UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
LABORATÓRIO DE INOVAÇÃO TECNOLÓGICA EM SAÚDE

LCD-OpenPACS: Sistema Integrado de Telerradiologia com Auxílio ao Diagnóstico de Nódulos Pulmonares em Exames de Tomografia Computadorizada

José Macêdo Firmino Filho

Orientador: Prof. Dr. Ricardo Alexsandro de Medeiros Valentim

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da UFRN (área de concentração: Engenharia de Computação) como parte dos requisitos para obtenção do título de Doutor em Ciências.

Número de ordem PPgEE: D158
Natal, RN, dezembro de 2015

Seção de Informação e Referência

Catalogação da Publicação na Fonte. UFRN / Biblioteca Central Zila Mamede.

Firmino Filho, José Macêdo.

LCD-OpenPACS: sistema integrado de telerradiologia com auxílio ao diagnóstico de nódulos pulmonares em exames de tomografia computadorizada / José Macêdo Firmino Filho. - Natal, RN, 2015.

77 f.

Orientador: Dr. Ricardo Alexsandro de Medeiros Valentim.

Tese (Doutorado em Engenharia Elétrica e de Computação) - Universidade Federal do Rio Grande do Norte. Centro de Tecnologia - Programa de Pós-Graduação em Engenharia Elétrica e de Computação.

Processamento de Imagens Médica - Tese. 2. Sistema CADe - Tese. 3. Câncer de Pulmão - Tese. 4. LCD-openpacs - Tese. I. Valentim, Ricardo Alexsandro de Medeiros. II. Costa, Flávio Bezerra. III.Título.

RN/UF/BCZM CDU 621:615.849

LCD-OpenPACS: Sistema Integrado de Telerradiologia com Auxílio ao Diagnóstico de Nódulos Pulmonares em Exames de Tomografia Computadorizada

José Macêdo Firmino Filho

Tese de Doutorado APROVADA em 04 de dezembro de 2015 pela banca examinadora composta pelos seguintes membros:

Prof. Dr. Ricardo Alexsandro de Medeiros Valentim (orientador) . DEB/UFRN

Prof. Dr. Aquiles Medeiros Filgueira Burlamaqui ..... ECT/UFRN

Prof[a] Dr[a] Karilany Dantas Coutinho ..... DEB/UFFN

Prof. Dr. Hertz Wilton de Castro Lins ..... DCO/UFRN

Prof. Dr. Jõao Paulo Queiroz dos Santos ..... IFRN

Prof. Dr. Bruno Gomes de Araújo ..... IFRN

Prof. Dr. Robinson Luis de Souza Alves ..... IFRN

AGRADECIMENTOS

Agradeço a Deus, por ter permitido que eu atingisse mais este objetivo;

Ao Prof. Dr. Ricardo Alexsandro de Medeiros Valentim, pela amizade, incentivo, apoio e preocupações sem as quais este trabalho não seria realizado;

Ao professor membros da banca (Dr. Aquiles Medeiros Filgueira Burlamaqui, Profª Drª Karilany Dantas Coutinho, Prof. Dr. Hertz Wilton de Castro Lins, Prof. Dr. João Paulo Queiroz dos Santos, Prof. Dr. Bruno Gomes de Araújo, Prof. Dr. Robinson Luis de Souza Alves) pelo envolvimento na realização deste trabalho, pelas suas inquietações, orientações e contribuições;

Aos colegas do LAIS (sediada na Universidade Federal do Rio Grande do Norte) que todos os dias me incentivaram com coragem, discernimento e amizade, contribuindo, assim, para que este trabalho pudesse ser realizado;

Aos amigos Giovani Angelo, Marcel Ribeiro Dantas e Bruno Marques pelo compartilhamento de experiências, pelas sugestões e apoios, alguns destes essenciais no desenvolvimento deste trabalho.

À minha esposa (Janaide Mª de Andrade Batista) por acreditar e me incentivar na realização deste trabalho e por estar ao meu lado, nas horas mais difíceis. Seu apoio, atenção e compreensividade foram fundamentais.

À minha filha, Luna Batista Firmino, que me proporcionou a base emocional e espiritual necessárias na finalização deste trabalho.

À minha mãe, Maria do Socorro F Firmino, e meu pai, José Macêdo Firmino, pelas bençãos, incentivos e por ter me mostrado a definição de amor, lealdade e respeito.

RESUMO

O câncer de pulmão é uma das principais causas de morte no mundo. Entretanto, esta mortalidade poderia ser reduzida se os pacientes fossem diagnosticados e tratados nos estágios iniciais da doença. De acordo com a literatura, a tomografia computadorizada é a modalidade de imagem mais indicada para a detecção precoce de nódulos pulmonares. No entanto, ela impacta diretamente na carga de trabalho dos radiologistas. Visando solucionar estas barreiras, o presente trabalho propõe um novo sistema de auxílio à detecção de nódulos pulmonares em exames de tomografia computadorizada. O sistema, chamado de LCD-OpenPACS (*Lung Cancer Detection* - OpenPACS), deverá ser integrado ao sistema OpenPACS e ter todos os requisitos necessários para ser utilizado no fluxo de trabalho das unidades de saúde pertencentes ao Sistema Único de Saúde Brasileiro. O LCD-OpenPACS fez uso de técnicas de processamento de imagem (segmentação por Crescimento por Regiões e Watershed), extrator de características (Histograma do Gradiente Orientado), redutor de dimensionalidade (Análise de Componentes Principais) e classificador de padrões (Máquina de Vetor de Suporte). O sistema foi testado com 220 exames de diferentes pacientes, totalizando 296 nódulos pulmonares, e obteve uma Sensibilidade de 94,4% com 7,04 Falsos Positivos por caso. O tempo de processamento foi de aproximadamente 10 minutos por exame. O sistema detectou nódulos pulmonares (solitários, opacidade em vidro fosco, justavascular e justapleural) entre 3 mm e 30 mm.

Palavras-chave: Sistemas CADe, Sistema de Auxílio ao Diagnóstico de Nódulos Pulmonares, Câncer de Pulmão, Processamento de Imagens Médica, LCD-OpenPACS.

ABSTRACT

Lung cancer is one of the most common types of cancer and has the highest mortality rate. Patient survival is highly correlated with early detection. Computed Tomography technology services the early detection of lung cancer tremendously by offering a minimally invasive medical diagnostic tool. However, the large amount of data per examination makes the interpretation difficult. This leads to omission of nodules by human radiologist. This thesis presents a development of a computer-aided diagnosis system (CADe) tool for the detection of lung nodules in Computed Tomography study. The system, called LCD-OpenPACS (Lung Cancer Detection - OpenPACS) should be integrated into the OpenPACS system and have all the requirements for use in the workflow of health facilities belonging to the SUS (Brazilian health system). The LCD-OpenPACS made use of image processing techniques (Region Growing and Watershed), feature extraction (Histogram of Gradient Oriented), dimensionality reduction (Principal Component Analysis) and classifier (Support Vector Machine). System was tested on 220 cases, totaling 296 pulmonary nodules, with sensitivity of 94.4% and 7.04 false positives per case. The total time for processing was approximately 10 minutes per case. The system has detected pulmonary nodules (solitary, juxtavascular, ground-glass opacity and juxtapleural) between 3 mm and 30 mm.

Keywords: CAD systems, Lung Cancer, Medical Image Processing, LCD-OpenPACS.

.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE SÍMBOLOS E ABREVIATURAS

CADe	*Computer-Aided Detection System*
DICOM	*Digital Imaging and Communications in Medicine*
DLF	Discriminante Linear de Fisher
FN	Falso Negativo
FP	Falso Positivo
HoG	Histograma do Gradiente Orientado
LCD-OpenPACS	*Lung Cancer Detection* - OpenPACS
LIDC	*Lung Image Database Consortium*
LIDC-IDRI	*Lung Image Database Consortium image collection*
PACS	Sistemas de Comunicação e Arquivamento de Imagens Médicas
PCA	Análise de Componentes Principais
SUS	Sistema Único de Saúde Brasileiro
SVM	Máquina de Vetor de Suporte
TC	Tomografia Computadorizada
UH	Unidades de Hounsfield
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

CAPÍTULO 1

INTRODUÇÃO

Câncer é um grupo de doenças caracterizadas por maturação, crescimento e/ou proliferação desordenada de grupos celulares anômalos [ACS 2011]. Segundo a OMS (2015), o câncer é responsável por 8,2 milhões de mortes por ano. Entre este grupo de doenças se destaca o câncer de pulmão, que juntamente com o câncer de traqueia e brônquios, apresentam uma taxa de aproximadamente 1,6 milhão de óbitos por ano, correspondendo a quinta maior causa de morte no mundo [OMS 2015]. O câncer de pulmão é a terceira neoplasia maligna mais frequentemente diagnosticada, representando 15% de todos os tumores. Entretanto, ela é primeira em taxa de mortalidade, sendo responsável por 30% de todas as mortes por câncer, porcentagem maior que a do câncer da mama, da próstata, do cólon e do ovário somadas [Zamboni & Carvalho 2005].

No Brasil, a estimativa para o ano de 2015 aponta para a ocorrência de aproximadamente 576 mil novos casos de câncer, sendo 27.330 incidências de câncer de pulmão que resultará em aproximadamente 23.503 óbitos [INCA 2014]. Os principais fatores de risco desta patologia são [Zamboni & Carvalho 2005]: tabagismo, presença de doença pulmonar preexistente, exposição ocupacional (por exemplo, asbesto, urânio, cromo e agentes alquilantes) e histórico familiar de câncer de pulmão. No entanto, muitos casos de câncer de pulmão não estão limitados a uma única causa, mas uma combinação de fatores. A maior parte dos casos acomete indivíduos entre 50 e 70 anos de idade.

A detecção em estágios iniciais é considerada a forma mais eficaz para melhorar a sobrevivência dos pacientes, pois nesse caso, a taxa de sobrevivência de cinco anos é de aproximadamente 54% [NIH 2015]. Por outro lado, quando essa patologia é detectada em estágios avançados a taxa de sobrevida em cinco anos é de apenas 4% [NIH 2015]. Normalmente, o câncer de pulmão é detectado em estágios avançados, uma vez que os

tumores iniciais não costumam produzir sintomas. Dessa forma, de cada 100 novos casos de câncer do pulmão, 80 são inoperáveis e a maioria deles morre dentro de três anos, somente três ou quatro casos sobrevivem após cinco anos [Barros et al. 2006].

O câncer de pulmão é diagnosticado normalmente através de exames de imagem de radiografia ou tomografia computadorizada (TC) do tórax [Walker 2006]. A radiografia de tórax utiliza radiação ionizante sob a forma de raios X para gerar uma imagem bi-dimensional dos pulmões. A TC é considerada a modalidade de imagem mais indicada para a detecção precoce e diagnóstico do câncer de pulmão. Essa utiliza uma série de raios X para permitir a reconstrução 3D da anatomia dos pulmões. A TC fornece imagens com alta resolução espacial, temporal e de contraste das estruturas anatômicas do tórax. Segundo Kazuo *et al.* (2004), a taxa de detecção de câncer de pulmão usando TC é de 2,6 a 10 vezes maior do que usando radiografia tradicional. Entretanto, a TC gera um grande número de imagens médicas que aliado a sobrecarga de trabalho dos radiologistas pode resultar em erros na detecção (falha em detectar um câncer) ou erro de interpretação (incapacidade de diagnosticar corretamente um tumor) [El-Baz et al. 2013]. Consequentemente, sistemas computacionais tornaram-se úteis para auxiliarem os radiologistas nas suas tomadas de decisão.

Sistemas CADe (*Computer-Aided Detection System*) é uma classe de sistemas computacionais que visam auxiliar na detecção de lesões nas imagens médicas através de uma "segunda opinião" [Suzuki 2012]. Jeon *et al.* (2012), visando comprovar a importância desses sistemas para a Radiologia, solicitaram que sete radiologistas analisassem 134 exames de TC e determinassem a presença de nódulos pulmonares. Depois, os mesmos radiologistas reviram as suas decisões após analisarem os resultados do sistema CADe. Como resultado, a sensibilidade na detecção de nódulos foi aumentada de 77% na avaliação inicial para 84% com o auxílio do sistema CADe [Jeon et al. 2012*a*].

Sistemas voltados para detecção de câncer tornou-se uma das principais áreas de pesquisa em imagiologia médica, pois o desenvolvimento desses sistemas pode resultar na aceleração do diagnóstico, redução de erros e melhoraria na avaliação quantitativa [van Ginneken et al. 2011]. Segundo van Ginneken *et al.* (2011), os atuais sistemas CADe não são amplamente utilizados na prática clínica pois não apresentam quatro requisitos principais: (a) melhorar o desempenho dos radiologistas; (b) reduzir o tempo necessário para o diagnóstico; (c) integrar perfeitamente com o ambiente de trabalho da equipe médica; e (d) apresentar custos insignificante ou reduzir custos hospitalares que justifique a sua implantação.

1.1 Revisão Bibliográfica

Para conhecermos melhor os sistemas CADe que auxiliam na detecção de nódulos pulmonares, será apresentado um levantamento bibliográfico das palavras chaves relacionadas ao tema nas base de dados da PubMed, Science Direct e IEEExplore. Um total de 420 artigos foram encontrados utilizando como palavras-chave: detecção de nódulos pulmonares, sistema de detecção de câncer, câncer de pulmão em computadorizada tomografias, CADe, detecção de câncer de pulmão e análise de imagem médica. Os resultados da pesquisa foram filtrados e *proceedings*; editoriais e cartas foram excluídos. Os artigos publicados entre 2009 e 2015 foram selecionados. No entanto, artigos que omitiram que o número de Falso Positivo (FP), número de nódulos usados na validação e sensibilidade foram excluídos. Finalmente, 70 artigos foram usados em nosso estudo. FP representa os resultados positivos quando a amostra não apresenta a doença e, sensibilidade é a capacidade que um sistema tem de discriminar, dentre os suspeitos de uma patologia, aqueles efetivamente doentes [Guimaraes 1985].

Os primeiros sistemas e patentes CADe para detecção de nódulos pulmonares surgiram no final da década de 80 [Giger et al. 1988] e [Doi et al. 1990]. Embora resultados interessantes tenham sido obtidos, essas primeiras tentativas não foram bem sucedidas devido a falta de recursos computacionais e de técnicas avançadas de processamento de imagem. No entanto, pesquisas já mostravam que o uso de sistemas CADe melhorava a precisão dos radiologistas nos diagnósticos, mesmo com um número grande de falsos positivos [Chan et al. 1990].

Visando melhorias nesses sistemas foram iniciadas várias pesquisas científicas, entre elas temos a de Xu *et al.* (1997). Eles utilizaram limiarização e redes neurais artificiais e obtiveram uma sensibilidade de 70%, com 1,7 FP por imagem. Com o intuito de melhorar o processo de automação e precisão, Armato *et al.* (1999), desenvolveram um dos primeiros sistemas para detecção de nódulos pulmonares que utilizavam imagens de tomografia computadorizada. Eles utilizavam técnicas de limiarização e *Linear Discriminant Analysis* e obtiveram uma sensibilidade de 70% com 9,6 FP por caso [Armato et al. 1999]. Esse sistema foi validado com 187 nódulos (solitários e justapleurais) com tamanhos entre 3,1 mm e 27,8 mm.

Lee *et al.* (2001) desenvolveram uma técnica que utilizava algoritmo genético e *template matching*, apresentando uma sensibilidade de 72% com o 25,3 falsos positivos por caso. Na validação do sistema foram utilizados 98 nódulos que possuíam dimensões menores do que 10 mm. Suzuki *et al.* (2003) desenvolveram uma técnica de reconhecimento de padrões baseado em uma rede neural artificial, chamada MTANN, e obtiveram uma

sensibilidade de 80,3% com 4,8 FP por caso, sendo testado com 121 nódulos (solitários, justapleurais, justavascular e opacidade em vidro fosco) com dimensões entre 4 mm e 25 mm [Suzuki et al. 2003*a*].

Armato *et al.* (2004) criaram o *Lung Image Database Consortium* (LIDC), um banco de imagens público de TC do tórax de pacientes normais e com câncer pulmonar em vários estágios. Esse banco de dados se tornou-se a fonte de dados mais utilizada no desenvolvimento, treinamento e avaliação de sistemas CADe de detecção de nódulos pulmonares.

Murphy *et al.* (2010) apresentaram um sistema CADe, chamado de ISI CAD, que utilizaram técnica de crescimento por região, filtros geométricos e classificador *k-nearest neighbor*, apresentando uma sensibilidade de 84% com 8,2 falsos positivos por caso sendo testado com 268 nódulos (solitários e justapleurais) com tamanhos entre 2 mm e 14 mm. Ye *et al.* (2009) propuseram um novo método que utilizava limiarização *fuzzy*, mapas de características, limiarização adaptativa e Máquina de Vetor de Suporte (SVM), apresentando uma sensibilidade de 90,2% e 8,2 falsos positivo por caso, sendo validado com 220 nódulos (solitários, justapleurais, justavascular e opacidade em vidro fosco) de tamanhos entre 2 mm e 20 mm.

Messay *et al.* (2010) apresentaram um sistema CADe que usava limiarização, processamento morfológico e Discriminante Linear de Fisher (DLF), obtendo uma sensibilidade de 82,66% com 3 falsos positivos por caso, sendo validado com 143 nódulos (solitários, justapleurais, justavascular e opacidade em vidro fosco), com dimensões de 3 mm a 30 mm. Liu *et al.* (2010) propuseram uma abordagem que utilizava limiarização com algoritmo *rolling ball* e Máquinas de Vetor de Suporte. Eles obtiveram uma sensibilidade de 97% e uma taxa de 4,3 FP por caso. Um ponto negativo em relação a esse trabalho diz respeito à validação do sistema no qual se limitou a testes com apenas 32 nódulos, sendo 31 nódulos solitários.

Gomatrhi & Thangara (2010) utilizaram técnicas de processamento de imagem (*Bit-Plane Slicing*, algoritmo *Fuzzy-C-Mean* e classificador neural. Esse sistema apresentou uma eficiência de 76,9% e 122 falsos positivos sendo validado com apenas 13 nódulos. Kumar *et al.* (2011) utilizaram transformada *Wavelet Biorthogonal*, crescimento por região e um sistema de inferência *fuzzy*. Esse sistema apresentava uma abordagem diferente, ele não apenas determinava a presença dos nódulos como também os classificava em nódulo benigno e neoplasia maligna. O sistema apresentou uma sensibilidade de 86% e 2,17 falsos positivos por caso, sendo validado com 538 nódulos. Tan *et al.* (2011) utilizaram limiarização e classificador neural, obtendo uma sensibilidade de 87,5% com uma média de 4 FP por caso, sendo testado com 574 nódulos (solitários, justapleurais e justavascular)

com diâmetro entre 3 mm e 30 mm.

Fantacci *et al.* (2011) e Camarlinghi *et al.* (2012) compararam o desempenho de sistemas CADe desenvolvidos pelo grupo de pesquisa Magic-5 [1]. Esse grupo desenvolveu sistemas CADe para a detecção de nódulos pulmonares, são eles: *Channeler Ant Model* (CAM), *Region Growing Volume Plateau* (RGVP) e *Voxel Based Neural Approach* (VBNA). De acordo com esses estudos, o CAM foi o sistema que apresentou o melhor desempenho. O CAM utiliza crescimento por regiões, colônia de formiga e redes neurais artificiais. Essa abordagem apresentou uma sensibilidade de 80% com 10 FP por caso quando testado com 114 nódulos.

Shao *et al.* (2012) utilizaram filtros *Wiener*, limiarização adaptativa e SVM (Máquina de Vetor de Suporte), obtendo uma sensibilidade de 89,47% com 11,9 de FP por caso quando testados com 44 nódulos pulmonares solitários. Cascio *et al.* (2012) fizeram uso de um classificador neural, técnica de crescimento por região com filtros morfológicos e *Mass-spring models* para eliminar falsos nódulos. O sistema obteve um desempenho de 97% com 6,1 falsos positivos por caso sendo validado com 148 nódulos (solitários e justapleurais).

