UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA

DAVID JONES FERREIRA DE LUCENA

Atualização local automática de pesos para recuperação de nódulos similares de câncer pulmonar

Maceió
2016

David Jones Ferreira de Lucena

Atualização local automática de pesos para recuperação de nódulos similares de câncer pulmonar

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Prof. Dr. Marcelo Costa Oliveira
Coorientador: Prof. Dr. Aydano Pamponet Machado

Maceió
2016

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

L935a Lucena, David Jones Ferreira de.
Atualização local automática de pesos para recuperação de nódulos similares de câncer pulmonar / David Jones Ferreira de Lucena. – 2016.
82 f. : il.

Orientador: Marcelo Costa Oliveira.
Coorientador: Aydano Pamponet Machado.
Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2006.

Bibliografia: f. 78-81.
Apêndice: f. 82.

1. Sistemas de recuperação da informação. 2. Conteúdo – Recuperação. 3. Suporte a tomada de decisão. 4. Algoritmo – Peso de atributo. 5. Câncer de pulmão – Diagnóstico por imagens. I. Título.

CDU: 004.8

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP 57 072-970 | Telefone: (082) 3214-1401

Membros da Comissão Julgadora da Dissertação de Mestrado de David Jones Ferreira de Lucena, intitulada: "*Atualização local automática de pesos para recuperação de nódulos similares de câncer pulmonar*", apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 12 de fevereiro de 2016, às 14h00min, na Sala 15 do Instituto de Computação da UFAL.

COMISSÃO JULGADORA

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

Prof. Dr. Aydano Pamponet Machado
UFAL – Instituto de Computação
Coorientador

Prof. Dr. Evandro de Barros Costa
UFAL – Instituto de Computação
Examinador

Prof. Dr. Marcelo Zanchetta do Nascimento
UFU – Universidade Federal de Uberlândia
Examinador

CONFERE COM
15 MAR. 2016
O ORIGINAL.

Floripes Teixeira Santos
Secretaria - IC/UFAL
Mat. SIAPE 1673751

Aos meus pais por todo o suporte dado a mim durante a formação e pelo incentivo para que eu continuasse sempre com a cabeça erguida.
À minha amada esposa pelo seu carinho e compreensão em todos os momentos difíceis.

AGRADECIMENTOS

A Deus primeiramente.

Ao meu orientador Marcelo Oliveira por mais uma orientação em seu laboratório e pela sua paciência em me ajudar a corrigir as minhas falhas.

Ao meu co-orientador Aydano Machado por sua orientação e pelas vezes em que me deu força para continuar na vida acadêmica.

A todos os colegas do laboratório LaTIM pelas grandes contribuições neste trabalho, em especial ao José Raniery.

A minha esposa e meus pais pelo apoio que recebi todo o tempo quando o desânimo e a incerteza batiam à porta.

A todos os colegas que cursaram as matérias da grade curricular comigo, com eles aprendi muito.

A todos os professores do PPGI.

Àqueles que ajudaram direta ou indiretamente para que eu pudesse concluir este trabalho.

