^2}{4\pi A} \quad (3)$$

onde “p” é o perímetro do objeto e “A” a área.

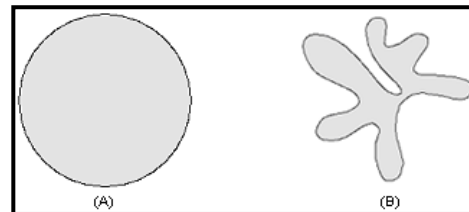


Figura 5 - Comparação entre objeto (a) Compacto e (b) não compacto. Fonte: Adaptado de (SAMPAIO, 2015).

3.2.3 Diâmetros de Feret

Os diâmetros de Feret mensuram as distâncias entre retas opostas que tocam a borda do objeto, demonstrando o seu comprimento em relação a uma direção específica. Pode-se observar estes parâmetros na Figura 6, que analisa um objeto bidimensional. Todas estas medidas são usadas como descritores e contribuem para a caracterização dos nódulos pulmonares (SAMPAIO, 2015).

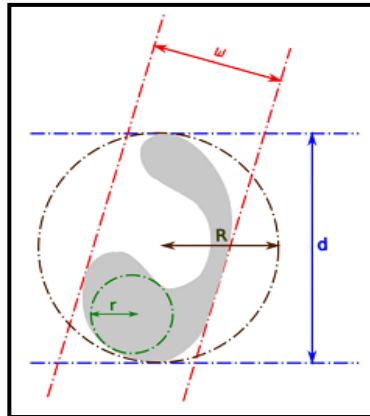


Figura 6 - Ilustração de diâmetro de Feret. Fonte: adaptado de (SAMPAIO, 2015).

onde “w” representa o diâmetro mínimo de Feret, “d”, o diâmetro máximo de Feret, “R”, raio da menor circunferência externa e “r”, raio da maior circunferência interna do objeto.

3.2.4 Esqueleto

Os esqueletos de uma forma geométrica são constituídos pelas coordenadas que representam os centros dos círculos que tocam pelo menos 2 pontos da borda de um objeto (SAMPAIO, 2015), como demonstrado na Figura 7:

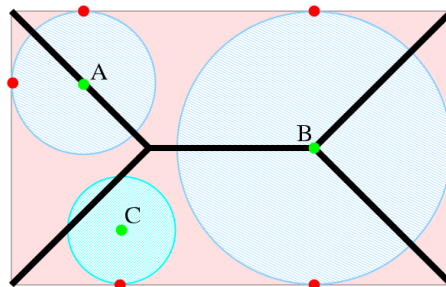


Figura 7 - Ilustração do esqueleto de um retângulo, onde os pontos “A” e “B” peretencem ao esqueleto e “C” não. Fonte: Adaptado de (SAMPAIO, 2015).

Portanto, o esqueleto de um objeto representa as características de sua forma e superfície num conjunto reduzido de pontos de coordenadas, Possibilitando compreender rapidamente o volume do objeto e o comportamento de sua superfície para cada região de seu esqueleto. Pode-se então extrair como característica o tamanho do esqueleto de determinado objeto, como também sua relação com outras características implementadas, como área, Feret máximo, Feret mínimo, dentre outras.

3.2.5 Medidas Estatísticas

Com o intuito de pontuar, com precisão, a variação entre as distancias dos pontos da borda ao centro do objeto, foram cálculos a média e desvio padrão dessas distâncias, que são medidas estatísticas que pontuam o grau de dispersão de valores, ou sua variação (BRITO *et al.*, 2003). Pode-se obter estas medidas através das Equações 4 e 5.

$$M = \frac{\sum_i^n X_i}{n} \quad (4)$$

$$DP = \sqrt{\frac{\sum_i^n (X_i - M)^2}{n}} \quad (5)$$

onde “M” é a média, “DP” o desvio padrão, “n” a quantidade de valores analisados (Quantidade de pontos da borda do objeto), “Xi” representa cada distancia de um ponto da borda ao centro de massa do objeto.