Orozco *et al.* (2012) apresentaram um sistema que utilizava a transformada discreta do cosseno, transformada rápida de Fourier e SVM. Esse sistema apresentou uma sensibilidade de 96,15% com 2 falsos positivos por caso, quando avaliado com 50 nódulos. Teramoto & Fujita (2013) utilizaram filtros cilíndricos e SVM para eliminar os falsos positivos. O sistema foi validado com 103 nódulos (solitários, justapleurais, justavascular e opacidade em vidro fosco), com diâmetro entre 5 mm e 20 mm, e obteve uma sensibilidade de 80% com 4,2 FP por caso. Han *et al.* (2015) propuseram um sistema CADe que utilizava *hierarchical vector quantization* e Máquina de Vetor de Suporte. O sistema foi validado em 205 casos e obteve uma sensibilidade de 82,7% com 4 FP por imagem.

Sistemas CADe, embora comprovadamente, melhoram o desempenho dos radiologistas na detecção de nódulos pulmonares, não são amplamente utilizados no cotidiano clínico [van Ginneken et al. 2011]. Na revisão bibliográfica foi observado que as soluções existentes não solucionaram esse problema porque elas não se preocupavam com a aplicabilidade dos sistemas. Os pesquisadores se preocupam apenas com o desenvolvimento do método visando o desempenho dos radiologistas e a redução do tempo necessário para o diagnóstico. Novas soluções deverão ser desenvolvidas visando também integrar o sistema de detecção ao fluxo de trabalho da equipe média e possuir baixo custo que justifique a sua implantação.

[1] `http://www.magic5.unile.it/`

1.2 Objetivos

O principal objetivo desse trabalho é desenvolver um sistema, chamado de LCD-OpenPACS que será incorporado ao Sistema OpenPACS permitindo o auxílio ao diagnóstico de detecção de nódulos pulmonares em exames de tomografia computadorizada. O sistema OpenPACS é um sistema de comunicação e arquivamento de imagens médicas desenvolvido pelo LAIS (Laboratório de Inovação Tecnológica em Saúde) da Universidade Federal do Rio Grande do Norte. O sistema LCD-OpenPACS deverá ter todos os requisitos necessários para ser utilizado no fluxo de trabalho das unidades de saúde pertencentes ao Sistema Único de Saúde Brasileiro (SUS). O LCD-OpenPACS é um sistema gratuito que será disponibilizado ao Sistema Único de Saúde Brasileiro (SUS), com o intuito de minimizar os gastos de sua implantação. O mesmo visa auxiliar aos radiologistasa otimizar o seu desempenho na detecção de nódulos pulmonares.

Objetivos Específicos

Os desenvolvimentos teóricos, a análise de desempenho da solução proposta e as implicações dos resultados observados compõem a maior parte das objetivos deste trabalho. Além disso, podemos discriminar os objetivos específicos a seguir:

- levantar o estado da arte relacionado a sistemas de auxílio ao diagnóstico de câncer de pulmão;
- mapear o conhecimento médico de radiologistas experientes para o desenvolvimento do sistema proposto;
- modelar um sistema que permita a telerradiologia para o pré-diagnóstico de nódulos pulmonares;
- utilizar a técnica Watershed para segmentar as estruturas pulmonares;
- utilizar a técnica Histograma do Gradiente Orientado (HoG) para caracterizar os nódulos pulmonares;
- comparar as técnicas de classificação Máquina de Vetor de Suporte (SVM), Discriminante Linear de Fisher (DLF) e Naive Bayse quando aplicada ao HoG de imagens médicas;
- validar o sistema de detecção através de cenários de teste baseados em dados reais provenientes de uma base de dados pública com mais de 200 exames de tomografia do tórax com pacientes normais e com câncer de pulmão em vários estados.

1.3 Etapas de Projeto

O procedimento utilizado no desenvolvimento deste trabalho possui uma abordagem dividida em quatro estágios. A seguir, será descrito cada um dos estágios.

1. Estudo bibliográfico: consiste na busca por bibliografia de referência e soluções anteriores para o problema. Uma vez encontra as soluções existentes é analisado se o problema foi totalmente resolvido. Caso o problema não tenha sido resolvido ou foi parcialmente resolvido, o próximo estágio é a criação de um modelo que vise solucionar o problema.
2. Modelagem do problema: com base no referencial teórico são criados representações e modelos visando entender o problema de forma eficaz.
3. Elaboração de soluções: a partir do entendimento do problema é possível elaborar soluções algorítmicas a fim de solucionar o problema.
4. Análise experimental: é realizado experimentos com dados reais para validar a solução. Caso os resultados não sejam satisfatórios deverá voltar para a etapa de elaboração de soluções ou de modelagem do problema a fim de verificar possíveis falhas que levaram ao resultado não satisfatório.

1.4 Estrutura da Tese

Além desta Introdução, o trabalho está organizado em mais cinco Capítulos, cujos conteúdos são individualmente discriminados a seguir.

O Capítulo 2, trata-se dos fundamentos teóricos relacionados ao desenvolvimento teórico do sistema proposto e que Capítulo este importante para o entendimento dos diversos aspectos envolvidos nesta tese. Nele é apresentado o sistema OpenPACS, alguns fundamentos de processamento de imagens médicas (incluindo segmentação de imagens, extração das características e redução de dimensionalidade), classificação de padrões, gerenciamento dos dados para classificação e métodos de validação dos classificadores.

No Capítulo 3 é apresentada a proposta, ou seja a contribuição científica do trabalho, de um sistema de detecção de nódulos pulmonares em imagem de tomografia computadorizada. Inicialmente, é apresentado o fluxo de trabalho da área de Radiologia em uma unidade de saúde que utiliza o OpenPACS e como o LCD-OpenPACS está incorporado a esse fluxo. Em seguida, é exibida a arquitetura do método, são descritas as funcionalidades e apresentado os algoritmos de cada módulo que ela contém. Foram utilizados dados de pacientes reais em experimentos para avaliar a aplicabilidade e a precisão do método proposto.

No Capítulo 4 serão apresentados os materiais que foram manipulados nos experimentos de validação e os resultados obtidos. Ainda nesse Capítulo, serão apresentadas comparações dos resultados obtidos pelo método com outros métodos existentes na literatura. A partir dessas avaliações poderemos inferir se o método proposto apresenta as características necessárias para utilização em um ambiente hospitalar. E, por último, são mostradas as limitações do método proposto.

Por fim, no Capítulo 5, são discutidas as conclusões obtidas nos resultados do trabalho e, serão ainda apresentados possíveis direcionamentos para novas pesquisas.

1.5 Publicações Relacionadas

Artigo em Periódico

Firmino, M.; Morais, A. H.; Mendonça, R.M.; Dantas, M. R.; Hekis, H.R. e Valentim, R. ***Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects.***. Biomedical engineering Online, 8, 13-41, 2014.

Firmino, M.; Giovani Angelo, G.; Morais, H.; Dantas, M. R. e Valentim, R. ***Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy.***. Biomedical engineering Online, 15, 1-16, 2016.

Firmino Filho, J., Valentim, R., Ribeiro, M., Cavalcanti, L. **OpenPACS - Sistema *Open Source* para Comunicação e Arquivamento de Imagens Médicas: Relato de Experiência em um Hospital Universitário** - DOI: 10.3395/reciis.v7i2. Sup1.772pt. Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, Brasil, 7, aug. 2013. Disponível em: `http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/772/1656`. Acesso em: 02 Sep. 2014.

Anais de Congresso

Firmino, Macedo; Pereira, Sheila; Valentim, Ricardo. **PACS - Sistema de Comunicação e Arquivamento de Imagens Médica: Visão Introdutória e Usabilidade no Sistema de Saúde Brasileiro**. Anais do VII CONNEPI - Congresso Norte Nordeste de Pesquisa e Inovação, Tocantins - 2012.

Capítulo de Livro

Macêdo Firmino, Marcel Ribeiro Dantas, Higor Morais, Bruno Gomes e Ricardo Valentim. **LCD-OpenPACS: Sistema Integrado de Telerradiologia com Auxílio ao Diagnóstico de Câncer de Pulmão em Exames de Tomografia Computadorizada.**. Editora UFRN. Natal – RN, 2014.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

O entendimento dos fundamentos de alguns conceitos de Sistemas de Comunicação e Arquivamento de Imagens Médicas (PACS), processamento de imagens médicas, extração das características, redução de dimensionalidade e classificação de padrões é condição necessária para o desenvolvimento teórico a que se propõe este trabalho. Dessa forma, esse Capítulo tem por objetivo apresentar esses e outros fundamentos no qual se enquadra sistemas LCD-OpenPACS. Além disso, serão apresentadas algumas definições que serão utilizadas ao longo da tese.

Esse Capítulo está organizado da seguinte forma. Na Seção 2.1 serão apresentados os fundamentos do sistema OpenPACS, sua arquitetura e vantagens. Na Seção 2.2 serão mostrados alguns fundamentos de processamento de imagens médicas visando à compreensão de conceitos e os principais algoritmos utilizados na segmentação de imagens médicas, extratores de características e redutores de dimensionalidade. Na Seção 2.3 será apresentado a definição de alguns algoritmos de classificação de padrões que serão utilizados no Capítulo de Experimentos e Resultados. Na sequência, na Seção 2.4 será mostrada as principais técnicas de gerenciamento de dados para a realização de classificação. E, por último, na Seção 2.5 será exibida a Matriz de Confusão e o cálculo da Sensibilidade que são os métodos de validação de classificadores.

2.1 Sistema OpenPACS

As imagens baseadas em filme (por exemplo, películas de raios X) vêm desempenhando um importante papel nos departamentos de imagem hospitalar nas últimas déca-

das. No entanto, atualmente, hospitais estão encontrando os seguintes problemas [Strickland 2000]:

- 20% das imagens em filmes se perdem. Isso resulta na repetição do exame ou na falta de informação do estado de saúde de um paciente. Exames repetidos desnecessariamente conduzem a exposição à radiação adicional, bem como ao desperdício de recursos: de tempo, humano e monetário;
- descarte de produtos químicos utilizados na Radiologia, pois alguns desses produtos são nocivos ao meio ambiente;
- tempo gasto para repetir exames ou encontrar um determinado arquivo pode resultar na piora do estado clínico do paciente e ainda provocar danos à reputação do hospital;
- tempo perdido por membros da equipe médica revelando películas ou procurando exames, enquanto eles poderiam estar realizando mais procedimentos;
- uma imagem impressa pode não estar disponível em vários locais simultaneamente. Dessa forma, se um médico desejar uma segunda opinião, o exame deverá ser encaminhado até outro especialista;
- para determinar a evolução clínica de um paciente, se faz necessário ter disponível tanto o exame atual quanto o anterior. No entanto, não é possível realizar essa comparação quando os exames de imagem anteriores não estão disponíveis em sua totalidade, seja por perda ou por não impressão.

Com o advento da tecnologia digital, surgiu um novo sistema de informação para a aquisição, o armazenamento, a distribuição e a exibição de imagens médicas, denominado PACS (Sistema de Comunicação e Arquivamento de Imagens Médicas). Caso os departamentos de imagem hospitalar não passem a utilizar esse sistema de informação, os problemas atualmente encontrados poderão se agravar, pois se espera que [Schulze et al. 2007]:

- a quantidade de exames de imagem aumente. Esse aumento será devido à disponibilidade crescente de modalidades de imagens especializadas, a mais exames não invasivos e ao aumento no número de pacientes (favorecido pelo envelhecimento global da população). Além disso, espera-se que a quantidade de dados gerados por exames de imagem aumente;
- a demanda por películas de raios X diminua, à medida que os hospitais optem por imagens digitais. Isso não só irá fazer aumentar o custo da película, mas também dos produtos químicos e da manutenção do sistema.

O sistema PACS é descrito por uma arquitetura formada por componentes integrados por redes de dados e *software* de aplicação que permite que imagens médicas digitais sejam recebidas, visualizadas e analisadas por especialistas em diferentes estações de trabalho [Law & Zhou 2003]. Em outras palavras, quando se utiliza este sistema computacional, em vez de existirem cópias de películas de raios X para serem processadas, transportadas e armazenadas manualmente, as imagens digitais são processadas, transportadas e armazenadas pelos computadores. A maioria dos PACS utiliza navegadores de Internet para permitir o fácil acesso às imagens pelos usuários em qualquer localidade. Os médicos podem observar as imagens radiológicas, por exemplo, em suas residências.

Os hospitais que utilizam sistema PACS apresentaram um aumento na qualidade e eficiência do atendimento [Law & Zhou 2003]. Segundo Schulzer et al. (2007) o principal beneficiário, dos hospitais que estão utilizando esta ferramenta, é o paciente, pois se observa redução significativa do tempo total desde a requisição do exame até à disponibilização para o médico, levando à diminuição da permanência hospitalar em 22%. A diminuição da permanência dos pacientes implica na diminuição do risco de infecções [Banta 1990].

Entre os sistemas PACS se destaca o OpenPACS que é um sistema de código fonte aberto e multiplataforma desenvolvido pelo Laboratório de Inovação Tecnológica em Saúde (LAIS) da Universidade Federal do Rio Grande do Norte (UFRN). O OpenPACS é composto por um conjunto de *software* e *hardware* formando uma arquitetura voltada para aquisição, armazenamento, distribuição e exibição de imagens médicas [Firmino et al. 2013]. Na sequência será apresentada a arquitetura do sistema OpenPACS e suas vantagens.

Arquitetura

O OpenPACS é formado por cinco componentes interligados por meio de redes de computadores e aplicações computacionais. Os componentes são: Modalidades, Servidor, Estações de Trabalho, *Gateway* de Internet e Servidor de *Backup* [Firmino et al. 2014], ver Figura 2.1.

As Modalidades são os equipamentos responsáveis pela aquisição de imagem, por exemplo, tomografia computadorizada, ultrassonografia, angiografia e radiologia digital. As principais funções das Modalidades são: aquisição de imagens de forma confiável e em tempo hábil, conversão dos dados para o padrão DICOM (*Digital Imaging and Communications in Medicine*) e o envio das imagens para o servidor PACS.

DICOM é um conjunto de normas a serem seguidas pelos dispositivos médicos para

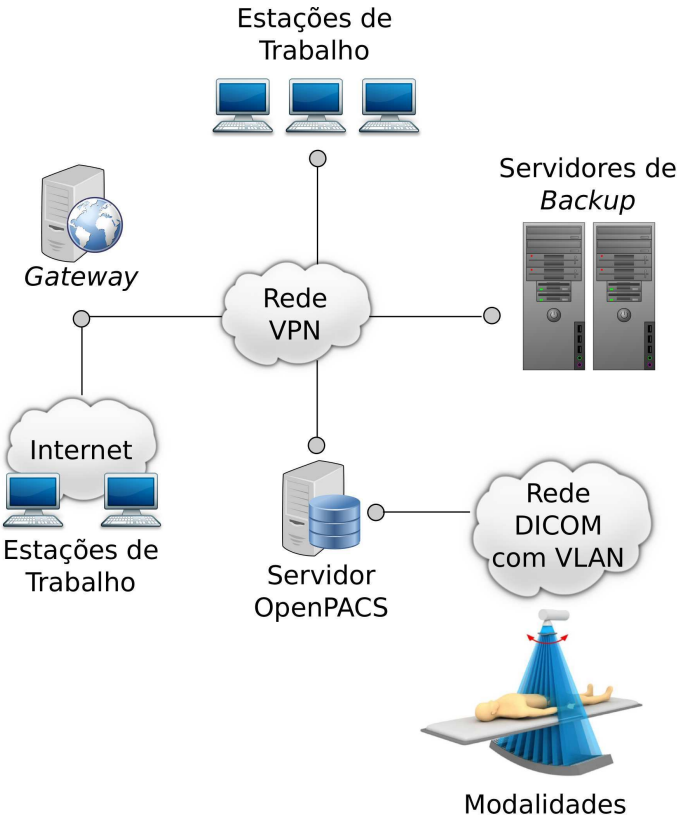

Figura 2.1: Arquitetura do sistema OpenPACS

armazenamento e transferência dos dados visando permitir a interoperabilidade [Pianykh 2008]. Ele define comandos e informações que podem ser trocadas pelos equipamentos e determina serviços de armazenamento de mídia a serem seguidos pelos dispositivos, bem como o formato do arquivo e sua estrutura de diretórios. Maiores informações sobre o protocolo DICOM consulte Pianykh (2008).

Uma vez que as imagens adquiridas, elas devem ser arquivadas para posterior avaliação pelos radiologistas e clínicos. O arquivamento ocorre no segundo componente do sistema, chamado de Servidor. Ele é um sistema gratuito, baseado no DCM4CHEE, multiplataforma, desenvolvido em Java e possui um banco de dados para armazenar informações DICOM e outros dados clínicos. O Servidor possui as seguintes funções [Firmino et al. 2014]:

- controlar a comunicação e o fluxo de dados no sistema. O Servidor recebe as imagens das modalidades, manda uma cópia para o Servidor de *Backup* e envia as imagens para as Estações de Trabalho (quando solicitado pelos radiologistas e médicos autorizados) através do protocolo DICOM;
- armazenar os dados garantindo a integridade por meio de um Sistema Gerenciador

de Banco de Dados (SGDB);

- garantir a disponibilidade dos dados, ou seja, importação/exportação dos exames médicos. A disponibilidade se faz importante porque longos períodos de inatividade não podem ser tolerados. Para garantir esta propriedade são utilizados de medidas de tolerância a falhas que incluem: detecção de erros, registro de *logs*, programas de auditoria, redundância dos dados, redundância de *hardware* e programas de monitoramento (estado da rede, espaço em disco, estado do banco de dados, utilização do processador e temperatura);
- permitir a interface com outros sistemas de informação hospitalar (por exemplo, sistema de informação radiológica e prontuário eletrônico);
- garantir a segurança das informações através de medidas para desencorajar, impedir, detectar e corrigir violações de segurança que envolva a transmissão das imagens médicas. São exemplos de mecanismos de segurança utilizados pelo Servidor: controle físico, criptografia dos dados, assinatura digital, *firewall*, rede virtual privada (VPN e VLANs) e protocolo seguro (HTTPS).

Em termos de *hardware*, o Servidor é um "*datacenter*" composto por computadores de alto desempenho e dispositivos de armazenamento (discos magnéticos, fitas magnéticas, CDs e DVDs) ligados por uma rede de alto desempenho. Devido à grande demanda por velocidade de acesso e confiabilidade, os discos utilizam a tecnologia RAID para armazenamento de dados [Firmino et al. 2014].

O terceiro componente é o Servidor de *Backup*. Esse tem a função de armazenar cópias dos dados para protegê-los de eventuais perdas. Os *backups* são salvos em vários discos e são agendados para serem executados automaticamente. Esse servidor deverá ser utilizado como Servidor PACS principal quando o principal estiver inativo, e, dessa forma, mantiver a disponibilidade do sistema [Firmino et al. 2014].

Uma parte importante do sistema OpenPACS são as Estações de Trabalho. Essas Estações são sistemas computacionais utilizados por radiologistas e clínicos para visualizar as imagens. Para isso, elas devem possuir um *software* para recuperar imagens, realizar o processamento e disponibilizá-la nos monitores. O OpenPACS disponibiliza um visualizador *Web* gratuito (chamado de Weasis), desenvolvido em Java, bidimensional (2D) projetado para proporcionar visualização Web e manipulação de imagens radiológicas. No entanto, outros clientes DICOM poderão ser utilizados [Firmino et al. 2014]. A Figura 2.2 ilustra o aplicativo Weasis mostrando um exame de tomografia computadorizada do tórax de um paciente de sexo masculino com 78 anos de idade.