Por fim, à Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

RESUMO

O câncer de pulmão se tornou a neoplasia maligna mais letal do mundo nas últimas décadas. E, apesar dos avanços na medicina, houve pouco progresso com relação à cura da doença. Segundo o INCA, na última estimativa mundial sobre a incidência de câncer pulmonar, em 2012, foram registrados 1,82 milhão de casos de câncer, sendo 1,24 milhão entre os homens e 583 mil entre as mulheres. O principal causador do câncer pulmonar é o tabagismo sendo responsável por 90% dos casos diagnosticados. O diagnóstico do câncer pulmonar é feito, principalmente, com base em imagens de TC e, hoje, é considerada a principal técnica de visualização para detecção de nódulos pulmonares. Entretanto, o processo de identificação e classificação de nódulos é complexo e envolve fatores subjetivos e qualitativos que acabam induzindo os especialistas ao erro. Este panorama exige o emprego de técnicas computacionais que permitam efetivamente manipular os dados e proporcionar meios para diagnósticos mais precisos. Sistemas computacionais têm sido desenvolvidos com o objetivo de buscar e recuperar imagens de exames já diagnosticados, que são similares a um novo caso com patologia ainda desconhecida segundo a similaridade entre as suas características. Essa propriedade é intrínseca aos sistemas CBIR. Os exames diagnosticados recuperados podem ser utilizados como uma segunda opinião para guiar os especialistas no momento do diagnóstico, fornecendo informações adicionais. Contudo, CBIR apresenta algumas limitações referentes ao processo de extração e representação de características das imagens, por meio de atributos, e a determinação de uma métrica de similaridade adequada. Este trabalho apresenta um algoritmo de ajuste local de pesos aplicado à DEP em uma arquitetura CBIR com o objetivo de verificar se a DEP com os pesos ajustados é mais precisa do que a DE na recuperação de imagens contendo nódulos de câncer pulmonar. Para isso, foram utilizados os AT 3D e os ANB 3D para representar os nódulos. O processo apresentado é composto por duas fases que são executadas de forma sequencial e cíclica sendo uma Fase de Avaliação e uma de Fase de Treinamento. A cada iteração os pesos são ajustados segundo os nódulos recuperados. Ao término do ciclo de execuções das fases, obtém-se um conjunto de pesos de atributos que otimizam a recuperação de nódulos semelhantes. Os resultados alcançados pela atualização dos pesos foram promissores aumentando a precisão em 10% e 6% em média para recuperação de nódulos benignos e malignos, respectivamente, com revocação de 25%. No melhor caso, o ANB 3D proporcionou 100% para recuperação das duas classes com revocação de 90%. Isso comprova a eficácia do algoritmo alcançando os objetivos almejados para o trabalho e confirmando a hipótese de que a DEP com os pesos ajustados proporciona maior precisão do que DE como métrica de similaridade em sistemas CBIR.

Palavras-chaves: Recuperação baseada em conteúdo; recuperação de informação; suporte à tomada de decisão; atualização de pesos de atributos; câncer pulmonar.

ABSTRACT

Lung cancer has become the most lethal malignancy in the world in recent decades. And despite advances in medicine, there has been little progress regarding the cure of the disease. According to the National Cancer Institute in the last global estimate of the incidence of lung cancer in 2012, there were 1.82 million cases of cancer, with 1.24 million among men and 583 thusand among women. The main cause of lung cancer is smoking that is responsible for 90 % of diagnosed cases. The diagnosis of lung cancer is done mainly based on CT images, and today it is considered the main visualization technique for detecting pulmonary nodules. However, the process of identifying and classification of nodules are complex and involves subjective and qualitative factors that lead experts to error. This scenario requires the use of computational techniques to effectively manipulate the data and provide the means for more accurate diagnoses. Computer systems have been developed in order to search and retrieve imaging exams already diagnosed which are similar to a new case with unknown pathology according to the similarity between their characteristics. This property is intrinsic to Content-Based Image Retrieval (CBIR). Diagnosed exams retrieved can be used as a second opinion to guide those specialists in the diagnosis, providing more information. However, CBIR presents some limitations regarding to the process of segmentation and representation of image characteristics through of attributes, as well as determine an appropriate similarity metric. This paper presents a local update weighing algorithm applied to the Weighted Euclidean Distance (WED) in a CBIR architecture in order to verify if the WED with adjusted weights is more accurate than the Euclidean Distance in image retrieval of pulmonary nodules. For this, the 3D Texture Attributes (3D AT) and 3D Margin Sharpness Attributes (3D MSA) were used to represent nodules. Presente process consists of two phases that are performed sequentially and cyclically being an Assessment Phase and Training Phase. At each iteration the weights are adjusted according to the retrieved nodules. At the end of cycles execution, it is obtained a set of attribute weights that optimize the recovery of similar nodes. The results achieved by updating the weights were promising and increase precision by 10% to 6% on mean for recovery of benign and malignant nodules respectively with recall 25%. In the best case, the 3D MSA provided 100% of precision for the two classes with recall 90%. This proves the effectiveness of the algorithm achieving the goals to this work and confirms the hypothesis that the DEP, with adjusted weights, provides greater precision than DE as a similarity metric in CBIR systems.