3.2.6 Descritores propostos

Ao relacionar os parâmetros e características de forma estudados neste trabalho, foi implementado um conjunto de descritores 3d que refletem as peculiaridades da forma de um tecido segmentado em um exame de TC, assim é constituído um vetor de características, ou conjunto de informações que dão base à inferência dos classificadores diante de novos objetos.

Além das características extraídas de desproporção esférica, compacidade diâmetros de Feret, tamanho do esqueleto e medidas estatísticas, foram traçadas relações entre estas características, gerando novos descritores que podem ser observados na Tabela 1.

Tabela 1 - Fórmulas obtidas através da relação entre descritores.

FÓRMULAS OBTIDAS ATRAVÉS DA RELAÇÃO ENTRE DESCRITORES	
1	r / R
2	$w / (2R)$
3	$A / 2\pi R^2$
4	$2r/d$
5	w/d

FÓRMULAS OBTIDAS ATRAVÉS DA RELAÇÃO ENTRE DESCRITORES	
6	$4A/\pi d^2$
7	$R^2\sqrt{3}/d$
8	$2\pi r/P$
9	$\pi w/P$
10	$4\pi A/P^2$
11	$2d/P$
12	$4R/P$
13	$\pi r^2/A$
14	$2r/w$
15	d/E
16	w/E

onde A , P , r , R , w , d e E são o volume, área da superfície, raio da maior circunferência interna, raio da menor circunferência externa, diâmetro mínimo de Feret, diâmetro máximo de Feret e tamanho do esqueleto, respectivamente.

3.3 SELEÇÃO DE ATRIBUTOS

Após a extração foi realizada a seleção entre as características que melhor discriminam as classes nódulos e não-nódulo. Para isso, utilizou-se o método de seleção *Principal Component Analysis* - PCA em conjunto à técnica de ranqueamento, com seus parâmetros em valores padrão, ambos disponíveis no *software Waikato Environment for Knowledge Analysis* - WEKA (HALL et al., 2009).

O método PCA tem por objetivo reduzir um conjunto de dados trabalhos, é um método estatístico de múltiplas variáveis de simples interpretação e é considerado uma transformação linear ótima. Além disso, o PCA é utilizado para identificar a relação entre características extraídas em determinado conjunto de dados, tornando-se útil para vetores de características de grandes dimensões (WESTAD et al., 2003).

Os descritores propostos neste trabalho produzem um vetor de características para cada estrutura analisada. O PCA, que foi selecionado pela grande utilização entre a comunidade científica e pela lógica relativamente simples, define os fatores determinantes na diferenciação de classes, propiciadas por características

específicas. Assim, seleciona os melhores descritores que ajudam a distinguir nódulos pulmonares e outras estruturas (WESTAD *et al.*, 2003).

3.4 CLASSIFICAÇÃO

O processo de classificação consiste em submeter as características extraídas dos tecidos encontrados nos exames a algoritmos capazes de aprender e inferir sobre os mesmos, de modo que, ao assimilar as características peculiares às classes estudadas (nódulos e não nódulos) estes algoritmos possam classificar um novo objeto, formando uma nova informação a ser analisada por um especialista ou uma nova condição para redução de falsos positivos em um sistema de auxílio ao diagnóstico.

Os algoritmos aos quais foram submetidos o conjunto de características serão apresentados na listagem de resultados, todos seguiram o método de *k-fold cross validation*, que divide o conjunto de dados em 10, utiliza os 9 primeiros para o treinamento do algoritmo e o último para testes, repetindo este processo 10 vezes ($k=10$) e apresentando o resultado.

Para tanto, foi utilizada a ferramenta *Waikato Environment for Knowledge Analysis* (WEKA), desenvolvida por pesquisadores da Universidade de Waikato, Nova Zelândia que é constituída por um emaranhado de técnicas de aprendizado de máquina e mineração de dados, de modo que disponibiliza ferramentas para pré-processamento de dados, classificação, regressão, agrupamento, seleção de características, dentre outros (HALL *et al.*, 2009).