O quinto componente do sistema PACS é o *Gateway* de Internet. Ele tem a função de

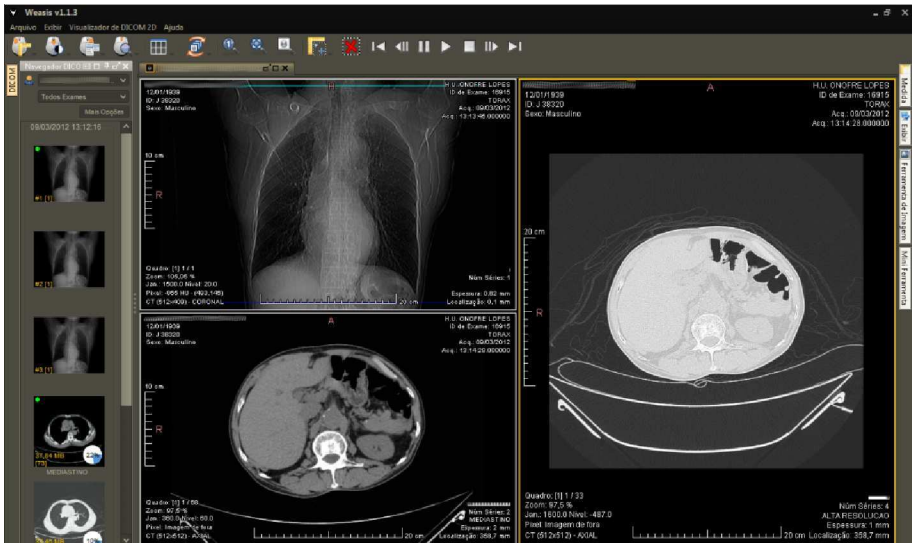

Figura 2.2: *Software* Weasis mostrando um exame de tomografia computadorizada do tórax.

permitir a Telemedicina. O Conselho Federal de Medicina define a Telemedicina como o exercício da medicina através da utilização de tecnologias da informação e comunicação, com o objetivo de análise de diagnóstico, prevenção e tratamento de doenças, educação e pesquisa em saúde [CFM 2002]. Com essa ferramenta, médicos de outras regiões podem emitir diagnósticos ou uma segunda opinião e acompanhar a evolução clínica dos pacientes em regiões do país que carecem de profissionais especializados. A telerradiologia é uma aplicação da Telemedicina que é caracterizada pelo envio remoto das imagens médicas com o objetivo de permitir o diagnóstico à distância (telediagnóstico) ou emitir uma segunda-opinião especializada (teleconsultoria).

O sistema OpenPACS está sendo validado através de sua utilização em uma pesquisa experimental, iniciada em janeiro de 2012, no Hospital Universitário Onofre Lopes (HUOL). Durante esse experimento, o sistema gerenciou mais de 80.000 exames, com mais de 70 usuários e aproximadamente 5.000 consultas por mês. A Figura 2.3 ilustra um usuário do sistema realizando consultas e realizando análises em exames armazenados pelo OpenPACS no HUOL.

Vantagens

As vantagens do sistema, obtidas através da sua utilização na pesquisa experimental, foram:

- disponibilizar ferramentas de processamento de imagem para permitir ao médico um diagnóstico preciso;

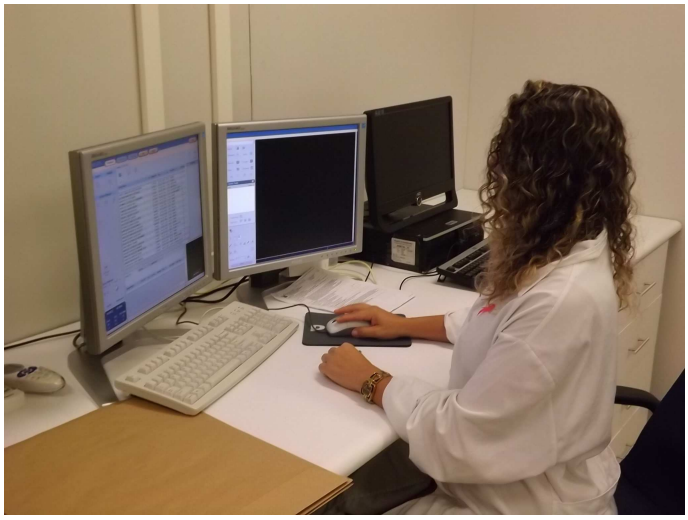

Figura 2.3: Usuário do sistema OpenPACS, na pesquisa experimental no HUOL, realizando consultas e análise de exames de pacientes.

- minimizar a taxa de impressão de exames resultando em economia de consumo de películas e produtos químicos, reduzindo o impacto ambiental;
- melhorar a acessibilidade aos exames, aos diagnósticos e aos resultados, uma vez que é possível consultar simultaneamente aos dados do PACS em várias estações de trabalho;
- facilitar a pesquisa e transmissão dos dados, através do uso de banco de dados e redes de comunicações;
- reduzir filmes extraviados, danificados ou desaparecidos.

2.2 Processamento de Imagens Médicas

Processamento de imagem médicas é um vasto conjunto de técnicas que visam a aquisição de imagens de estruturas anatômicas, reconstrução, compressão, armazenamento, visualização e análise das imagens médicas [Duncan & Ayache 2000]. As técnicas de processamento de imagens médicas são utilizadas nas áreas: educacional (por exemplo, no estudo da anatomia, simulação de estruturas anatômicas para o treinamento de cirurgia), diagnóstico (por exemplo, no auxílio ao diagnóstico de patologias cardíacas), tratamento (por exemplo, visualizações 3D de estruturas anatômicas para o planejamento de intervenções cirúrgicas ortopédicas e cirurgia craniofacial) e procedimentos operatórios (por exemplo, a colocação de *clips* intracranianos em aneurisma cerebral com o auxílio por imagens de angiografia).

Para o nosso propósito iremos apresentar, na sequência, alguns conceitos de segmentação de imagens e duas técnicas de segmentação que serão importantes para o desenvolvimento da presente pesquisa científica. Essas técnicas são: Crescimento por Regiões e *Watershed*.

2.2.1 Segmentação de Imagens

Atualmente, as pessoas têm recorrido a imagens para armazenar, exibir e fornecer informações sobre o mundo que as rodeiam. O objetivo das técnicas de processamento de imagens é extrair informação a partir de dados brutos de imagens, ou seja, converter imagens em informação. Uma questão central na extração de informações, a partir de uma imagem digital, é a redução das informações dessa imagem em regiões. Isto consiste no problema de segmentação da imagem. Conceitualmente, segmentação de imagens refere-se ao processo de decomposição de uma imagem digital em vários segmentos (regiões) que a formam [Jain 1989]. Com efeito, uma vez a imagem segmentada, é possível efetuar cálculos de características geométricas como área e volume de estruturas, ajudando a análise e a diagnose de doenças.

Pode-se dizer que a elaboração de algoritmos visando à automatização de processos de segmentação é uma das tarefas mais difíceis dentro da área de processamento de imagens médicas, pois a segmentação de imagem não é uma solução puramente analítica [Gonzales & Wintz 1987]. Dessa forma, se fazem necessárias informações a priori dos componentes que formam as imagens que estão sendo processadas.

No caso das imagens geradas pelas tomografias computadorizadas, cada *pixel* recebe valores que refletem a densidade dos vários tecidos. Esses valores são normalmente expressos na forma de coeficiente de atenuação relativa, ou Unidade de *Hounsfiled* (UHs) [Preim & Bartz 2007]. A Figura 2.4 lista valores de densidade Hounsfield para diferentes substâncias, órgãos e tecidos.

Na Figura 2.4 é possível observar que os tecidos pulmonares se apresentam com baixos valores de UH quando comparado com outros órgãos, tais como tecidos moles e tecido ósseo. Dessa forma, as técnicas de segmentação podem utilizar essa informação para separar os pulmões de outros tecidos e órgãos nas imagens de TC.

O desenvolvimento dos algoritmos de segmentação levam em consideração duas propriedades básicas dos *pixels*, são elas: descontinuidade e similaridade. A descontinuidade é uma propriedade de destacar objetos na imagem através de variações em tons de cinza entre o objeto e a região ao qual ele está inserido. Por outro lado, a similaridade tem como fundamento a observação das características dos *pixels* e não se preocupam com

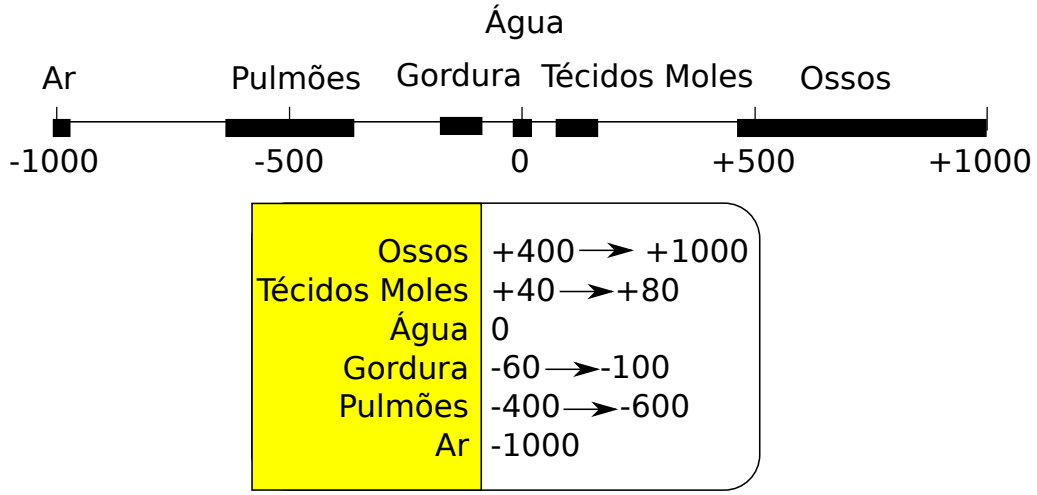

Figura 2.4: Valores de Hounsfield para algumas substâncias, tecidos e órgãos em imagens de TC.

a região ao qual o objeto está inserido. Para tanto, parte da premissa que os *pixels* que compõe um objeto têm propriedades similares, mesmo estando em várias regiões da imagem, enquanto que *pixels* de objetos distintos têm propriedades distintas [Gonzales & Wintz 1987].

Na sequência, iremos apresentar duas técnicas de segmentação que utilizam a abordagem detecção de similaridade e são utilizadas para segmentação de imagens médicas, são elas: Crescimento por Regiões e *Watershed*.

Crescimento por Regiões

O princípio do funcionamento da técnica de Crescimento por Regiões é agrupar *pixels* ou sub-regiões em regiões maiores. Seu início se dá com a adoção de um conjunto de *pixels*, chamados de sementes. A escolha dessas sementes geralmente é feita baseando-se na natureza do problema. Inicialmente, a região é formada somente pelas sementes. Na sequência, *pixels* vizinhos que tenham atributos similares (tais como, intensidade, textura, e cor) são incluídos na região até que se atinja um ponto de parada, que geralmente é baseado em algum critério de dessemelhança entre as regiões [Gonzales & Wintz 1987].

Por exemplo, observe a Figura 2.5 que mostra a utilização da técnica de Crescimento por Regiões para segmentar a massa branca cerebral de uma imagem de ressonância magnética usando a biblioteca SimpleITK. A massa branca é a massa que fica no interior do cérebro, que tem a função de transmitir as informações entre as regiões diferentes do cérebro. Inicialmente, na imagem à esquerda é mostrada o ponto semente (ponto branco na imagem) que será utilizado pelo algoritmo. Partindo desse ponto, o algoritmo analisa os

pixels vizinhos. Caso os mesmos possuam valores de cinza (intensidade) similares esses *pixels* são acrescentados a região segmentada. Na Figura 2.5 à direita é mostrado o resultado final da segmentação. A similaridade nesse caso foi o valor de intensidade entre 130 e 170 UH.

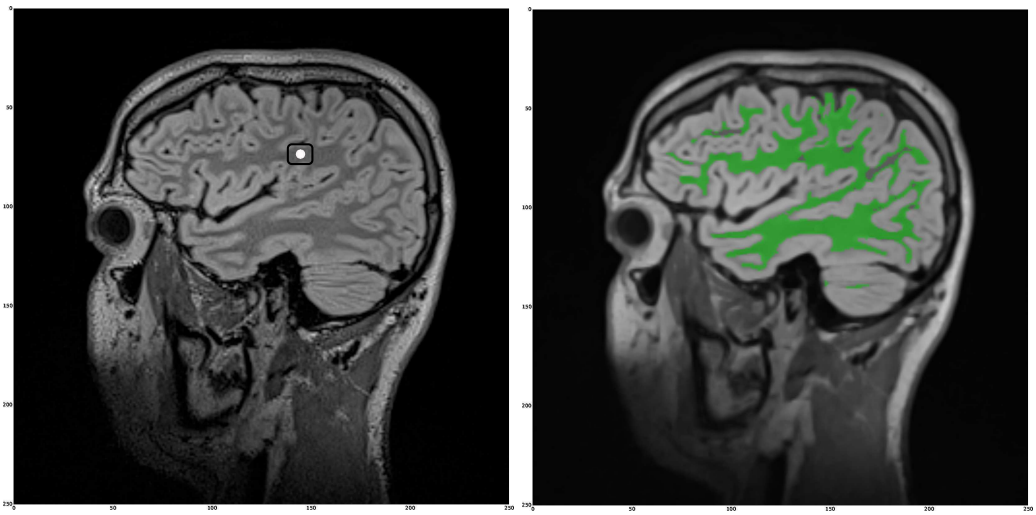

Figura 2.5: Exemplo de segmentação por Crescimento por Regiões da massa branca cerebral em uma imagem de ressonância magnética. Na imagem à esquerda é mostrado o ponto de semente e na imagem à direita o resultado da segmentação.

Watershed

O método de segmentação *Watershed* calcula o módulo do gradiente para todos os *pixels* da imagem. Os valores de gradiente formam uma pseudo superfície topográfica com vales e montanhas, fazendo uma analogia entre valores de *pixel* (nível de cinza) e altitude. As regiões mais baixas seriam correspondentes as de menor valor de *pixel*, as mais altas as de maior valor de *pixel* e regiões com valores constantes são chamadas de platôs [Audigier 2004]. Essa analogia é visualizável com uma imagem bidimensional em níveis de cinza, onde se acrescenta uma terceira dimensão que seria a altitude.

O funcionamento do método *Watershed* baseia-se no princípio da inundação de vales [Beare & Lehmann 2006]. Para entender, considere uma imagem bidimensional em níveis de cinza, representada por uma superfície topográfica. Os mínimos regionais da imagem são as regiões conexas na superfície que apresentam intensidade constante (zonas planas) e menor do que a intensidade dos pontos vizinhos destas regiões [Vincent & Soille 1991]. Imagine que esses mínimos regionais sejam perfurados e que a superfície seja imersa na água. A água vai penetrar pelos furos e encher as bacias hidrográficas até

que fluxos provenientes de mínimos regionais diferentes possam se unir. Nesse momento a inundação desses mínimos regionais para separando os dois mínimos locais. A segmentação acaba quando for possível identificar todos os mínimos locais. A vantagem do *Watershed* é a sua rapidez de processamento, permitindo que seja utilizado em aplicações interativas, mesmo quando as imagens a serem processadas sejam complexas [Beare & Lehmann 2006].

Um exemplo da utilização da segmentação *Watershed* é ilustrada na Figura 2.6. Nesse exemplo é segmentado regiões internas do tórax em uma imagem de tomografia computadorizada usando a biblioteca SimpleITK. Nessa mesma figura a imagem da esquerda mostra a imagem original que será utilizada no processo de segmentação. A partir da imagem original é calculado o valor do gradiente de cada *pixel*. Os valores do módulo do gradiente resultante são mostrados na imagem do centro. Na sequência, as regiões da imagem são separadas através dos mínimos locais dos valores do gradiente. Intuitivamente são como se as regiões fossem inundadas formando pequenos lagos que são os mínimos locais. A imagem à direita mostra o resultado final do processo de segmentação.

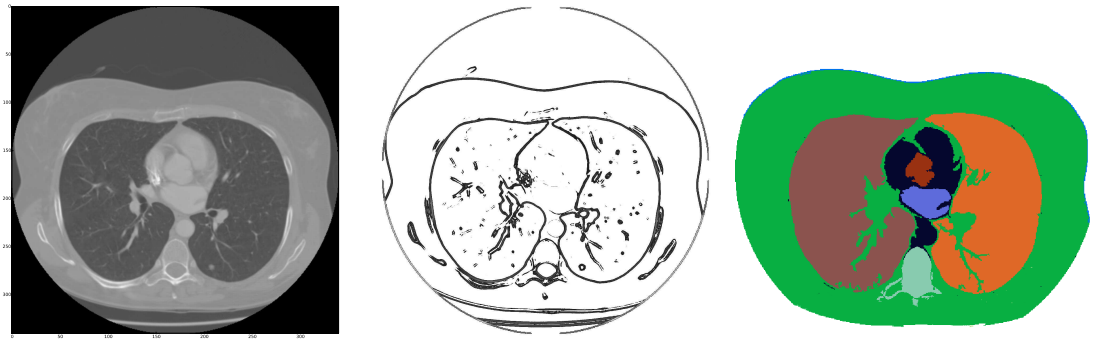

Figura 2.6: Exemplo de segmentação por *Watershed* da região toráxica em uma imagem de tomografia computadorizada.

2.2.2 Extração das Características

Extração das características são técnicas que buscam modelos matemáticos no qual um espaço de dados é transformado em um espaço de características. Dessa forma, o conjunto de dados pode ser representado por um número reduzido de características e ainda reter a maioria do conteúdo de informação intrínseco dos dados [Gonzales & Wintz 1987]. A escolha dessas características tem uma influência importante sobre: (1) a precisão de classificação, (2) tempo necessário para a classificação, (3) o número de exemplos necessários para a aprendizagem, e (4) custo da classificação. Consequentemente, a extração

de características é muito importante no diagnóstico médico, pois informações errôneas podem implicar na piora do quadro clínico dos pacientes e novos testes de diagnóstico. Na sequência será apresentado um extrator de características chamado de Histograma do Gradiente Orientado (HoG).

Histograma do Gradiente Orientado

O sistema LCD-OpenPACS adota uma nova abordagem para a extração das características de possíveis nódulos pulmonares visando diferencia-los de outras patologias pulmonares ou diferentes órgãos e tecidos. Esta nova abordagem faz uso do Histograma do Gradiente Orientado (HoG). O algoritmo HoG foi proposto por Dalal & Triggs (2005) aplicado ao problema de detecção de pedestres em imagens estáticas.

O HoG calcula as ocorrências de uma determinada orientação do gradiente em certas porções da imagem, a que mais ocorrer naquela região será considerado como a gradiente daquela partição da imagem. O algoritmo baseia-se na ideia de que a forma e a aparência de um objeto podem ser descritas muitas vezes pela intensidade dos gradientes ou a direção das bordas, sem um conhecimento prévio da posição de tais bordas [Dalal & Triggs 2005]. Dessa forma, partimos da premissa que o HoG seria um bom descritor para nódulos pulmonares, pois os radiologistas analisam a presença de tumores baseado na forma e aparência de objetos (manchas) presentes na imagem de tomografia.

Inicialmente, o algoritmo realiza uma normalização global da imagem, de forma a reduzir a influência de efeitos de iluminação na imagem, através de compressões gamma em cada canal de cor da imagem. Em seguida, calculam-se os gradientes de primeira ordem da imagem, que capturam contornos e algumas informações de textura. O gradiente é calculada da seguinte fórmula:

$$\begin{aligned} \nabla G(x,y) = [&G(x+1,y) - G(x-1,y), \\ &G(x,y+1) - G(x,y-1)] \end{aligned} \tag{2.1}$$

onde: G é uma imagem discreta em níveis de cinzas e (x,y) corresponde a localização dos *pixels*.

Na sequência, é realizado uma divisão da imagem em células que combinarão os histogramas de orientação do gradientes naquela região. Cada histograma de orientação divide a amplitude do ângulo do gradiente em um número predeterminado de *bins*. Cada célula gera se o histograma de orientações, onde se utiliza a magnitude do gradiente como peso.

Posteriormente, criam-se grupos de células, chamadas de blocos. Os blocos são nor-

malizados de modo a aumentar a invariância de cada célula à iluminação às sombras e contraste. Os blocos de descritores então normalizados são denominados Histogramas de Gradientes Orientados, que são finalmente combinados em um vetor de características para posterior classificação. Mais detalhes sobre o Método de Histogramas de Gradientes Orientados podem ser vistos em Dalal e Triggs (2005).

A Figura 2.7 ilustra a utilização do HoG para caracterização de imagem de um nódulo pulmonar e de uma artéria intrapulmonar. Na imagem à esquerda é mostrado à imagem DICOM original e na imagem à direita apresenta o resultado final do processo de extração das características. Visualmente o HoG apresenta-se como um tipo de textura, mas os valores que realmente os representa o extrator são os valores dos módulos dos gradientes. Neste cálculo foi utilizada uma orientação de quatro *bins*, grupamento de 5x5 *pixels* por células e blocos formados por 2x2 células.

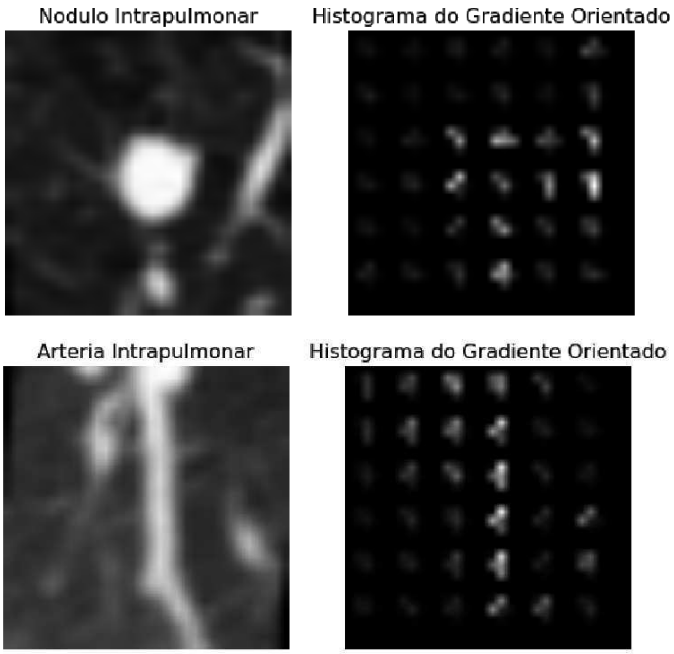

Figura 2.7: Histograma do Gradiente Orientado aplicado a caracterização das estruturas de nódulo (à acima) e de uma artéria intrapulmonar (à baixo) em exame de tomografia computadorizada.

2.2.3 Redução de Dimensionalidade

Alguns extratores, como é o caso do HoG, apresentam características com altas dimensionalidade. Essa propriedade resulta em uma grande quantidade de exemplos ne-

cessários para o treinamento e tempo para classificação. Para contornar esse problema existem técnicas matemáticas que realizam a redução de dimensionalidade mantendo o conteúdo de informação intrínseco dos dados. Entre essas técnicas matemáticas destaca-se a Análise de Componentes Principais (PCA) que será apresentado a seguir.

Análise de Componentes Principais

Análise de Componentes Principais (PCA) é uma técnica de redução de dimensionalidade que transforma um número de variáveis, possivelmente correlacionadas, em um número pequeno de variáveis não correlacionadas (chamados componentes principais) visando reduzir a redundância das informações [Haykin 1998]. Para isso, o PCA determina combinações lineares que propiciam uma maior variância empírica dos dados resultantes. Ou seja, é uma transformação linear ortogonal de um espaço q-dimensional para um espaço n-dimensional, com n $\leq$ q. A coordenada dos dados no novo espaço não são correlacionadas e a maior quantidade de variância dos dados originais é preservada.

Portanto, dadas p variáveis $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_p$ encontram-se combinações lineares para produzir novas variáveis $\mathbf{Z}_1$, $\mathbf{Z}_2$, ..., $\mathbf{Z}_p$ que não sejam correlacionadas, onde $\mathbf{Z}_i$ são as componentes principais e são ordenadas de forma que var($\mathbf{Z}_1$) $\geq$ var($\mathbf{Z}_2$) $\geq$ ... $\geq$ var($\mathbf{Z}_p$) e var($\mathbf{Z}_i$) representa a variância de $\mathbf{Z}_i$.

Inicialmente é realizada uma centralização dos dados em torno da média. Essa transformação é feita pela aplicação da seguinte equação:

$$\mathbf{Z}_i = \frac{(\mathbf{X}_i - \mu)}{\sigma} \tag{2.2}$$

onde $\mathbf{Z}_i$ são os valores transformados, $\mathbf{X}_i$ são as amostras para i = 1, 2,... n, μ a média da variável $\mathbf{X}_i$, σ é o desvio padrão. Desta forma, todas as variáveis aleatórias são distribuídas com média zero e desvio padrão unitário.

Na sequência é calculada a matriz de covariância $\mathbf{C}_z$:

$$\mathbf{C}_z = Z \cdot Z^T \tag{2.3}$$

onde, Z^T é a transposta de Z. Os elementos da diagonal principal desta matriz se referem às variâncias das colunas (variáveis independentes). Já os elementos fora da diagonal principal representam a covariância entre as variáveis.

O passo seguinte é a determinação dos autovalores (λ) e autovetores (ν_n) correspon-

dentes da matriz $\mathbf{C}_z$. Os autovetores são arranjados de modo decrescentes de acordo com os valores dos autovalores. Encontrados os autovetores ν_n, estes formarão as colunas de uma matriz **P**.

$$\mathbf{P} = \nu_1, \nu_2, ..., \nu_n \tag{2.4}$$

O último passo é a diagonalização. A matriz **P** é utilizada para mudar a base de $\mathbf{C}_z$ obtendo uma matriz diagonal **D** de autovalores de $\mathbf{C}_z$.

$$\mathbf{D} = \mathbf{P}^{-1}\mathbf{C}_z\mathbf{P} \tag{2.5}$$

onde a matriz **D** apresenta elementos iguais aos autovalores na diagonal principal.

Geometricamente, os dados passam por um processo de deslocamento e rotação do sistema de coordenadas. Para isso, é realizado um procedimento de remoção da média de todos os pontos disponíveis, seguido da combinação linear das variáveis originais, para assim, produzir vetores que descrevam a maior parte da variação do conjunto de dados originais. Esses vetores, chamados autovetores, são ordenados de acordo com variabilidade em que representam. O primeiro é o autovetor que corresponde à maior variância, sendo os demais ortogonais a este, ordena dos pela direção da maior variância dos resíduos do primeiro [Haykin 1998].

2.3 Classificação de Padrões

A classificação de padrões é definido como o processo pelo qual um padrão é atribuído a uma classe dentre um número pré-determinado de classes (categorias) [Haykin 1998]. Existem alguns algoritmos que realizam o reconhecimento de padrões passando inicialmente por uma seção de treinamento, durante o qual se apresenta repetidamente um conjunto de padrões de entrada junto à categoria a qual cada padrão pertence. Na sequência, apresenta-se ao algoritmo um novo padrão que não foi utilizado antes, mas que pertence à mesma população de padrões utilizada para treinar o algoritmo. Ele deverá ser capaz de identificar a classe daquele padrão particular por causa da informação que ela extraiu dos dados de treinamento. Atualmente existem vários algoritmos de classificação de padrões, dentre eles iremos conhecer a Máquina de Vetor de Suporte, Naive Bayse e Discriminante Linear de Fisher.

2.3.1 Máquina de Vetor de Suporte

Máquinas de Vetores Suporte (SVMs) é um algoritmo de classificação de padrões introduzido por Vapnik & Cortes (1995) originalmente utilizadas para classificação de dados em duas classes, ou seja, na geração de dicotomias. O objetivo do SVM é elaborar uma forma computacional de criar hiperplanos de separação em um espaço de características de alta dimensão, onde esses hiperplanos otimizam os limites de generalização [Vapnik & Cortes 1995]. Para entendermos o funcionamento do SVM, considere um conjunto de treinamento linearmente separável (mostrado na Figura 2.8). Linearmente separável significa que é possível separar os padrões das classes diferentes por pelo menos um hiperplano.

A SVM constrói um hiperplano em um espaço dimensional elevada, o que pode ser usado para a classificação, regressão ou outras tarefas. Intuitivamente, uma boa separação é conseguida através do hiperplano que tem a maior distância com a aproximação de pontos de qualquer classe de dados de treino, uma vez que, em geral, quanto maior for a margem mais baixo o erro de generalização do classificador.

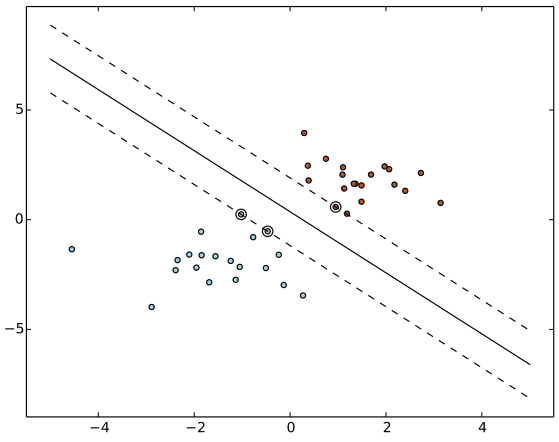

Figura 2.8: Plotando um hyperplano de separação de duas classes linearmente separáveis usando Máquina de Vetor de Suporte.

A Figura 2.8 ilustra um conjunto de dados formados por 2 classes separáveis (bolinhas pretas e brancas). Nesse conjunto de dados é utilizado o SVM para classificar as duas classes. Como pode ser observado, o SVM consegue criar o hiperplano de separação que permite distinguir as duas classes.

Para entendermos a matemática do algoritmo, considere a entrada $\mathbf{x}_i \in \mathbb{R}^p, i = (1,...,n)$ em duas classes, e o vetor $\mathbf{y} \in \{1,-1\}^n$ que atribuída a uma classe positiva, se f(x) > 0 e

atribuída a uma classe negativa caso contrário. Classificadores que separam os dados por meio de um hiperplano são denominados lineares, podendo ser definidos pela equação:

$$\mathbf{W}\,\mathbf{x}+b \;=\; 0 \tag{2.6}$$

onde $\mathbf{W}\,\mathbf{x}$ é o produto escalar entre os vetores $\mathbf{W}$ e $\mathbf{x}$, em que $\mathbf{W}$ é o vetor normal ao hiperplano e b é um termo "compensador". O par $(\mathbf{W}, b)$ é determinado durante o treinamento do classificador. Esta equação divide o espaço de entradas em duas regiões: $\mathbf{W}\,\mathbf{x}+b>0$ e $\mathbf{W}\,\mathbf{x}+b<0$, levando à equação:

$$\begin{cases} y_i = +1 & \text{se } \mathbf{W}\,\mathbf{x}+b>0 \\ y_i = -1 & \text{se } \mathbf{W}\,\mathbf{x}+b<0. \end{cases} \tag{2.7}$$

A interpretação geométrica deste tipo de hipótese é que o espaço de entrada $\mathbf{X}$ é dividido em duas partes pelo hiperplano definido pela equação $(\mathbf{W}\,\mathbf{x})+b=0$. Um hiperplano é um subespaço afim de dimensão $n-1$ que divide o espaço em duas metades que correspondem às entradas das duas classes distintas. Por exemplo, na Figura 2.8 o hiperplano é a linha escura, com a região positiva acima e a negativa abaixo. O vetor $\mathbf{W}$ define uma direção perpendicular ao hiperplano, enquanto variar o valor de b move o hiperplano paralelamente a ele mesmo.

Para se lidar com classes que não são linearmente separáveis, utiliza-se as funções *kernels*. A funções de *kernel* têm a finalidade de projetar os vetores de características de entrada em um espaço de características de alta dimensão para classificação de problemas que se encontram em espaços não linearmente separáveis [Haykin 1998]. Isso é feito, pois à medida que se aumenta o espaço da dimensão do problema, aumenta também a probabilidade desse problema se tornar linearmente separável em relação a um espaço de baixa dimensão. Pode-se encontrar na literatura *kernels* do tipo: polinomial, função de base radial e sigmoidal [Haykin 1998].

2.3.2 Naive Bayse

O algoritmo Naive Bayes, também chamado de classificador Bayesiano, é um aprendizado supervisionado com base na aplicação teorema de Bayes com o pressuposto de independência entre cada par de classes. O algoritmo tem como objetivo calcular a probabilidade que uma amostra desconhecida pertença a cada uma das classes possíveis. Este tipo de predição é chamado de classificação estatística, pois é completamente baseada em

probabilidades [Haykin 1998].

Essa classificação considera que o efeito do valor de variáveis sobre uma determinada classe é independente dos valores das outras variáveis. Apesar dessa simplicidade, Naive Bayes pode muitas vezes superam métodos de classificação mais sofisticados [Haykin 1998].

Para entendermos o algoritmo, considere uma classe variável y, um vetor de características dependente $[x_1,\ldots,x_n]$ e $P(y \mid X)$ sendo a probabilidade que uma classe seja y dado que a as características de entrada foram $\mathbf{X}$, o teorema de Bayes calcula as probabilidades das classes, através da seguinte relação:

$$P(y \mid x_1,\ldots,x_n) = \frac{P(y)P(x_1,\ldots x_n \mid y)}{P(x_1,\ldots,x_n)} \tag{2.8}$$

Usando a suposição que as características são independentes, temos:

$$P(y \mid x_1,\ldots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1,\ldots,x_n)} \tag{2.9}$$

Desde $P(x_1,\ldots,x_n)$ é constante dada a entrada, podemos usar a seguinte regra de classificação:

$$P(y \mid x_1,\ldots,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y) \tag{2.10}$$

$$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{n} P(x_i \mid y), \tag{2.11}$$

e podemos usar Estimativa Máxima A Posteriori para estimar $P(y)$ e $P(x_i \mid y)$, que forma a frequência relativa da classe y no conjunto de treinamento. Classificadores Naive Bayes diferem principalmente pelos pressupostos que fazem com relação à distribuição de $P(x_i \mid y)$. O Gaussian Naive Bayes assume que a probabilidade das características é uma função Gaussiana.

$$P(x_i \mid y) \quad = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \tag{2.12}$$

onde os parâmetros de variância (σ_y) e média (μ_y) são estimados usando máxima verossimilhança.

2.3.3 Discriminante Linear de Fisher

Discriminante Linear de Fisher (DLF) é um algoritmo que descreve uma transformação linear de um problema multidimensional em um problema unidimensional visando separar subgrupos de indivíduos. Enquanto a Análise de Componentes Principais (PCA) aplicada uma transformação nos dados que respondem por mais variância nos dados, a Discriminante Linear de Fisher utilizada uma transformação visando uma maior distinção entre classes [Haykin 1998].

Para entendermos o algoritmo, considere uma variável **y** formada por uma combinação linear dos elementos de um vetor de entrada **x**, isto é, ela é definida como o produto interno de x e um vetor de parâmetros ajustáveis w, como mostrado por:

$$\mathbf{y} = \mathbf{w}^T \mathbf{x} \tag{2.13}$$

o vetor **x** é retirado de duas populações, β_1 e β_2, que diferem entre si pelos seus vetores médios μ_1 e μ_2, respectivamente. O critério de Fisher para discriminar entre estas duas classes é definido por:

$$J(w) = \frac{w^T \mathbf{C}_b w}{w^T \mathbf{C}_t w} \tag{2.14}$$

onde $\mathbf{C}_b$ é a matriz de covariância entre classes (chamada de interclasses) definida por:

$$\mathbf{C}_b = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T \tag{2.15}$$

e $\mathbf{C}_t$ é a matriz de covariância no interior das classes (chamada de intraclasse) definida por:

$$\mathbf{C}_t = \sum_{n \in \beta_1} (x_n - \mu_1)(x_n - \mu_1)^T + \sum_{n \in \beta_2} (x_n - \mu_2)(x_n - \mu_2)^T \tag{2.16}$$

O objetivo é encontrar uma combinação linear adequada para salientar a estrutura de subgrupos será um vetor que minimize a variabilidade intraclasses e, ao fazê-lo, estará simultaneamente a maximizar a variabilidade interclasses [Haykin 1998].

Simplificando a Equação 2.14 temos:

$$w = \mathbf{C}_t^{-1}(\mu_1 - \mu_2) \tag{2.17}$$

que é referido como o discriminante linear de Fisher.

O ponto médio entre as duas médias populacionais univariadas μ_1 e μ_2 é:

$$m = \frac{1}{2}(\mu_1 - \mu_2)^T \mathbf{C}_t(\mu_1 + \mu_2) \tag{2.18}$$

A regra de classificação baseada na função discriminante de Fisher é:

$$\begin{cases} \mathbf{x}_i \in \beta_1 & \text{se } (\mu_1 - \mu_2)^T \mathbf{C}_t \mathbf{x}_i \geq m \\ \mathbf{x}_i \in \beta_2 & \text{se } (\mu_1 - \mu_2)^T \mathbf{C}_t \mathbf{x}_i < m \end{cases} \tag{2.19}$$

2.4 Gerenciamento dos Dados para Classificação

O gerenciamento de dados para classificação são técnicas que visam dividir o conjunto de dados em conjunto de treinamento e conjunto de teste. A motivação é validar o classificador com um conjunto de dados diferente daquele usado para estimar os parâmetros no treinamento. Se o conjunto de treinamento for uma amostra representativa do universo do problema, suas estimativas de desempenho para um conjunto de validação composto por exemplos não visto anteriormente podem ser muito boas. Caso contrário, o classificador poderá apresentar muitos erros de generalização durante os testes, seja por problemas de excesso de complexidade do classificador, que costuma causar *overfitting* [Haykin 1998].

Na sequência iremos conhecer as duas principais formas de realizar a divisão dos conjuntos de treinamento e validação, que são: *Cross Validation* e *Holdout*.

2.4.1 *Cross Validation*

No *Cross Validation* o conjunto de treinamento contendo N exemplos é dividido em k subconjuntos, onde $k > 1$, isto presume que k é divisível por N. O classificador é treinado com todos os subconjuntos, exceto um, e o erro de validação é medido testando-o com esse subconjunto não utilizado no treinamento. Este procedimentos é repetido para um total de k tentativas, cada uma usando um subconjunto diferente para a validação, como ilustrado na Figura 2.9. O desempenho do classificador é avaliado pela média do erro quadrado obtido na validação sobre todas as tentativas do experimento [Haykin 1998].

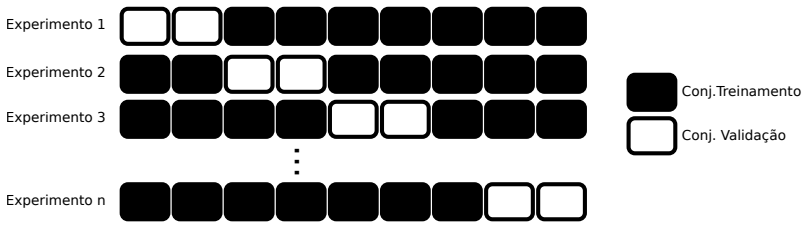

Figura 2.9: *Cross Validation* com o conjunto de treinamento e validação sendo separado para cada experimento.

2.4.2 Holdout

O holdout é o método mais simples para teste de classificadores. Nele, a base de dados é dividida em dois conjuntos: um conjunto de treinamento e conjunto de teste. O conjunto de treinamento fornece os dados para o treinamento e o conjunto de teste fornece dados novos para testar o classificador [Quilici-Gonzalez & de Assis Zampirolli 2014].

Sua vantagem é a simplicidade, mas deve-se ter um cuidado especial na representação das classes nos dois conjuntos para obter bons resultados. Em outras palavras, o resultado da avaliação depende, por exemplo, da quantidade de padrões existentes de cada classe em cada conjunto. Por exemplo, se for utilizado uma grande quantidade de uma determinada classe A e poucas da classe B para treinar o classificador, o mesmo terá uma melhor generalização para os dados da classe A do que para a classe B. Dessa forma, recomenda-se dividir as classes de forma equilibrada tanto no conjunto de treinamento quanto no de validação. Adotar a relação de 2/3 para o conjunto de treinamento e 1/3 para o conjunto de validação é uma prática comumente utilizada [Quilici-Gonzalez & de Assis Zampirolli 2014]. Uma limitação desse método está no fato de que menos exemplos são usados no treinamento, podendo ter impacto negativo no desempenho do classificador.

Figura 2.10: *Holdout* com o conjunto de treinamento e validação sendo separado para o único experimento.