3.5 VALIDAÇÃO DE RESULTADOS

Os resultados obtidos entre os classificadores são avaliados através das métricas *Kappa* (K) (LANDIS *et al.* 1977), área sob a curva *Receiver-Operating Characteristics* (ROC) (METZ *et al.* 1086), Sensibilidade (S), Especificidade (E) e Acurácia (A) (MARTINEZ *et al.* 2003).

A sensibilidade demonstra quão boa é a metodologia na detecção da doença, neste caso a proporção de não nódulos classificados corretamente. A especificidade demonstra o quão bom o método é na classificação de imagens saudáveis, ou nódulos, neste caso. O *kappa* avalia a concordância entre classificações distintas. E a curva ROC mensura quão boa é a metodologia na distinção entre duas classes,

sendo elas verdadeiros positivos, medidos através da especificidade, e falsos positivos medidos pela sensibilidade.

4 RESULTADOS E DISCUSSÃO

Neste trabalho foram analisadas e submetidas aos classificadores um conjunto de 24.156 imagens pertencentes a 833 exames da base pública LIDC-IDRI, sendo elas 6.414 nódulos e 17.742 não nódulos.

As características extraídas (Seção 3.2) foram submetidas às métricas de avaliação *Kappa* (K), área sob a curva *Receiver-Operating Characteristics* (ROC), Sensibilidade (S), Especificidade (E) e Acurácia (A), a partir dos seguintes classificadores: Máquina de Vetor de Suporte – MVS; *Multilayer Perceptron* (MLP); *Randon Forest* (RF); J48; IBK; e *Simple Logistic* (SL), sendo estes selecionados por sua utilização em trabalhos relacionados e por apresentar resultados satisfatórios em outros contextos.

Pode-se analisar na Tabela 2 os resultados utilizando múltiplos classificadores.

Tabela 2 – RESULTADOS UTILIZANDO MÚLTIPLOS CLASSIFICADORES

Classificador	Resultado comparativo entre classificadores				
	<i>KAPP A</i>	<i>Curva ROC</i>	<i>S (%)</i>	<i>E (%)</i>	<i>A (%)</i>
MVS	0,5873	0,761	85,3	85,8	85,7
MLP	0,8327	0,978	84,9	96,7	93,3
RF	0,8616	0,985	88,5	96,8	94,5
J48	0,8234	0,929	86,2	95,6	93,1
IBK	0,7745	0,891	82,5	94,4	91,1
SL	0,8264	0,975	85,5	96,1	93,1

Os resultados apresentados na Tabela 2 demonstram que o conjunto de classificadores obteve resultados satisfatórios, com todas as medições de S, E e A superiores a 80%. Com destaque ao RF, que obteve o melhor resultado dentro do contexto da classificação entre nódulos e não nódulos. Este classificador, associado aos descritores implementados, obteve 88,5% de Sensibilidade, 96,8% de Especificidade, 94,5% de Acurácia, 0,8616 no índice Kappa e 0,985 de área sobre a curva ROC. Em contrapartida, o MVS obteve o pior resultado, com 85,3% de Sensibilidade, 85,8% de Especificidade, 85,7% de Acurácia, 0,5873 no índice Kappa e 0,761 de área sobre a curva ROC.

Após a seleção de atributos com o PCA, o conjunto de classificadores tiveram seus resultados otimizados, tendo selecionado 9 dentre as 20 características extraídas. A Tabela 3 mostra este resultado.

Tabela 3 – RESULTADOS APÓS A SELEÇÃO DE ATRIBUTOS

Classificadores	Resultados após a seleção de atributos			
	Curva ROC	S (%)	Sp(%)	A(%)
SVM	0.928	90.9	94.7	93.7
MLP	0.970	91.0	92.8	92.3
RF	0.979	88.2	95.1	93.3
J48	0.934	82.2	94.9	91.4
IBK	0.888	84.1	93.1	90.7
SL	0.967	83.7	94.6	91.7

Os resultados apresentados na Tabela 3 mostram que a seleção de atributos otimizou os resultados da metodologia proposta e destacou o SVM e MLP que, após a seleção, obteve resultados superiores a 90% em todas as métricas de avaliação. O melhor resultado foi apresentado pelo MLP. Este classificador, combinado aos descritores e sua seleção, apresentou 91% de sensibilidade, 92,8% de especificidade, 92,3% de exatidão e 0,970 de área sob a curva ROC.