2.5 Métodos de Validação dos Classificadores

A essência dos classificadores é codificar uma relação entre a entrada e saída (representado por um conjunto de exemplos rotulados) em parâmetros estatísticas intrínsecos do método. É esperado que o classificador torne-se bem treinado de modo que aprenda o

suficiente sobre o passado para generalizar no futuro. Ou seja, o processo de aprendizado se transforma em uma escolha de parametrização para um conjunto de dados. Métodos de validação são maneiras de determinar se o classificador foi bem treinado e aprendeu a relação entre entradas e saídas [Haykin 1998]. Entre esses métodos de validação destaca-se a Matriz de Confusão e o cálculo da Sensibilidade [Guimaraes 1985]. Esses dois métodos são os mais utilizados para na avaliação de classificadores de detecção de patologias médicas. Dessa forma, na sequência os mesmos serão apresentados.

2.5.1 Matriz de Confusão

Após a divisão do conjunto de dados e treinamento, é de grande importância fazer uma avaliação do desempenho do classificador. Em algumas aplicações o número total de erros não é um parâmetro adequado para analisar o desempenho de um sistema. Por exemplo, é preferível cometer o erro de classificar um não nódulo como sendo nódulo do que cometer o erro de classificar um nódulo como não nódulo, pois dessa forma poderá influenciar na decisão do radiologista de não investigar um paciente com a patologia. Para contornar esse problema, os resultados são descritos através de uma na matriz, chamada de Matriz de Confusão [Haykin 1998].

Para compreender o método, considere que nosso classificador esteja sendo usado para fazer um diagnóstico médico e que as respostas possíveis para este diagnóstico são: “Positivo” e “Negativo”. Dessa forma, quatro possibilidades de predição podem ocorrer:

- Se o paciente for portador de uma doença e o classificador acertar no diagnóstico, o caso é um Verdadeiro Positivo (VP);
- Se o paciente não for portador da doença e o classificador acertar no diagnóstico, o caso é um Verdadeiro Negativo (VN);
- Se o paciente for portador da doença, mas o classificador errar no diagnóstico indicando que ele está saudável, o caso é um Falso Negativo (FN);
- Se o paciente não for portador da doença, mas o classificador errar no diagnóstico indicando que ele está doente, o caso é um Falso Positivo (FP).

Essas quatro combinações de resultados estão representadas em uma matriz chamada de Matriz de Confusão, ilustrada na Tabela 2.1.

A Matriz de Confusão representa melhor os resultados de um sistema de classificação. Os elementos fora da diagonal podem ser somados para se obter o número de erros, enquanto que os elementos da diagonal podem ser somados para se obter o número de classificações corretas.

Tabela 2.1: Matriz de Confusão.

	Classificado com doente	Classificado como não doente
Doente	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Não doente	Falso Positivo (FP)	Verdadeiro Negativo (VN)

2.5.2 Sensibilidade

A Sensibilidade é a capacidade que um sistema tem de discriminar, dentre os suspeitos de uma patologia, aqueles efetivamente doentes [Guimaraes 1985]. Para isso, ele faz uma relação entre os Verdadeiros Positivos e todos os efetivamente doentes (que são os VP mais os FN). Esta propriedade é dada pela fórmula [Haykin 1998]:

$$\texttt{Sensibilidade} \;=\; \frac{VP}{(VP+FN)} \tag{2.20}$$

A Sensibilidade é o parâmetro mais utilizada para mensurar a qualidade de classificadores utilizados na detecção de câncer, pois esse parâmetro leva em consideração que o custo de um Falso Positivo é diferente de um Falso Negativo [El-Baz et al. 2013]. Por exemplo, pense nos danos, do ponto de vista da saúde pública, entre fornecer um falso diagnóstico positivo para um paciente saudável e um falso diagnóstico negativo para um paciente com um câncer tratável. Nesse caso é preferível errar com um Falso Positivo.

CAPÍTULO 3

MÉTODO PROPOSTO

Nas últimas décadas vêm surgindo vários avanços na área da Radiologia. Esses avanços ocorrem não só no desenvolvimento de novos equipamentos e técnicas, mas também no suporte tecnológico para auxiliarem os serviços. Um dos grandes avanços foi o desenvolvimento do sistema PACS, o que permitiu não só uma melhoria no sistema de arquivamento de imagens, como também aperfeiçoou o fluxo de trabalho nos serviços de Radiologia [Strickland 2000].

Atualmente um dos maiores desafios na pesquisa na área da Radiologia é o desenvolvimento de sistemas de detecção de patologias médicas. Esses sistemas deverão melhorar o desempenho dos radiologistas, reduzir o tempo necessário para o diagnóstico, ser integrado com o ambiente de trabalho da equipe médica e possuir baixo custo de implantação e utilização [van Ginneken et al. 2011]. Visando superar este desafio para a detecção de nódulos pulmonares, iremos propoe neste Capítulo um sistema inteligente, que será integrado ao OpenPACS, chamado de LCD-OpenPACS (*Lung Cancer Detection* - OpenPACS). O sistema irá auxiliar aos radiologistas através da realização de uma pré-análise semiautomática de exames de TC do tórax visando detectar nódulos pulmonares e, caso detectado, gerar alertas. Dessa forma, os radiologistas autorizados poderão ter acesso remoto aos exames originais e a pré-análise, através do OpenPACS, para realizar a sua tomada de decisão.

Visando melhor entendimento, o capítulo está subdividido nas seguintes seções. Na Seção 3.1 será mostrado o fluxo de trabalho da Radiologia em uma unidade de saúde que utiliza o OpenPACS e como o LCD-OpenPACS está incorporado a esse fluxo. Na Seção 3.2 é exibido a arquitetura do método e descrito a funcionalidade de cada módulo que ela contém. E, por último, na Seção 4.4 serão apresentadas as limitações do método proposto.

3.1 Contextualização

A geração de exames de imagens médicas normalmente não é o primeiro passo de um processo de diagnóstico, com exceções são casos de urgência ou de acidentes graves com pacientes inconscientes. Normalmente, a equipe de saúde questiona o paciente sobre os seus sintomas, histórico de doenças anteriores e realiza diagnóstico simples, como a palpação, auscultação com um estetoscópio e medidas de pressão arterial. Em alguns casos, faz-se necessário testes clínicos em que os líquidos do corpo, tais como sangue e urina, são analisados. A geração exames de imagens médicas normalmente é realizada quando as etapas de diagnóstico descritas anteriormente não foram suficientes para realizar o diagnóstico.

No fluxo de trabalho nos setores de Radiologia que utilizam o sistema OpenPACS a realização do exame de imagem é realizado pelos radiologistas ou técnicos de radiologia, dependendo da complexidade do exame. Na sequência, as imagens são automaticamente enviadas para o Servidor OpenPACS e um técnico administrativo entra em contato com o radiologista para analisar o exame e realizar o laudo. Caso o médico tenha alguma dúvida, ele solicita a realização de novos exames. De outro modo, o médico elabora o laudo. Ele é um documento técnico no qual o radiologista, com base na sua formação especializada e em experiências anteriores, descreve os elementos de um determinado exame de imagem e finaliza com a sua interpretação [S.Silva et al. 2014].

O método proposto é totalmente integrado ao ambiente de trabalho que utiliza o sistema OpenPACS, conforme ilustrado no diagrama presente na Figura 3.1. Será acrescentado uma aplicação para conectar ao servidor e verificará se existem novos exames de tomografia computadorizada do tórax. Caso afirmativo, esses exames irão ser enviados para o módulo de processamento do LCD-OpenPACS. Após o processamento, os resultados serão enviados para serem armazenados no servidor OpenPACS. O sistema ainda gera mensagens de alertas para a equipe médica caso algum tumor seja detectado.

O sistema proposto também permite a utilização da telerradiologia com auxílio ao diagnóstico de detecção de nódulos pulmonares. A telerradiologia consiste na transmissão eletrônica de imagens radiológicas de um local para outro com finalidades permitir um melhor acesso aos exames. Através da telerradiologia, na ausência do médico no hospital, as imagens do paciente e o resultado do LCD-OpenPACS podem ser visualizadas por computadores pessoais, com acesso a Internet, de usuários autorizados. Dessa forma, é possível reduzir o tempo necessário para o diagnóstico.

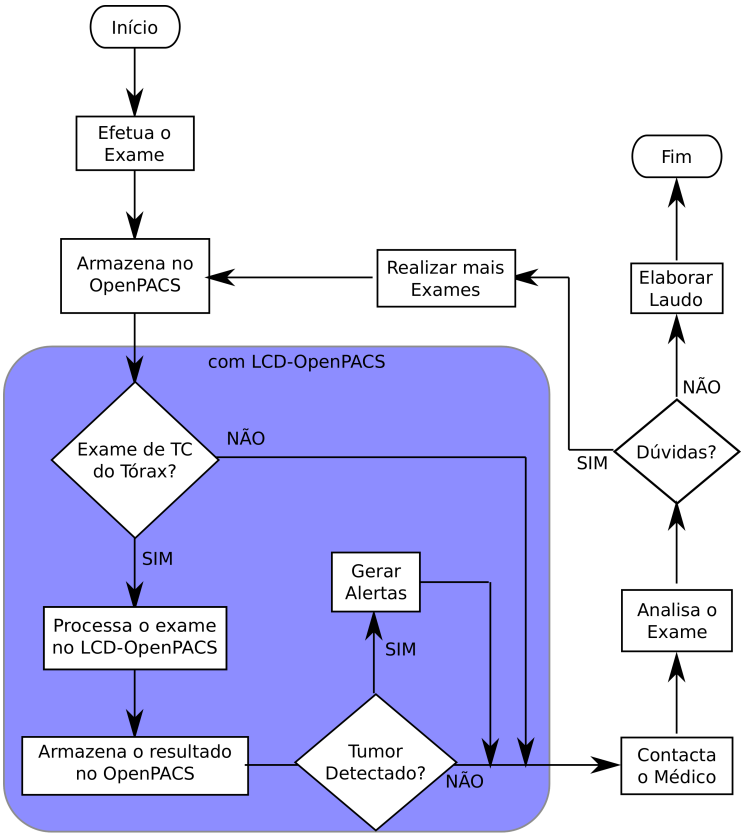

Figura 3.1: Fluxo de trabalho nos serviços de radiologia com a utilização do sistema LCD-OpenPACS.

3.2 Arquitetura

O LCD-OpenPACS é formado por sete módulos: Aquisição das Imagens, Segmentação, Detecção de Nódulos Suspeitos, Extração das características, Eliminação de Falsos Positivos, Envio dos Resultados e Geração de Alertas. Um diagrama esquemático das etapas que compõem o método é mostrado na Figura 3.2. Na sequência será apresentado detalhadamente cada módulo do sistema.

3.2.1 Aquisição das Imagens

As imagens radiologistas são geradas pelas Modalidades que tem a capacidade de descrever as funções anatômicas e fisiológicas do corpo humano. Nas unidades de saúde que utilizam o sistema OpenPACS, após a geração das imagens, as mesmas são enviadas para o Servidor. Uma vez as imagens armazenadas, o sistema proposto pode adquiri-las.

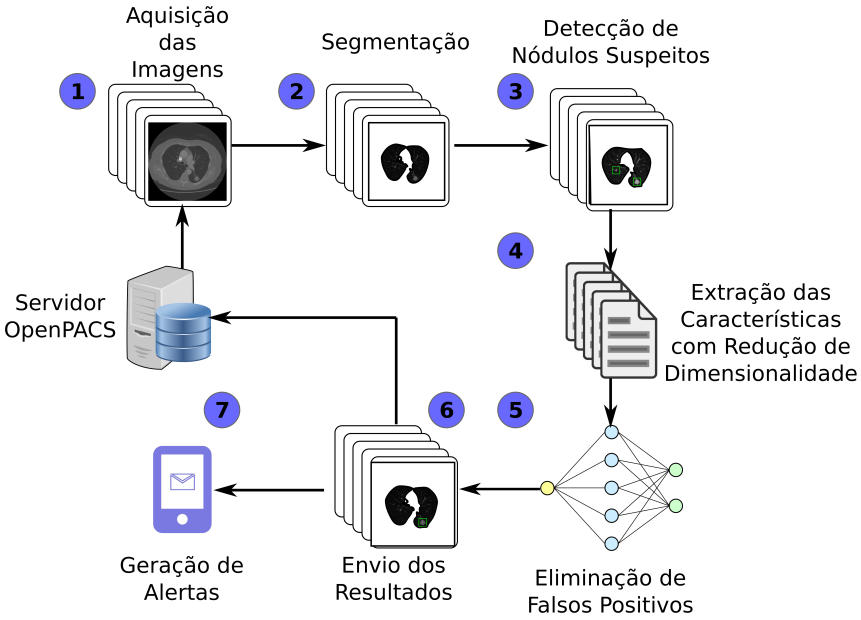

Figura 3.2: Diagrama dos módulos que compõem o sistema LCD-OpenPACS.

Para isso, o módulo de aquisição se conecta ao servidor comportando-se como um cliente DICOM. Ao detectar um novo exame é verificado se o mesmo trata-se de um exame de tomografia computadorizada do tórax, conforme mostrado no Algoritmo 1. Toda a consulta e transmissão das imagens do servidor OpenPACS e para o sistema LCD-OpenPACS se dá através do protocolo DICOM implementado pela biblioteca pynetdicom[1].

Algoritmo 1: Módulo de Aquisição das Imagens

```
Entrada: ([IP, Porta, AET], UV, TP) {Informações do Servidor OpenPACS, data da última verificação e tempo de pausa)
início
    C ← netdicom.applicationentity.AE(); {Cria um Cliente DICOM.}
    S ← (IP, Porta, AET); {Configura o Servidor OpenPACS.}
    se não C.RequestAssociation(S) então
        retorna (Erro! Servidor não localizado.)
    fim
    VA ← UV; {Tempo da Verificação Anterior.}
    repita
        T ← Tempo.Agora();
        NovosExames[] = C.ConsultarExames(T, VA); {Obtér Exames Entre o Período de Tempo [T, VA].}
        para Exame em NovosExames faça em paralelo
            se Exame.Modalidade = 'TC' & Exame.Descrição = 'Tórax' então
                C.Enviar(Exame);
            fim
        fim
        VA ← T;
        Pausar(TP); {Parar o programa por TP minutos.}
    até sempre;
fim
```

[1] http://pypi.python.org/pypi/pynetdicom

Para o funcionamento do módulo de Aquisição das Imagens, o usuário deverá informar as configurações do servidor OpenPACS ao qual se deseja conectar. Essas informações são o endereço IP, endereço da Porta e o identificador AET. Além disso, o usuário deverá informar quando foi realizada a última verificação e qual o tempo de pausa (em minutos) que o sistema deverá aguardar entre as verificações. Esse tempo de pausa se faz necessário para evitar de sobrecarregar a rede de dados hospitalar.

Conforme descrito no Algoritmo 1, o módulo irá se conectar ao servidor e ficar periodicamente verificando a existência de novos exames de tomografia do tórax. Caso detectado, o módulo envia uma cópia para o módulo de segmentação.

3.2.2 Segmentação

O módulo de segmentação é divido em dois submódulos, são eles: segmentação das imagens pulmonares e segmentação das estruturas pulmonares. Na sequência, serão apresentadas as funcionalidades de cada submódulo.

Segmentação das Imagens Pulmonares

A segmentação dos pulmões em imagens pode ser definida como um processo de delinear a extensão espacial dos pulmões que aparecem em imagens do tórax [Gonzales & Wintz 1987]. Esse processo é possível, em imagens de TC, pois os valores de atenuação gerada na tomografia refletem a densidade dos vários tecidos, conforme discutido na Seção 2.2. Baseado nisso, é proposto um novo método semiautomático para segmentar os pulmões em imagens de TC que combina algoritmo de Crescimento por Regiões e filtros morfológicos, conforme mostrado no Algoritmo 2.

No início do processo de segmentação, deverá ser fornecido ao módulo de segmentação, o exame de TC do tórax e dois pontos (chamados de sementes). Esses pontos correspondem a dois *pixels* que estão no interior do pulmão direito e esquerdo, observe a Figura 3.3 na etapa 1. Na sequência é aplicado um filtro de pré-processamento, chamado de *Curvature Flow*, para eliminar ruídos na imagem. Esse filtro trata-se de um algoritmo de diferenças finitas proposto por J.A. Sethian [Sethian 1999] e implementado por SimpleITK [2]. Além de eliminar os ruídos, a utilização desse filtro visa deixar a distribuição dos *pixels* mais uniforme, facilitando assim as etapas posteriores. SimpleITK é uma interface Python para a biblioteca ITK [BC et al. 2013]. O *Insight Toolkit* (ITK) é uma biblioteca de código fonte aberta e multi-plataforma que fornece um amplo conjunto de ferramentas para análise de imagem.

[2] `http://www.simpleitk.org`

Algoritmo 2: Módulo de Segmentação das Imagens

```
Entrada: (Exame, PS) {Exame de TC do Tórax e Pontos Sementes.}
Saída: (ImgSeg) {Imagens Segmentados dos Pulmões.}
início
    Imagens[] ← Ler(Exame);
    P[] ← Ler(PS);
    Imagens[] ← SimpleITK.CurvatureFlow(Imagens); {Filtro de Pré-processamento.}
    T ← 1; {Critério de Parada.}
    enquanto T < 6 faça
        ImgSeg[] ← SimpleITK.ConnectedThreshold(Imagens, P); {Segmentação por Crescimento por Regiões.}
        ImgSeg[] ← SimpleITK.BinaryMorphologicalClosingImageFilter(ImgSeg); {Filtros Morfológicos.}
        V = Volume(ImgSeg);
        se V > 3,5(10^6) então
            retorna (ImgSeg) {Imagens Segmentadas.};
        senão
            T ← T + 1; {Tentativas.}
            Mostrar(ImgSeg); {Mostrar ao Usuário o Resultado da Segmentação.}
            P = Ler(); {Solicita ao Usuário Novos Pontos Sementes.}
        fim
    fim
fim
```

Em seguida, é utilizado um algoritmo de segmentação baseado em Crescimento por Regiões 3D, chamado de *Connected Threshold* [BC et al. 2013]. Esse algoritmo agrupa os *voxels* vizinhos de acordo com a sua intensidade. Foram utilizados como limites de similaridade os valores: -1000 UH e -200 UH. Esses valores foram escolhidos, pois englobam os tecidos pulmonares, vasos pulmonares e o ar interno aos pulmões [Preim & Bartz 2007]. Nas imagens resultantes da segmentação de crescimento por regiões é comum o surgimento de pequenas estruturas não agrupadas, observe a Figura 3.3 na etapa 2. Para eliminar essas estruturas foi utilizado o filtro de fechamento morfológico que executa uma dilatação binária seguida por uma erosão. Como resultado, é criada uma máscara binária com *voxel* internos aos pulmões tendo o valor 1 e externos com o valor 0, observe a etapa 3 da Figura 3.3.

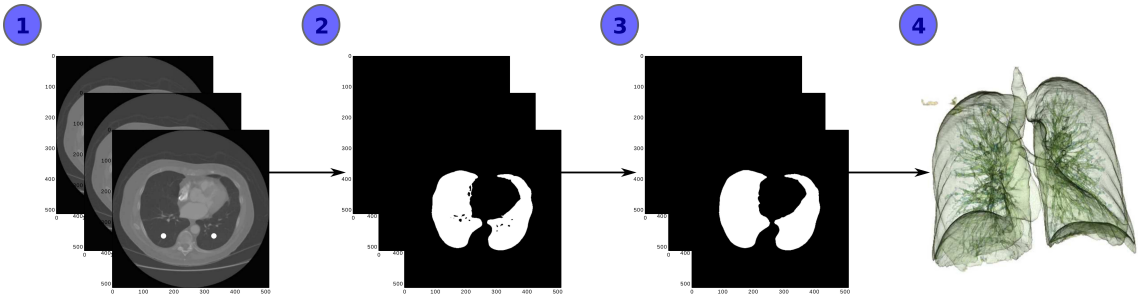

Figura 3.3: Principais etapas do processo de segmentação. (1) Definição dos pontos de sementes (pontos brancos), (2) Segmentação baseado em Crescimento por Regiões, (3) Aplicação de filtros morfológicos e (4) Reconstrução 3D do resultado final.