A análise comparativa entre os resultados da metodologia proposta e trabalhos relacionados analisando características de forma e/ou textura para descrição de nódulos pulmonares são apresentados na Tabela 4, sendo que todos os trabalhos utilizaram a mesma base de imagens, mas não se tem acesso aos códigos e especificação dos exames utilizados para uma análise comparativa consistente, sendo esta meramente ilustrativa.

Tabela 4 - RESULTADOS COMPARATIVOS ENTRE MÉTODOS.

	Resultados comparativos entre métodos				
	Base de Imagens	S (%)	Es(%)	A(%)	Nº Exames
Fernandes et. al.	LIDC-IDRI	87.94	94.32	91.05	754
Santos et. al.	LIDC-IDRI	90.6	85	88.4	28
Carvalho Filho et. al.	LIDC-IDRI	85.91	97.7	97.55	833
Metodologia proposta	LIDC-IDRI	91	92.8	92.3	833

A metodologia proposta por Fernandes *et al.* (2016), que utilizaram apenas características da forma de nódulos pulmonares solitários, obteve resultados menores nas métricas de avaliação de sensibilidade, especificidade e exatidão, dentre os trabalhos relacionados. O presente trabalho, por sua vez, obteve resultados significativos em relação a Fernandes *et al.* (2016), trazendo contribuições importantes para o estudo do tema.

Dada a análise comparativa apresentada na Tabela 4, podemos classificar a metodologia proposta como satisfatória, pois traz resultados significativos, com avaliação de métricas balanceadas, além do sucesso quando comparado a outros estudos que também utilizam apenas análise de forma, como Fernandes *et al.* (2016), por exemplo.

5 CONSIDERAÇÕES FINAIS

O câncer é um conjunto de doenças substancialmente perigosas e o impacto da sua detecção tardia pode ser irreversível. O câncer de pulmão se tornou uma das principais causas evitáveis de morte à medida que seu diagnóstico seja realizado o mais breve possível e as medidas cabíveis possam ser tomadas de imediato.

Neste trabalho foi proposta uma metodologia para a redução de falsos positivos em exames de TC, através de descritores de forma que relacionam as características geométricas de área, perímetro, desproporção esférica, diâmetros de Feret e esqueleto, além de medidas estatísticas relacionadas as distancias dos pontos da superfície ao centro de massa de um objeto.

Outro aspecto pertinente deste trabalho é a base de imagens utilizadas, a LIDC-IDRI. Através dela um grande acervo de exames e anotações médicas são capazes de proporcionar a implementação, testes e validação de diversas técnicas relacionadas ao auxílio à detecção e diagnóstico do câncer de pulmão.

Os resultados obtidos demonstraram que a análise morfológica é satisfatória e pode apresentar contribuições significativas aos sistemas CAD e, conseqüentemente, à vida das pessoas que enfrentam estes problemas, pois apresentam bom desempenho na caracterização das peculiaridades de nódulos pulmonares e outros tecidos encontrados nos exames de TC.

Para trabalhos futuros pretende-se testar a capacidade de generalização dos descritores propostos, submetendo-os a outras bases de imagens, outros contextos e doenças ou outras etapas de um sistema de auxílio ao diagnóstico. Além disso, para enriquecer o conjunto de características extraídas, pretende-se acrescentar outros atributos morfológicos. Todos os acréscimos e alterações poderão otimizar os resultados, contribuindo para a constatação da eficácia da metodologia proposta.

6 REFERÊNCIAS

BRITO, NATILENE MESQUITA, et al. Validação de métodos analíticos: estratégia e discussão. **Pesticidas: revista de ecotoxicologia e meio ambiente**, v. 13, p. 129-146. 2003.