Essa máscara é utilizada para determinar o volume dos pulmões segmentados, isto é,

o número de *voxel* interno ao pulmão. Se o volume ficar maior do que o menor valor de volume pulmonar encontrado em testes experimentais (igual a $3,5(10^6)$ *voxels*), será aplicada a máscara na imagem original. Caso contrário, se o número de tentativas não atingir um limite (de cinco tentativas), será solicitado ao usuário que informe outros pontos de sementes para realizar uma nova tentativa de agrupamento de *voxels*. Por outro lado, se o volume da máscara ficar maior do que o liminar esperado, as imagens serão enviadas para o próximo módulo e uma imagem dos pulmões segmentados será mostrado ao usuário, observe a Figura 3.3 na etapa 4.

Segmentação das Estruturas Pulmonares

Após o processo de segmentação dos pulmões é realizado a segmentação das estruturas pulmonares internas. Nessa segmentação as estruturas internas (por exemplo, traquéia, brônquios e vasos pulmonares) são separadas, visando distinguir nódulos pulmonares das outras estruturas. Nessa segmentação foi utilizado a transformada *Watershed* proposta por Vincent e Soille [Vincent & Soille 1991] e implementada pela biblioteca SimpleITK. Esse método define uma função $f(x,y,z)$ para agrupar um conjunto de *voxels* que são mínimos locais, conforme mostrado na Seção 2.2. No método proposto foi utilizado como função $f(x,y,z)$ o cálculo da magnitude do gradiente mostrada na Equação 3.1 e 3.2.

$$f(x,y,z) = \sqrt{\left(\frac{\partial J}{\partial x}\right)^2 + \left(\frac{\partial J}{\partial y}\right)^2 + \left(\frac{\partial J}{\partial z}\right)^2} \tag{3.1}$$

$$J_w = I \odot \left(\frac{\partial}{\partial_w} \odot G\right) \tag{3.2}$$

onde: $\frac{\partial J}{\partial x}$ é a derivada parcial da função J em relação a x, I é a imagem 3D original e G é uma função gaussiana 3D.

Através dessa segmentação é possível agrupar os tecidos que possuem intensidades semelhantes permitindo que as estruturas pulmonares sejam separadas. O Algoritmo 3 mostra resumidamente o processo de segmentação utilizado pelo método proposto. Nele é possível observar que as imagens segmentadas são recebidas e o módulo da função gradiente é calculado através de uma função da biblioteca SimpleITK. Na sequência, é realizada a segmentação *Watershed* resultando em imagens formadas por rótulos, onde para cada estrutura distinguível é atribuído um rótulo. Como estamos interessados nas

estruturas internas, os rótulos correspondentes aos tecidos que formam os contornos pulmonares deverão ser eliminados. Essa subtração é realizada através da eliminação da maior estrutura rotulada.

Algoritmo 3: Módulo de Segmentação das Estruturas Pulmonares

Entrada: (ImgSeg) {Imagens Segmentadas de um Exame de TC do Tórax.}
Saída: (Rotulos) {Identificadores das várias estruturas pulmonares.}
início
    Imagens[] ← Ler(ImgSeg);
    ImgGrad[] ← SimpleITK.GradientMagnitudeRecursiveGaussian(Imagens); {Cálculo da função gradiente $f(x,y,z)$.}
    Rotulos[] ← SimpleITK.MorphologicalWatershed(ImgGrad); {Encontra a segmentação de *Watershed*.}
    Rotulos[] ← Ordenação(Rotulos); {Ordenando os rótulos encontrados na segmentação.}
    Rotulos[] ← Rotulos - max(Rotulos); {Eliminando os rótulos dos contornos dos pulmões.}
    retorna *Rotulos*
fim

A Figura 3.4 ilustra um exemplo de utilização do método proposto. Nela é possível observar uma reconstrução das imagens dos pulmões segmentados, uma reconstrução do agrupamento das estruturas pulmonares segmentadas pelo *Watershed* e uma reconstrução de um nódulo pulmonar. Nesse exemplo foram detectados 82 rótulos (estruturas distinguíveis) em toda a região pulmonar, destes um corresponde a um tumor pulmonar (mostrado na etapa 3 da Figura 3.4). Para a reconstrução 3D foi utilizada a biblioteca VTK[3]. VTK é uma biblioteca de código fonte aberta para computação gráfica, modelagem, processamento de imagem, renderização de volume e visualização científica.

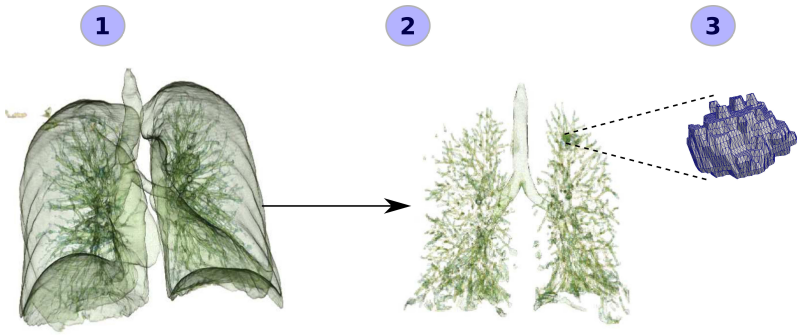

Figura 3.4: Principais etapas do processo de segmentação das estruturas pulmonares. (1) Reconstrução 3D dos pulmões segmentados, (2) Reconstrução 3D do agrupamento das várias estruturas internas, (3) Reconstrução 3D do tumor.

[3] http://www.vtk.org

3.2.3 Detecção de Nódulos Candidatos

Para entendermos o processo de detecção de nódulos candidatos, iremos inicialmente conhecer as características dos nódulos pulmonares. Esses nódulos normalmente surgem com forma esférica, mas com o crescimento tendem a perder um pouco essa forma e assumem configurações mais irregulares e com limites pouco definidos [Zamboni & Carvalho 2005]. Os nódulos podem ser classificados em: pequenos nódulos, nódulos justavascular, nódulos justapleurais e nódulos com opacidade em vidro fosco [El-Baz et al. 2013].

Pequenos nódulos representam os nódulos com diâmetro menor do que 5 mm. Nódulos justavascular referem-se nódulos que estão ligados aos vasos sanguíneos, enquanto que, nódulos justapleurais referem-se a casos que estão ligados à parede do parênquima ou diafragma. Nódulos com opacidade em vidro fosco refere-se a um tipo de nódulos onde os valores de intensidade dos *pixels* são significativamente inferiores aos dos nódulos sólidos e próximos aos valores dos tecidos pulmonares [Min et al. 2010]. A Figura 3.5 apresenta exemplos de nódulos com: (1) opacidade em vidro fosco, (2) nódulo justapleural, (3) pequeno nódulo e (4) nódulo justavascular.

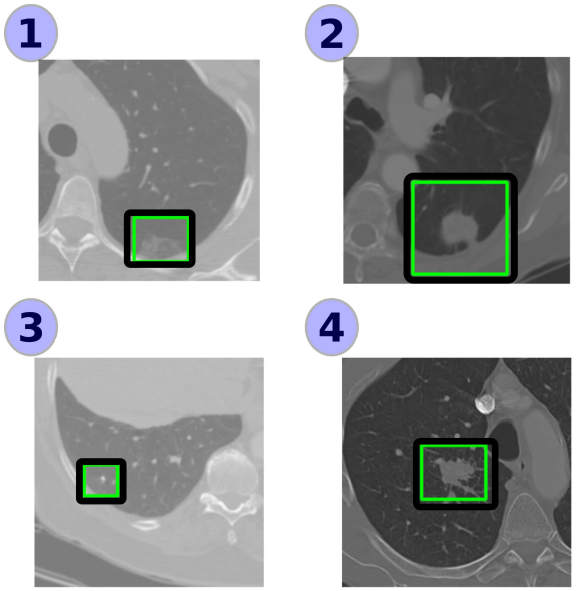

Figura 3.5: Exemplos de diferentes tipos de nódulos pulmonares. (1) nódulo com forma irregular e com opacidade em vidro fosco, (2) nódulo justapleural sólido com formato ovóide, (3) nódulo esférico sólido de 4 mm de diâmetro e (4) nódulo justavascular sólido em formato ovóide.

Sabendo que os nódulos tendem a serem esféricos ou semiesféricos e que outras estruturas pulmonares (por exemplo, vasos, traquéia e brônquios) tendem a ser cilíndricos, foi

utilizado um classificador baseado em regras para encontrar os nódulos candidatos. A primeira regra, chamada de *Roundness*, foi aplicada a estrutura segmentada visando detectar objetos esféricos ou semiesféricos, o seu cálculo é mostrada na Equação 3.3. Sempre que a *Roundness* for maior do que um limiar, o objeto segmentado é considerado como nódulo candidato.

$$(R1)Roundness = \frac{A_n(r)}{a} > 8,3(10^4) \tag{3.3}$$

onde: A_n é a área de uma hiperesfera (de raio r) que tem o mesmo volume do nódulo e a é a área do nódulo.

A segunda regra, chamada *Elongation*, visa detectar estruturas cilíndricas através do cálculo da relação de suas dimensões (comprimento, largura e profundidade) através do momento principal da imagem, Equação 3.4. Quanto mais cilíndrico for o objeto maior será o seu valor. Dessa forma, sempre que o *Elongation* for menor do que um limiar, o objeto é considerado nódulo candidato.

$$(R2)Elongation = \frac{MP_{max}}{MP_{min}} < 6,8(10^4) \tag{3.4}$$

onde: MP_{max} é o maior momento principal e MP_{min} é o menor momento principal.

A terceira regra é baseada na textura da imagem através do cálculo da energia. A energia, calculada através da matriz de co-ocorrência, expressa a uniformidade da textura na imagem visando eliminar regiões que não possuem nódulos. A energia é calculada pela Equação 3.5 e caso o valor medido seja inferior a um limiar, o objeto é considerado nódulo candidato. As equações R1, R2 e R3 são dadas pelas fórmulas [Lehmann 2007]:

$$(R3)Energia = \sqrt{\sum_{1}^{ns} \sum_{levels-1}^{i,j=0} P_{i,j}^2} < 3,3 \tag{3.5}$$

onde: *ns* é o número de *slices* que o nódulo aparece, *levels* corresponde ao valor de intensidade máxima na escala de cinza da imagem e $P_{i,j}$ é o histograma de co-ocorrência de nível de cinza da imagem.

O classificador baseado em regras foi utilizado a fim de remover rapidamente algumas estruturas que são Falsos Positivos facilmente distinguíveis (por exemplo, brônquios, traqueia, vasos pulmonares), de modo eliminar a influência dessas estruturas sobre as etapas posteriores. Neste trabalho, as regras foram projetadas com base no conhecimento adquirido com radiologistas, tais como as características anatômicas e morfológicas de nódulos pulmonares. Além disso, foi realizado um estudo estatístico das características morfológicas, intensidade e textura de nódulos pulmonares e demais estruturas nos exames de tomografia do tórax disponíveis na LIDC-IDRI [Armato SG 2011]. Baseado nesses estudos foi encontrado os valores de limiares utilizados nas regras. Entretanto, as regras foram definidas com critérios relativamente tênues para que eles não sejam específicos para o conjunto de dados utilizado.

O Algoritmo 4 mostra o processo de detecção de nódulos candidatos utilizado pelo método proposto. Esse módulo recebe como entrada as estruturas pulmonares segmentadas e utilizam filtros das bibliotecas *SimpleITK* e *Skimage*[4] para calcular as características das estruturas. Na sequência, é utilizado um classificador baseado em regras onde será eliminado as estrutura que não atenderem as regras. No final, restarão os nódulos candidatos. A biblioteca *Skimage* provê um conjunto de técnicas de processamento de imagem. Ela é escrita em *Python* e disponível de forma código fonte aberta.

Algoritmo 4: Módulo de Detecção de Nódulos Candidatos

```
Entrada: (ER) {Estruturas Rotuladas das Estruturas Pulmonares.}
Saída: (NodulosCandidatos) {Identificação e Localização de Nódulos Candidatos.}
início
    Estruturas[] ← Ler(ER);
    Filtro1 ← SimpleITK.LabelShapeStatisticsImageFilter();
    Filtro2 ← Skimage.feature();
    NodulosCandidatos ← []
    para cada estrutura em Estruturas faça
        R1 ← Filtro1.Roundness(estrutura);
        R2 ← Filtro1.Elongation(estrutura);
        R3 ← Filtro2.Energy(estrutura);
        se (R1 < 8,3(10^4)) & R2 < 6,8(10^4)) & R3 < 3,3) então
            NodulosCandidatos ← NodulosCandidatos + estrutura;
        fim
    fim
    retorna NodulosCandidatos
fim
```

[4]http://scikit-image.org

3.2.4 Extração das Características com Redução de Dimensionalidade

Na seção 3.2.3, foram selecionadas estruturas esféricas e semiesféricas. Entretanto, nem todas as estruturas que possuem essa forma são nódulos, por exemplo, alguns vasos com curvatura acentuada e que rapidamente muda de diâmetro, após segmentados, apresentou estruturas semelhantes aos nódulos. Dessa forma, essas estruturas similares deverão ser eliminadas. Para isso, os dados presentes nas imagens deverão ser transformados em um conjunto reduzido de características para serem analisadas. A extração das características das imagens é a etapa mais crítica para a identificação de objetos na imagem. Um dos principais problemas é como selecionar um conjunto reduzido de características que represente parte relevante da informação com precisão suficiente para sua identificação [da Silva 2007].

As características mais usadas para descrever uma imagem de modo sucinto são baseadas principalmente em distribuições de intensidades, textura e forma [Demir & Yener 2005]. Uma propriedade desejável para o extrator de características é que ele se comporte da mesma maneira para objetos de mesma natureza, e que esteja em posição, rotação e escalas diferentes (invariância a transformações geométricas). Ademais, os extratores deverão descrever adequadamente o objeto mesmo quando a imagem contém ruídos [da Silva 2007].

Devido o HoG [Dalal & Triggs 2005] apresentar essas propriedades, o método proposto utilizou o mesmo como extrator de características dos nódulos candidatos. Esse extrator ressalta as informações de aparência e forma dos objetos na imagem. Essas mesmas informações são utilizadas pelos radiologistas na analise de detecção de tumores em imagens de tomografia. Maiores informações sobre o extrator HoG pode ser encontradas na Seção 2.2.2.

Entretanto, o HoG apresentou o problema de alta dimensionalidade. Os histogramas encontrados nos nódulos candidatos ficaram com dimensões entre 77 e 2.380.848 unidades. Dessa forma, o número de elementos de treinamento requeridos, para que um classificador tenha um bom desempenho, deverá ser extremamente alta. Para contornar esse problema, é recomendável a utilização de técnicas para reduzir a dimensionalidade necessária para representar as características [Haykin 1998]. A redução da dimensionalidade também é benéfica uma vez que tende a reduzir o superajustamento (*overfitting*) [Haykin 1998]. Classificadores que aprenderam com o superajustamento são bons na reclassificação dos dados usados no treinamento, porém tendem a classificar incorretamente dados que ainda não foram vistos.

Desse modo, o módulo de extrator de características utiliza a Análise de Componente Principal (PCA) no HoG visando reduzir a dimensionalidade. O PCA é um método matemático que utiliza transformação ortogonal para converter um conjunto de variáveis, possivelmente correlacionadas, a um conjunto de valores de variáveis linearmente descorrelacionadas chamadas componentes principais. Maiores informações sobre o PCA pode ser encontradas na Seção 2.2.2.

O Algoritmo 5 apresenta um pseudocódigo da implementação do extrator de características com redutor de dimensionalidade. A entrada são as imagens dos nódulos candidatos segmentadas. O HoG é calculado a partir de cada fatia do imagem, depois um histograma resultante é gerado agrupando o HoG de cada fatia. Nos experimentos foi utilizado uma orientação de 27 *bins*, células formadas por 4x4 *pixels* e blocos formados por 2x2 células. Esses valores forma encontrados empiricamente. Na sequência, a dimensão do HoG é reduzida pelo método PCA mantendo 80% da variância dos dados originais. O PCA utilizado foi proposto por Thomas P Minka [Minka 2001] e implementado por Sklearn. Ele permite a utilização do PCA através da definição de um percentual mínimo de variância que deverá ser mantido, sem a necessidade de determinar previamente o número de componentes. Com o PCA o vetor de características resultantes ficou com a dimensão de 73 unidades. A biblioteca Sklearn[5] [Pedregosa et al. 2011] é uma formada por um conjunto de ferramentas, de código fonte aberta, voltadas para a análise e mineração de dados.

Algoritmo 5: Módulo de Extração de Características com Redução da Dimensionalidade

```
Entrada: (NC) {Imagens Segmentadas de um Determinado Nódulos Candidatos.}
Saída: (Caracteristicas) {Características dos Nódulos Candidatos.}
início
    Imagens[] ← Ler(NC);
    HoG ← [];
    para cada imagem em Imagens faça
        HoG ← HoG + skimage.feature.hog(imagem); {Calculando o HoG.}
    fim
    PCA ← sklearn.decomposition.PCA(0.8); {Definindo o PCA.}
    PCA.fit(HoG); {Ajustando o PCA.}
    Caracteristicas ← PCA.transform(HoG); {Aplicando o PCA para reduzir a dminensionalidade do HoG.}
    retorna Caracteristicas
fim
```

[5] http://scikit-learn.org

3.2.5 Eliminação de Falsos Positivos

Nessa etapa iremos eliminar os Falsos Positivos (FPs) remanescentes enquanto preservamos os Verdadeiros Positivos. No contexto de sistemas CADe, o termo Falso Positivo significa lesões que são identificados pelo algoritmo, mas não são nódulos. Típicos FPs encontrados foram: vasos com curvatura acentuada, vasos grossos com bifurcações, manchas geradas pelo movimento respiratório ou cardíaco, patologias infecciosas e cicatrizes no tecido parênquima.

Para eliminar os falsos positivos foi usado um classificador chamado de Máquinas de Vetor de Suporte (SVMs). O SVM é uma técnica embasadas na Teoria de Aprendizado Estatístico do tipo treinamento supervisionado capazes de generalizar problemas de classificação binária a partir de um conjunto de dados [Haykin 1998]. Maiores informações sobre o funcionamento do algoritmo SVM pode ser encontrado na Seção 2.3. Optou-se por utilizar o classificador SVM, pois o mesmo obteve melhores respostas quando comparado com outros classificadores, conforme irá ser mostrado no Capítulo 4. Além disso, Segundo Haykin (1998) o SVM apresenta melhores resultados quando aplicado à classificação de dados em duas classes (por exemplo, nódulo e não nódulo), ou seja, na geração de dicotomias. O SVM utilizado foi o *C-Support Vector Classification* da biblioteca Sklearn com *kernel* radial.

O Algoritmo 6 mostra um pseudocódigo do módulo de eliminação de Falsos Positivos. Como entrada do algoritmo deverá ser fornecido o exame contendo as imagens DICOM originais, uma estrutura de dados contendo as imagens segmentadas e as características dos nódulos candidatos, e por último, a estrutura SVM previamente treinada. Foi utilizada a biblioteca Joblib[6] para salvar e ler a estrutura de dados contendo o SVM. Joblib é um conjunto de ferramentas de código fonte aberta, em Python, para manipulação de dados e computação paralela.

Para cada nódulo candidato é verificado se o mesmo é um nódulo Verdadeiro Positivo através da classificação do SVM. Uma vez que, identificado como nódulo se faz necessário determinar sua localização. Para isso, foi utilizado o filtro StatisticsFilter da biblioteca SimpleITK. Uma vez localizados os nódulos, é necessário destaca-los na imagem original para que o radiologista possa analisar e tomar as duas decisões de tratamento e acompanhamento. Foi utilizada a biblioteca OpenCV[7] para destacar os nódulos nas imagens. OpenCV é uma biblioteca, de código fonte aberta, voltado para visão computacional e aprendizado de máquina.

[6] `https://pypi.python.org/pypi/joblib`
[7] `http://opencv.org/`

Algoritmo 6: Módulo de Eliminação de Falsos Positivos

```
Entrada: (Exame, NC, SVMFilename) {Exame Dicom Original, Estrutura de Dados Contendo as Imagem Segmentadas dos
         Nódulos Candidatos e suas Características Reduzida, e o Nome do Arquivo do SVM Previamente Treinado.}
Saída: (Imagens) {Imagens com os Nódulos Destacados.}
início
    Imagens[] ← Ler(Exame);
    SVM ← joblib.load('SVMFilename');
    para cada nodulo em NC faça
        Resultado ← SVM.predict(nodulo.Caracteristicas); {Determinar se é um Verdadeiro Positivo.}
        se Resultado == 1 então
            x0,xf,y0,yf,z0,zf = SimpleITK.StatisticsFilter.GetBoundingBox(nodulo.Imagem); {Determinar a
            Localização do Nódulo.}
            para cada z in [z0:zf] faça
                cv2.rectangle(Imagem[:,:,z], (x0,y0), (xf, yf), (0, 255, 0), 2); {Destacar os Nódulo nas Imagens.}
            fim
        fim
    fim
    retorna Imagens
fim
```

3.2.6 Envio dos Resultados

Uma vez os nódulos detectados, o próximo passo é enviar o exame para ser armazenado e distribuído pelo Servidor OpenPACS. Para isso, o módulo de envio dos resultados deverá receber as informações do Servidor OpenPACS para conexão e os exames com os nódulos destacados. O módulo irá estabelecer uma associação com o servidor e solicitar o armazenamento do exame, conforme mostrado no Algoritmo 7. Esse exame deverá ser anexado ao conjunto do exame original armazenado no Servidor através da inserção de uma nova série no mesmo exame. Dessa forma, o radiologista ao solicitar o exame irá ser enviado o conjunto de imagens originais juntamente com as imagens processadas pelo LCD-OpenPACS. Todas as transmissões das imagens do sistema LCD-OpenPACS para o servidor OpenPACS se dá através do protocolo DICOM implementado pela biblioteca pynetdicom[8].

Algoritmo 7: Módulo de Envio dos Resultados

```
Entrada: ([IP, Porta, AET], Exame) {Informações do Servidor OpenPACS, e Exame com os Nódulos Destacados)
início
    C ← netdicom.applicationentity.AE(); {Cria um Cliente DICOM.}
    S ← (IP, Porta, AET); {Configura o Servidor OpenPACS.}
    A ← C.RequestAssociation(S) {Tenta Estabelecer Associação com o Servidor.}.
    se não A então
        retorna (Erro! Servidor não localizado.)
    fim
    A.SCU(Exame);
fim
```

[8] `http://pypi.python.org/pypi/pynetdicom`

3.2.7 Geração de Alertas

O módulo de geração de alertas é responsável por controlar o envio de mensagens entre os usuários cadastrados e o sistema LCD-OpenPACS. A ideia principal é que a equipe médica possa receber informações de pacientes com possíveis nódulos pulmonares detectados pelo LCD-OpenPACS. As mensagens serão enviadas aos usuários previamente cadastrados através da emissão de alarmes para dispositivos (móveis ou não), podendo, conforme configurações pré-definidas, serem enviadas para tela de *desktop*, por e-mail, por SMS e outros.

A Figura 3.6 ilustra um exemplo de uma mensagem de alerta enviado via SMS para um dispositivo móvel de um usuário do sistema. Através dessa mensagem, o usuário ficará sabendo que foram identificados nódulos pulmonares em um determinado paciente em uma determinada unidade de saúde. É informado ainda o número do prontuário e o identificador do exame.

Figura 3.6: Exemplo de mensagem de alerta de SMS para um dispositivo móvel.

CAPÍTULO 4

RESULTADOS E DISCURSÕES

O câncer de pulmão é responsável por mais de 1,5 milhão de óbitos por ano [OMS 2015]. Essa alta taxa de mortalidade está ligada a dificuldade inerente da detecção de nódulos pulmonares. Sistemas CADe vêm sendo desenvolvidos para auxiliar os radiologistas na detecção visando diminuir essa taxa. O objetivo desse Capítulo é avaliar um sistema CADe, chamado de LCD-OpenPACS, proposto na presente tese. Para isso será analisado a precisão dos seus principais módulos que são: segmentação e detecção de nódulos (união dos módulos de extração de características e eliminação de falsos positivos). A partir dessas avaliações poderemos inferir se o método proposto apresenta as características necessárias para utilização em um ambiente hospitalar.

Nesse Capítulo, a Seção 4.1 apresentará os materiais que foram manipulados nos experimentos de validação. A Seção 4.2 traz os relatos dos resultados obtidos nos experimentos. Finalmente, na Seção 4.3 é apresentado uma comparação dos resultados obtidos pelo método com outros métodos existentes na literatura.

4.1 Materiais

O banco de dados utilizado nos experimentos de validação consistiu de 220 exames de diferentes pacientes, obtidos aleatoriamente do LIDC-IDRI (*Lung Image Database Consortium*). O banco de dados está disponível publicamente no *Cancer Imaging Archive* (TCIA)[1]. Atualmente, esse banco é composto por 1.010 exames, de tomografias computadorizadas do tórax, coletados em diferentes equipamentos e com distintos parâmetros

[1] `http://www.cancerimagingarchive.net/`

de configuração (por exemplo, espessura do corte, tamanho do *pixel* e número total de fatias). Todos os nódulos foram avaliados por quatro radiologistas experientes. Mais detalhes sobre o banco de dados, tais como os métodos e protocolos usados para adquirir dados de imagem e o processo de anotação das lesões pode ser encontrado em Armato *et al.* (2011).

Os 220 exames, de 220 pacientes, utilizados continham 296 nódulos que foram diagnosticados consensualmente por pelo menos dois radiologistas. Foram encontrados pacientes com até cinco nódulos. O tamanho do nódulo variou de 3 mm a 30 milímetros, podendo ser de câncer primário de pulmão, doença metastática, nódulo benigno ou de natureza indeterminada. A base de dados testada era formada por nódulos isolados, justapleurais, justavascular, pequenos nódulos e nódulos com opacidade em vidro fosco.

4.2 Resultados

O objetivo dessa seção é avaliar o desempenho do sistema LCD-OpenPACS na segmentação (pulmões e nódulos) e na detecção de nódulos pulmonares. O desempenho foi avaliado através de testes experimentais com exames de pacientes reais. Na sequência serão apresentados os resultados obtidos nos testes.

4.2.1 Segmentação

Após o desenvolvimento do sistema, os módulos de segmentação dos pulmões e segmentação das estruturas pulmonares foram submetidos a testes para validação com os exames da base da dados. O módulo de segmentação dos pulmões, que é baseado em Crescimento por Regiões (apresentado no Capítulo 3, seção 3.2.2), foi capaz de segmentar 96,9% dos exames de tomografia. Nos casos onde não foi obtido sucesso satisfatório, a principal causa está relacionada a pacientes que possuíam patologias graves, que alteravam a opacidade dos contornos pulmonares.

A Figura 4.1 ilustra um exemplo segmentação de um exame. Esse exame possui 261 imagens DICOM, cada imagem contendo 512x512 *pixels*. Após o processamento de segmentação dos pulmões, foram separadas algumas imagens resultante, conforme mostrado na Figura 4.1, juntamente com as suas respectivas imagens originais acima. Nessa figura, a etapa 1 corresponde a parte superior do tórax, onde é possível visualizar parte dos pulmões e parte da traqueia. Por outro lado, a etapa 5 mostra a parte inferior do tórax, onde é possível visualizar a parte inferior dos pulmões.

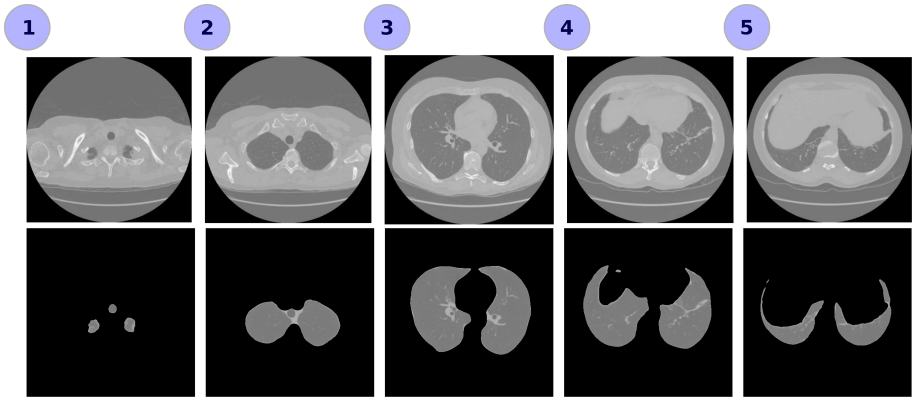

Figura 4.1: Exemplo de segmentação dos pulmões em um exame de TC pelo sistema LCD-OpenPACS.

O módulo de segmentação das estruturas pulmonares (apresentado no Capítulo 3, seção 3.2.2) tem como principal objetivo a distinção entre nódulos e as demais estruturas pulmonares, através da segmentação das imagens. Consequentemente, a nossa avaliação só considera se o módulo obteve êxito na separação dos nódulos, ou seja, na sua segmentação. Dessa forma, o módulo de segmentação das estruturas internas, baseada na técnica *Watershed*, foi capaz de segmentar 82,1% dos nódulos presentes nos exames de validação. Não foi obtido sucesso satisfatório em alguns pequenos nódulos (menores do que 5 mm) e com opacidade em vidro fosco.

A Figura 4.2 ilustra um exemplo segmentação de um nódulo pulmonar em um exame de validação. Na imagem é possível visualizar os pulmões segmentados (acima) e os nódulos segmentados (abaixo). Esse exame possui 135 imagens DICOM, cada imagem contendo 512x512 *pixels*. Foi separado na Figura 4.2 algumas imagens resultantes do processo de segmentação do nódulo. Esse nódulo possui dimensão de 26,5 mm e apareceu em nove imagens consecutivas.

4.2.2 Detecção dos Nódulos

Uma vez os nódulos candidatos segmentados, o processo de detecção dos nódulos foi submetido a testes de validação. A detecção dos nódulos é formada pelos módulos de extração das características e eliminação de falsos positivos (apresentados no Capítulo 3, seção 3.2.4 e 3.2.5). Os principais critérios utilizados para essa avaliação em sistemas CAD de detecção (CADe) são: taxa de Falso Positivo e Sensibilidade. Falso Positivo (FP) representa os resultados positivos quando a amostra não apresenta a doença, enquanto que a Sensibilidade é a capacidade que um sistema tem de discriminar dentre os suspeitos de

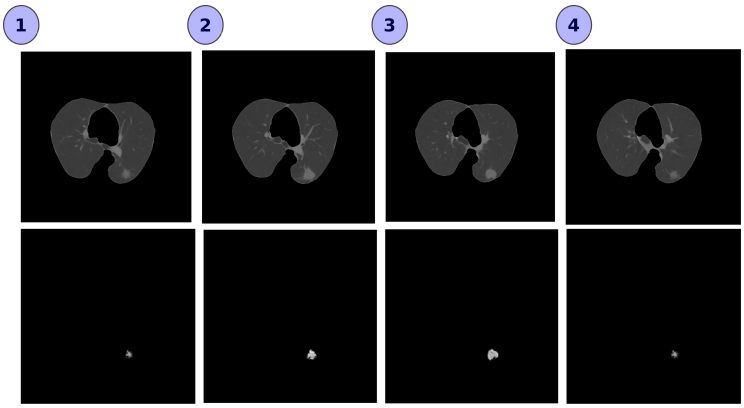

Figura 4.2: Exemplo de segmentação de um nódulo pulmonar em um exame de TC pelo sistema LCD-OpenPACS.

uma patologia, aqueles efetivamente doentes. Maiores informações sobre esses critérios podem ser encontradas na Seção 2.5.

Os métodos utilizados para validar a capacidade de generalização dos módulos foram o *Cross Validation* e *Holdout*. A eficácia do SVM foi verificada através de uma análise comparativa com os métodos DLF (Discriminante Linear de Fisher) e Naive Bayes. Utilizamos as implementações do DLF e Naive Bayes da biblioteca Sklearn. Os três classificadores foram utilizados tendo como entrada o HoG após a sua redução realizada pelo PCA. Discriminante Linear de Fisher é um classificador supervisionado que projeta dados de alta-dimensional para o espaço unidimensional para a classificação [Hastie et al. 2008]. Naive Bayes é um classificador supervisionado baseado na aplicação do teorema de Bayes com estimação MAP (máxima a posteriori) [Zhang 2004]. Maiores informações sobre esse métodos pode ser obtidas na Seção 2.3. Na sequência serão apresentados os resultados de testes de validação.

Cross Validation

Inicialmente, foi realizada a validação utilizando o método de gerenciamento de dados *Cross Validation*. Esse método, apresentado na Seção 2.5, divide aleatoriamente os dados original em K subconjuntos de mesmo tamanho. Desses K subconjuntos, $K-1$ são usados para treinamento e 1 subconjunto é utilizado para os testes. Este procedimento é repetido até que todos os k subconjuntos sejam usados nos testes. O objetivo da repetição é aumentar a confiabilidade na determinação da precisão do classificador. Nos testes realizados, foi utilizado $k = 10$. Esse método é comumente usado na avaliação dos sistemas CADe [Suzuki et al. 2003*b*] [Messay et al. 2010] [Tan et al. 2011] [Cascio et al. 2012]. Os resultados dos testes mostram que o SVM obteve melhor desempenho apresentando

uma Sensibilidade de 94,4% com uma taxa de FP de 7,04 por caso, observe a Tabela 4.1.

Tabela 4.1: Comparação de desempenho dos classificadores para detecção de nódulos pulmonares usando a validação *Cross Validation*.

Classificador	Sensibilidade	FP
DLF	89,2%	6,89
Naive Bayes	93,9%	7,03
SVM	94,4%	7,04

Holdout

Na sequência, foi realizado a comparação de desempenho dos métodos com o *Holdout*. O *Holdout* é uma técnica estatística utilizada para determinar, durante o treinamento, a capacidade de generalização dos classificadores [Haykin 1998]. O conjunto de dados é dividido aleatoriamente em dois grupos distintos, um usado para treinar e um utilizado para validar. Este método é também utilizado em validação de sistemas CAD [Gurcan et al. 2002] e [Matsumoto et al. 2006]. Nos testes, foi adotado: o subconjunto de treinamento contendo 154 exames de tomografia computadorizada (com 207 nódulos) e um subconjunto de validação contém 66 exames (com 89 nódulos). A eficácia da SVM foi verificada através da comparação com DLF e Naive Bayes, ambos implementados pela biblioteca Sklearn. Os resultados são mostrados na Tabela 4.2. Através dessa Tabela, pode-se inferir que o desempenho do SVM foi o melhor, apresentando uma Sensibilidade de 93,9%, com uma taxa de FP igual a 7,21 por caso.

Tabela 4.2: Comparação de desempenho dos classificadores para detecção de nódulos pulmonares usando a validação *Holdout*.

Classificador	Sensibilidade	FP
DLF	88,99%	7,47
Naive Bayes	90,47%	7,44
SVM	93,9%	7,21

A Tabela 4.3 ilustra a Matriz de Confusão obtida nos testes de validação do classificador DLF na avaliação *Holdout*. Nessa matriz é possível verificarmos que o número de VP, FN, FP e VN foram respectivamente: 79; 10; 7,47; e 17,8.

A Tabela 4.4 mostra a Matriz de Confusão obtida nos testes de validação do classificador Naive Bayes na avaliação *Holdout*. Através dela podemos determinar que o número VP, FN, FP e VN foram respectivamente: 81; 8; 7,44; e 15,2.

Tabela 4.3: Matriz de Confusão do Classificador DLF na validação *Holdout*.

	Nódulo Detectado	Não Nódulo Detectado
Nódulo	79	10
Não Nódulo	7,47 (p/ exame)	17,8 (p/ exame)

Tabela 4.4: Matriz de Confusão do Classificador Naive Bayes na validação *Holdout*.

	Nódulo Detectado	Não Nódulo Detectado
Nódulo	81	8
Não Nódulo	7,44 (p/ exame)	15,2 (p/ exame)

Por último, na Tabela 4.5 é apresentado a Matriz de Confusão resultante do processo de validação do método com o classificador SVM. Como pode ser visualizado, o número de VP, FN, FP e VN foram respectivamente: 84; 5; 7,21; e 15,7.

Tabela 4.5: Matriz de Confusão do Classificador SVM na validação *Holdout*.

	Nódulo Detectado	Não Nódulo Detectado
Nódulo	84	5
Não Nódulo	7,21 (p/ exame)	15,7 (p/ exame)

4.3 Discussões

A presente tese propõe um novo sistema CADe, chamado LCD-OpenPACS, que foi desenvolvido para atender a quatro requisitos: (a) melhorar o desempenho dos radiologistas, (b) reduzir o tempo necessário para o diagnóstico, (c) ser perfeitamente integrados com o ambiente de trabalho da equipe médica, e (d) apresentar custos insignificantes, ou uma redução nos custos hospitalares que justifiquem a sua implantação.

O LCD-OpenPACS foi modelado visando a sua completa integração com o sistema OpenPACS e com o fluxo de trabalho da equipe médica que a utiliza. Além disso, o sistema pode ser integrado com qualquer outro sistema PACS que tenha suporte DICOM. Com relação ao custo, o método proposto foi desenvolvido com bibliotecas de código fonte abertas e com licenças que permitam a disponibilização do LCD-OpenPACS gratuitamente para o Sistema Único de Saúde (SUS).

Para mensurarmos o desempenho do método proposto, foi realizada uma comparação relativa do nosso método (usando SVM com os resultados obtidas na validação *Cross Validation*) com outros métodos existentes na literatura. Essa comparação é mostrada na Tabela 4.6. Nessa tabela, cada linha representa uma sistema CADe de detecção de

nódulos pulmonares em exames de TC. Para cada sistema é apresentado à Sensibilidade obtida, número de Falsos Positivos e números de nódulos usados para validação. Com base nesta comparação, pode-se inferir que o desempenho do sistema proposto está entre os melhores em relação à Sensibilidade com 94,4% e 7,04 FP/caso. Enquanto que a taxa média da Sensibilidade obtida pelos radiologistas sem a utilização de sistemas CADe é de aproximadamente 77% [Jeon et al. 2012*b*]. Dessa forma, acreditamos que o desempenho dos radiologistas pode ser melhorado com a utilização do método proposto.

Tabela 4.6: Comparação de desempenho de métodos de detecção de nódulos pulmonares através da Sensibilidade, FP e número de nódulos obtidas na validação *Cross Validation*.

Métodos	Sensibilidade	FP	Nº Nódulos
Armato *et al*. [**?**]	70%	9,6/caso	187
Lee *et al*. [Lee et al. 2001]	72%	25,3/caso	98
Suzuki *et al*. [Suzuki et al. 2003*b*]	80,3%	4,8/caso	121
Murphy *et al*. [Murphy et al. 2007]	84%	8,2/caso	268
Ye *et al*. [Ye et al. 2009]	90,2%	8,2/caso	220
Messay, H. e Rogers [Messay et al. 2010]	82,66%	3/caso	143
Cascio *et al*. [Cascio et al. 2012]	97%	6,1/caso	148
Teramoto e Fujita [Teramoto & Fujita 2013]	80%	4,2/caso	103
Han *et al*. [Han et al. 2015]	82,7%	4/imagem	490
Método Proposto	94,4%	7,04/caso	296

Os resultados experimentais realizados com um conjunto de dados, não pertencentes a base de dados utilizado nos testes, demonstraram a generalização do método proposto. Entretanto, o sistema não detecta nódulos pulmonares menores do que 3 mm e não deverá ser utilizado nos casos que ocorre a presença de patologias graves, que alteram a opacidade dos contornos pulmonares.

Com relação ao tempo necessário para o diagnóstico, o tempo de processamento do sistema por exame foi de aproximadamente 10 minutos, o que não acarretaria em atrasos significativos no diagnóstico. Além disso, o como sistema LCD-OpenPACS permite que a equipe médica tenha acesso aos resultados remotamente e envia alertas de detecção. Dessa forma, acreditamos que o método proposto pode diminuir o tempo necessário para o diagnóstico.