CARVALHO FILHO, A.O., DE SAMPAIO, W.B., SILVA, A.C., DE PAIVA, A.C., NUNES, R.A., GATTASS, M.: Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. **Artificial Intelligence in Medicine**, 2013. DOI <http://dx.doi.org/10.1016/j.artmed.2013.11.002>. URL <http://www.sciencedirect.com/science/article/pii/S0933365713001541>

CARVALHO FILHO, A. O.; SILVA, A. C.; PAIVA, A. C.; GATTAS, R. A. N. M. Lung-nodule classification based on computed tomography using taxonomic diversity indexes and an SVM. **Journal of Signal Processing Systems for Signal, Image, and Video Technology**, v. 83, may 2016.

FERNANDES, V. P. M. ; KANEHISA, R. F. A. ; BRAZ JÚNIOR, Geraldo ; SILVA, A. C. ; PAIVA, A. C. . Lung Nodule Classification Based on Shape Distributions. In: **SAC - Symposium on Applied Computing, 2016, Pisa**. SAC '16:Proceedings of the 31th Annual ACM Symposium on Applied Computing, 2016.

HALL, M. FRANK, E. HOLMES, G. PFAHRINGER, B. REUTEMANN, P. WITTEN, I. H. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v. 11, n. 1, p. 10-18, 2009.

Instituto Nacional do Câncer - INCA. Ministério da Saúde. **Câncer de pulmão**. Disponível em: <<http://www2.inca.gov.br/wps/wcm/connect/cancer/site/oquee>>. Acesso em: 23 de maio de 2016.

LANDIS JR, GARY GK. The measurement of observer agreement for categorical data. **Biometrics**; 1977. 159-74.

METZ C. E. ROC methodology in radiologic imaging. **Invest. Radiol**, v. 21, n.9, p. 720-33, 1986.

MARTINEZ E. Z; LOUZADA-NETO, F.; PEREIRA, B. B. A curva ROC para testes diagnósticos. **Cadernos Saúde Coletiva**, v.11, n. 1, p.7-31, 2003.

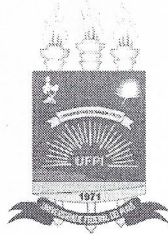
National Cancer Institute - NCI. "**Lung Cancer. U.S. National Institute of Health**". Disponível em: www.cancer.gov/about-cancer/what-is-cancer. Acessado em 20 jun. 2016.

SAMPAIO, W.B. **Deteção de massas em imagens mamográficas usando uma metodologia adaptada à densidade da mama**. Dissertação de Mestrado (Programa de Pós-Graduação em Engenharia de Eletricidade) – Universidade Federal do Maranhão, 2015.

SANTOS, A. M.; CARVALHO FILHO, A. O.; SILVA, A. C.; PAIVA, A. C.; NUNES, R. A.; GATTASS, M. Automatic detection of small lung nodules in 3D CT data using Gaussian mixture models, Tsallis entropy and SVM. **Engineering Applications of Artificial Intelligence**, vol. 36, p. 27–39, 2014.

SILVA, D.R.; BAGLIO, P.T.; GAZZANA, M.B. Nódulo pulmonar solitário. **Rev Bras Clin Med**, v.7; p.132-139, 2009.

WESTAD, F.; HERSLETH, M.; LEA, P.; MARTENS, H. Variable selection in PCA in sensory descriptive and consumer data. **Food Quality and Preference**. v. 14, p. 463–472, 2003.



TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA
“JOSÉ ALBANO DE MACEDO”

Identificação do Tipo de Documento

- () Tese
() Dissertação
(X) Monografia
() Artigo

Eu, Alexandre Ribeiro Cajazeira Ramos,
autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de
02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar,
gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação
“**Desenvolvimento de descritores de forma para redução de falsos positivos na detecção
automática de nódulos pulmonares**” de minha autoria, em formato PDF, para fins de leitura
e/ou impressão, pela internet a título de divulgação da produção científica gerada pela
Universidade.

Picos-PI 27 de Janeiro de 2017.

Alexandre Ramos

Assinatura

Alexandre Ramos

Assinatura