4.4 Limitações

O sistema não deverá ser utilizado em pequenos nódulos pulmonares (menores do que 3 mm) e não foi testado para nódulos maiores do que 30 mm. Além disso, a etapa de

segmentação ainda não apresentou resultados satisfatórios para segmentação das imagens pulmonares nos casos que ocorre a presença de patologias graves, que alteram a opacidade dos contornos pulmonares, tais como enfisemas pulmonares. Outra limitação do método é a necessidade de um operador que auxilie na segmentação dos pulmões (fornecendo os pontos de sementes) e recebendo os alertas dos resultados do sistema.

CAPÍTULO 5

CONCLUSÕES

O câncer de pulmão é a quinta maior causa de morte no mundo, apresentando uma taxa de aproximadamente 1,6 milhão de óbitos por ano. Normalmente, o câncer de pulmão é detectado em estágios avançados. Um dos motivos é a dificuldade inerente na sua detecção. Sistemas CADe vêm sendo desenvolvidos visando auxiliarem os radiologistas na detecção de nódulos pulmonares. Entretanto, as soluções propostas ainda não são amplamente utilizadas na prática clínica.

A presente tese realizou um estudo bibliográfico visando detectar os problemas das atuais soluções. Como resultado, foi notado que os atuais sistemas não apresentavam quatro requisitos básicos, são eles: (a) melhorar o desempenho dos radiologistas, (b) reduzir o tempo necessário para o diagnóstico, (c) ser perfeitamente integrado com o ambiente de trabalho da equipe médica, e (d) apresentarem custos insignificantes ou que justifiquem a sua implantação.

A partir do entendimento do problema, foi modelado e implementado um novo sistema CADe para detecção de nódulos pulmonares em exames de tomografia computadorizada visando a sua disponibilização para as unidades básicas de saúde pertencentes ao Sistema Único de Saúde Brasileiro. Esse novo sistema, chamado de LCD-OpenPACS, foi submetido a testes de validação com exames de pacientes reais. O módulo de segmentação das imagens pulmonares foi capaz de segmentar 96,9% dos exames. O módulo de segmentação das estruturas pulmonares foi capaz de segmentar 82,1% dos nódulos. O módulo de detecção dos nódulos obteve uma Sensibilidade de 94,4% com 7,04 FP por exame. Enquanto que a taxa média de detecção obtida pelos radiologistas é de aproximadamente 77%.

O sistema foi desenvolvido de forma integrada com o fluxo de trabalho da equipe

médica e para ser disponibilizado gratuitamente para o SUS, visando reduzir os custos de sua implantação. Com relação ao requisito de redução do tempo necessário para o diagnóstico, o sistema apresentou um tempo de processamento de aproximadamente 10 minutos por exame.

Baseado nos resultados experimentais e na concepção do sistema, acreditamos que o mesmo tem potencial para ser utilizado no cotidiano clínico das unidades de saúde pertencentes ao SUS. Ele pode ser utilizado em triagens de pacientes de risco e auxiliando o trabalho dos radiologistas na detecção de nódulos pulmonares em exames de tomografia computadorizada.

5.1 Trabalhos Futuros

Como trabalhos futuros, deixamos como sugestão o desenvolvimento de melhorias no módulo de segmentação das imagens pulmonares. Essas melhorias devem permitir a segmentação automatizada, incluir pacientes com patologias que alterem a opacidade do contorno pulmonar e a detecção de nódulos menores do que 3 mm. Além disso, realizar um estudo de caso da aplicação do sistema proposto em uma unidade de saúde e analisar o desempenho em um ambiente real. Outros trabalhos podem analisar diferentes técnicas ou realizar modificações no módulo proposto visando melhorar o desempenho. Por último, poderá ser desenvolvido um módulo no LCD-OpenPACS para mensurar características dos nódulos e avaliar a evolução do tratamento oncológico com possível prognóstico.

REFERÊNCIAS BIBLIOGRÁFICAS

ACS, American Cancer Society (2011), *Global Cancer Facts and Figures, 2nd edition*, American Cancer Society Inc, Atlanta.

Armato, S. G., G. McLennan, M.F. McNitt-Gray, C.R. Meyer, D. Yankelevitz, D.R. Aberle, C.I. Henschke, E.A. Hoffman, E.A. Kazerooni, H. MacMahon, A.P. Reeves, B.Y. Croft, L.P. Clarke & For the Lung Image Database Consortium Research Group (2004), 'Lung image database consortium: Developing a resource for the medical imaging research community1', *Radiology* **232**(3), 739–748.

Armato, S.G., M.L. Gieger, C.J. Moran, J.T. Blackburn, K. Doi & H. Macmahan (1999), 'Computerized Detection of Pulmonary Nodules on CT scans', *Radiographics* **19**(5), 1303–11.

Armato SG, et al. (2011), 'The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans.', *Medical Physics* **38**(2), 915–931.

Audigier, Romaric Matthias Michel (2004), 'Segmentacao e Visualizacao Tridimensional Interativa de Imagens de Ressonancia Magnetica.'. Dissertacao (Mestre em Engenharia Eletrica), Faculdade de Engenharia Eletrica e de Computacao da Universidade Estadual de Campinas, Brasil.

Banta, H. David (1990), 'Future health care technology and the hospital', *Health Policy* **14**(1), 61–73.

Barros, Joao Adriano, Geraldo Valladares, Adriane Reichert Faria, Erika Megumi Fugita, Ana Paula Ruiz, Andre Gustavo Daher Vianna, Guilherme Luas Trevisan & Fabricio

Augusto Martinelli de Oliveira (2006), 'Diagnostico precoce do cancer de pulmao: o grande desafio. Variaveis epidemiologicas e clinicas, estadiamento e tratamento', *Jornal Brasileiro de Pneumologia* **32**, 221 – 227.

BC, Lowekamp, Chen DT, Ibanez L & Blezek D (2013), 'The design of simpleitk', *Frontiers in Neuroinformatics* **7**, 45.

Beare, R. & G. Lehmann (2006), 'The watershed transform in itk - discussion and new developments'.

Camarlinghi, Niccolo, Ilaria Gori, Alessandra Retico, Roberto Bellotti, Paolo Bosco, Piergiorgio Cerello, Gianfranco Gargano, Ernesto Lopez Torres, Rosario Megna, Marco Peccarisi & MariaEvelina Fantacci (2012), 'Combination of computer-aided detection algorithms for automatic lung nodule identification', *International Journal of Computer Assisted Radiology and Surgery* **7**(3), 455–464.

Cascio, D., R. Magro, F. Fauci, M. Iacomi & G. Raso (2012), 'Automatic detection of lung nodules in ct datasets based on stable 3d mass-spring models', *Comput. Biol. Med.* **42**(11), 1098–1109.

CFM (2002), Resolucao cfm n 1.643/2002 - define e disciplina a prestacao de servicos atraves da telemedicina, Relatório técnico, Conselho Federal de Medicina, Brasilia - DF.

Chan, HP, K Doi, CJ Vyborny, RA Schmidt, CE Metz, KL Lam, T Ogura, YZ Wu & H. MacMahon (1990), 'Improvement in radiologists' detection of clustered microcalcifications on mammograms. the potential of computer-aided diagnosis.', *Invest Radiol.* **25**(10), 1102–1112.

da Silva, Carolina Yukari Veludo Watanabe (2007), Extracao de caracteristicas de imagens medicas utilizando wavelets para mineracao de imagens e auxilio ao diagnostico, Tese de doutorado, USP â Sao Carlos, Sao Carlos - SP.

Dalal, N. & B. Triggs (2005), Histograms of oriented gradients for human detection, *em* 'Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on', Vol. 1, pp. 886–893 vol. 1.

Demir, Cigdem & Bulent Yener (2005), *Automated cancer diagnosis based on histopathological images: a systematic survey*, Technical report, Rensselaer Polytechnic Institute, Department of Computer Science.

Doi, Kunio, Heang-Ping Chan & Maryellen L. Giger (1990), 'Method and system for enhancement and detection of abnormal anatomic regions in a digital image'.

Duncan, J.S. & N. Ayache (2000), 'Medical image analysis: progress over two decades and the challenges ahead', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(1), 85–106.

El-Baz, Ayman, Garth M. Beache, Georgy Gimel'farb, Kenji Suzuki, Kazunori Okada, Ahmed Elnakib, Ahmed Soliman & Behnoush Abdollahi (2013), 'Computer-aided diagnosis systems for lung cancer: Challenges and methodologies', *International Journal of Biomedical Imaging* **2013**, 1–46.

Fantacci, M.E., N. Camarlinghi, I. Gori, R. Bellotti, G. Gargano, R. Megna, E.L. Torres, P. Cerello, C. Peroni, I. De Mitri, G. De Nunzio & A. Retico (2011), Algorithms for automatic detection of lung nodules in ct scans, *em* 'Medical Measurements and Applications Proceedings (MeMeA), 2011 IEEE International Workshop on', pp. 623–627.

Firmino, Macedo, Antonio H. Morais, Roberto M. Mendoca, Marcel R. Dantas, Helio R. Hekis & Ricardo Valentim (2014), 'Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects.', *Biomedical engineering online* **8**(1), 13–41.

Firmino, Macedo, Ricardo Valentim, Marcel Ribeiro & Leila Cavalcanti (2013), 'Openpacs - sistema open source para comunicacao e arquivamento de imagens medicas: Relato de experiencia em um hospital universitario - doi: 10.3395/reciis.v7i2.sup1.772pt', *RECIIS* **7**(2).
URL: *http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/772*

Giger, Maryellen Lissak, Kunio Doi & Heber MacMahon (1988), 'Image feature analysis and computer-aided diagnosis in digital radiography. 3. automated detection of nodules in peripheral lung fields', *Medical Physics* **15**(2), 158–166.

Gomatrhi, M. & P. Thangara (2010), A computer aided diagnosis system for detectionof lung cancer nodules using extreme learning machine, *em* 'International Journal of Engineering Science and Technology', Vol. 2, pp. 5770–5779.

Gonzales, Rafael C. & Paul Wintz (1987), *Digital Image Processing (2Nd Ed.)*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Guimaraes, M. Carolina S. (1985), 'Exames de Laboratorio: Sensibilidade, Especificidade, Valor Preditivo Positivo', *Revista da Sociedade Brasileira de Medicina Tropical* **18**(2), 117–120.

Gurcan, M.N., B. Sahiner, N. Petrick, Chan H.P., E. A. Kazerooni, P.N. Cascade & L. Hadjiiski (2002), 'Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system.', *Medical Physics* **28**, 2552–2558.

Han, Hao, Lihong Li, Fangfang Han, Bowen Song, W. Moore & Zhengrong Liang (2015), 'Fast and adaptive detection of pulmonary nodules in thoracic ct images using a hierarchical vector quantization scheme', *Biomedical and Health Informatics, IEEE Journal of* **19**(2), 648–659.

Hastie, Trevor, Robert Tibshirani & Jerome Friedman (2008), *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2nd[a] edição, Springer, Stanford, California.

Haykin, Simon (1998), *Neural Networks: A Comprehensive Foundation*, 2nd[a] edição, Prentice Hall PTR, Upper Saddle River, NJ, USA.

INCA (2014), Estimativa 2014: Incidencia de cancer no brasil, Relatório técnico, Instituto Nacional de Cancer Jose Alencar Gomes da Silva, Coordenacao de Prevencao e Vigilancia, Rio de Janeiro.

Jain, Anil K. (1989), *Fundamentals of Digital Image Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jeon, Kyung Nyeo, Jin Mo Goo, Chang Hyun Lee, Youkyung Lee, Ji Yung Choo, Nyoung Keun Lee, Mi-Suk Shim, In Sun Lee, Kwang Gi Kim, David S. Gierada & Kyongtae T. Bae (2012*a*), 'Computer-aided nodule detection and volumetry to reduce variability between radiologists in the interpretation of lung nodules at low-dose screening computed tomography', *Investigative Radiology* **47**, 457â461.

Jeon, Kyung Nyeo, Jin Mo Goo, Chang Hyun Lee, Youkyung Lee, Ji Yung Choo, Nyoung Keun Lee, Mi-Suk Shim, In Sun Lee, Kwang Gi Kim, David S. Gierada & Kyongtae T. Bae (2012*b*), 'Computer-aided nodule detection and volumetry to reduce variability between radiologists in the interpretation of lung nodules at low-dose screening computed tomography', *Investigative Radiology* **47**, 457â461.

Kazuo, Awai, Murao Kohei, Ozawa Akio, Komi Masanori, Hayakawa Haruo, Hori Shinichi & Nishimura Yasumasa. (2004), 'Pulmonary nodules at chest ct: effect of computer-aided diagnosis on radiologists' detection performance.', *Radiology* **230**, 347–352.

Kumar, S.A., J. Ramesh, P. T. Vanathi & K. Gunavathi (2011), Robust and automated lung nodule diagnosis from ct images based on fuzzy systems, *em* 'Process Automation, Control and Computing (PACC), 2011 International Conference on', pp. 1–6.

Law, Maria Y. Y. & Zheng Zhou (2003), 'New direction in PACS education and training', *Computerized Medical Imaging and Graphics* **27**(2-3), 147–156.

Lee, Yongbum, Takeshi Hara, Hiroshi Fujita, Shigeki Itoh & Takeo Ishigaki (2001), 'Automated detection of pulmonary nodules in helical ct images based on an improved template-matching technique', *IEEE Trans. Medical Imaging* **20**, 595–604.

Lehmann, G. (2007), 'Label object representation and manipulation with ITK', *The Insight Journal* **08**.
URL: *http://hdl.handle.net/1926/584*

Liu, Yang, Jinzhu Yang, Dazhe Zhao & Jiren Liu (2010), A method of pulmonary nodule detection utilizing multiple support vector machines, *em* 'Computer Application and System Modeling (ICCASM), 2010 International Conference on', Vol. 10, pp. V10–118–V10–121.

Matsumoto, Sumiaki, Harold L. Kundel, James C. Gee, Warren B. Gefter & Hiroto Hatabu (2006), 'Pulmonary nodule detection in {CT} images with quantized convergence index filter', *Medical Image Analysis* **10**(3), 343 – 352. Special Issue on The Second International Workshop on Biomedical Image Registration (WBIRâ03).

Messay, Temesguen, Russell C. Hardie & Steven K. Rogers (2010), 'A new computationally efficient {CAD} system for pulmonary nodule detection in {CT} imagery', *Medical Image Analysis* **14**(3), 390 – 406.

Min, Ji Hye, Ho Yun Lee, Kyung Soo Lee, Joungho Han, Keunchil Park, Myung-Ju Ahn & Su-Jin Lee (2010), 'Stepwise evolution from a focal pure pulmonary ground-glass opacity nodule into an invasive lung adenocarcinoma: An observation for more than 10 years', *Lung Cancer* **69**(1), 123 – 126.

Minka, T. P. (2001), 'Automatic choice of dimensionality for PCA', *Advances in Neural Information Processing Systems* **15**, 598–604.

Murphy, K., A. Schilham, H. Gietema, M. Prokop & B. van Ginneken (2007), 'Automated detection of pulmonary nodules from low-dose computed tomography scans using a two-stage classification system based on local image features', *Proc. SPIE* **6514**, 651410–651410–12.

NIH (2015), Seer cancer statistics review, 1975-2011, NIH Specification Acessado em: 2015-07-21, U.S. National Institutes of Health.
URL: *http://seer.cancer.gov/archive/csr/1975_2011/*

OMS (2015), International agency for research on cancer. globocan 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012, OMS Specification Acessado em: 2015-07-21, Organizacao Mundial da Saude.
URL: *http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx?cancer=lung*

Orozco, H.M., O. Osiris Vergara Villegas, L.O. Maynez, V.G.C. Sanchez & H. de Jesus Ochoa Dominguez (2012), Lung nodule classification in frequency domain using support vector machines, *em* 'Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on', pp. 870–875.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Pianykh, Oleg S. (2008), *Digital Imaging and Communications in Medicine: A Practical Introduction and Survival Guide*, 1ª edição, Springer Publishing Company, Incorporated.

Preim, Bernhard & Dirk Bartz (2007), Chapter 03 - acquisition of medical image data, *em* B. P.Bartz, ed., 'Visualization in Medicine', The Morgan Kaufmann Series in Computer Graphics, Morgan Kaufmann, Burlington, pp. 35 – 64.

Quilici-Gonzalez, Jose Artur & Francisco de Assis Zampirolli (2014), *Sistemas Inteligentes e MineraÃ§Ã£o de Dados*, Triunfal Grafica e Editora, Santo Andre - SP.

Schulze, Otto Carl, Jaco Greyling, Hofmeyr Viljoen & Savvas Andronikou (2007), 'Talking pacs: Part 2 - why should we change to pacs?', *South African Journal of Radiology* pp. 86 – 90.

Sethian, J.A. (1999), *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK.

Shao, Hong, Li Cao & Yang Liu (2012), A detection approach for solitary pulmonary nodules based on ct images, *em* 'Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on', pp. 1253–1257.

S.Silva, C.Isabela, Giuseppe D'Ippolito & Antonio Jose da Rocha (2014), *Oncologia, Serie Colegio Brasileiro de Radiologia e Diagnostico por Imagem*, Elsevier, Rio de Janeiro - RJ.

Strickland, N H (2000), 'Pacs (picture archiving and communication systems): filmless radiology.', *Archives of Disease in Childhood* **83**(1), 82–86.
URL: *http://adc.bmj.com/cgi/doi/10.1136/adc.83.1.82*

Suzuki, Kenji (2012), 'A review of computer-aided diagnosis in thoracic and colonic imaging', *Quantitative Imaging in Medicine and Surgery* **2**(3), 163–176.

Suzuki, Kenji, Samuel G. Armato III, Feng Li, Shusuke Sone & Kunio Doi (2003*a*), 'Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography', *Medical Physics* **30**(7), 1602–1617.

Suzuki, Kenji, Samuel G. Armato III, Feng Li, Shusuke Sone & Kunio Doi (2003*b*), 'Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography', *Medical Physics* **30**(7), 1602–1617.

Tan, Maxine, Rudi Deklerck, Bart Jansen, Michel Bister & Jan Cornelis (2011), 'A novel computer-aided lung nodule detection system for ct images', *Medical Physics* **38**(10), 5630–5645.

Teramoto, Atsushi & Hiroshi Fujita (2013), 'Fast lung nodule detection in chest ct images using cylindrical nodule-enhancement filter', *International Journal of Computer Assisted Radiology and Surgery* **8**(2), 193–205.

van Ginneken, Bram, Cornelia M. Schaefer-Prokop & Mathias Prokop (2011), 'Computer-aided diagnosis: How to move from the laboratory to the clinic', *Radiology* **261**(3), 719–732.

Vapnik, Vladimir & Corinna Cortes (1995), 'Support-vector networks', *Mach. Learn.* **20**(3), 273–297.

Vincent, L. & P. Soille (1991), 'Watersheds in digital spaces: an efficient algorithm based on immersion simulations', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **13**(6), 583–598.

Walker, J. (2006), *Lung Cancer: Current and Emerging Trends in Detection and Treatment*, Cancer and modern science, Rosen Publishing Group.
URL: *https://books.google.com.br/books?id=0VfKAKwDIW4C*

Xu, Xin-Wei, Kunio Doi, Takeshi Kobayashi, Heber MacMahon & Maryellen L. Giger (1997), 'Development of an improved cad scheme for automated detection of lung nodules in digital chest images', *Medical Physics* **24**(9), 1395–1403.

Ye, Xujiong, Xinyu Lin, J. Dehmeshki, G. Slabaugh & G. Beddoe (2009), 'Shape-based computer-aided detection of lung nodules in thoracic ct images', *Biomedical Engineering, IEEE Transactions on* **56**(7), 1810–1820.

Zamboni, Mauro & Walter Roriz De Carvalho (2005), *Cancer do Pulmao*, Atheneu, Sao Paulo - SP.

Zhang, Harry (2004), The Optimality of Naive Bayes., *em* V.Barr & Z.Markov, eds., 'FLAIRS Conference', AAAI Press.