Keywords: Content-based image retrieval; information retrieval; decision support; update weighing attributes; lung cancer.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AI Atributos integrados

ANB Atributos de Nitidez de Borda

AT Atributos de Textura

AUC do inglês *Area Under Curve*

CAD do inglês *Computer-Aided Diagnosis*

CADe do inglês *Computer-Aided Detection*

CADx do inglês *Computer-Aided Diagnosis*

CBIR do inglês *Content-Based Image Retrieval*

DE Distância Euclidiana

DEP Distância Euclidiana Ponderada

DICOM do inglês *Digital imaging and communications in medicine*

DP Desvio Padrão

FDA do inglês *Food and Drug Administration*

FDM Família de Distâncias *Minkowski*

FNIH do inglês *Foundation for The National Institutes of Health*

HU *Hounsfield*

IDP Inverso do Desvio Padrão

INCA Instituto Nacional do Câncer

KNN K-vizinhos mais próximos

LIDC do inglês *Lung Image Database Consortium*

LOO do inglês *Leave-One-Out*

LWL do inglês *Locally Weighted Learning*

LWR do inglês *Locally Weighted Regression*

MCO Matriz de Coocorrência

NCI do inglês *National Cancer Institute*

NDCG do inglês *Normalized Discounted Cumulative Gain*

NoSQL do inglês *Not only Structured Query Language*

PN Precisão(n)

PR Precisão $\times$ Revocação

RBF do inglês *kernel Radial Basis Function*

RD Radiografia

SVM do inglês *Support Vector Machine*

TC Tomografia Computadorizada

WA do inglês *Weighted Average*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

O câncer de pulmão se tornou a neoplasia maligna mais letal do mundo nas últimas décadas. Contudo, apesar dos avanços na medicina, houve pouco progresso com relação à cura da doença (LIMA; PIMENTA, 2015).

Segundo o Instituto Nacional do Câncer (INCA), em 2012, na última estimativa mundial sobre a incidência de câncer pulmonar, foram registrados 1,82 milhão de casos da doença, sendo 1,24 milhão entre os homens e 583 mil entre as mulheres. No Brasil, em 2014 foram contabilizadas quase 23 mil mortes devidas a esta doença, e foram estimados aproximadamente 28 mil novos casos, sendo 17 mil entre homens e 11 mil entre mulheres (INCA, 2015). Os índices de mortalidade são similares em magnitude aos da incidência do câncer devido à alta fatalidade da doença, ou seja, a maioria dos casos diagnósticados acaba em óbito do paciente (WENDER et al., 2013).

O principal causador do câncer pulmonar é o tabagismo. O INCA aponta que 90% dos casos diagnosticados tem como causa esta prática (INCA, 2015). Assim, a melhor forma de combate a esta doença é o estímulo para que os fumantes parem de fumar e que outras pessoas não venham a fazer uso do cigarro (ZAMBONI, 2002).

O diagnóstico do câncer pulmonar é feito principalmente com base em imagens de Tomografia Computadorizada (TC). Ela é considerada por autores como (GONZALEZ; WOODS, 2008), (BUSHBERG; BOONE, 2011) e (DICIOTTI et al., 2010) como a principal ferramenta de visualização para a detecção de nódulos pulmonares. O seu uso na rotina radiológica permite alcançar resultados efetivos na redução de casos de morte pela doença devido à detecção dos nódulos em estágios iniciais, quando têm menos de 3mm de diâmetro (PASTORINO, 2010; TEAM, 2011).

Entretanto, o uso da TC não pode ser considerado uma panaceia, dado que existem limitações que envolvem o sistema computacional (entendido como o conjunto de *hardware* e *software*) e o especialista (usuário). As limitações do sistema podem estar relacionadas as formas de visualização e análise das imagens, além das dificuldades referentes a indexação da grande quantidade de imagens geradas diariamente ao redor do mundo, tendo em vista a intensificação do uso da TC por causa redução dos custos de equipamentos de geração e armazenamento das imagens (OLIVEIRA; CIRNE; AZEVEDO-MARQUES, 2007). Já os especialistas estão sujeitos a fatores inter e intrapessoais, que podem levá-los a erro no diagnóstico da doença como, por exemplo, distrações, capacidade técnica, fadiga e limitações de memória (EADIE; TAYLOR; GIBSON, 2012).

Tendo em vista este panorama, pesquisadores como Reeves e Kostis (2000), Oliveira, Cirne e Azevedo-Marques (2007), Xu et al. (2012) e Ferreira Junior e Oliveira (2014)

têm se voltado para a busca por técnicas computacionais que minimizem o impacto das dificuldades que envolvem o diagnóstico do câncer para que seja cada vez mais precoce e preciso. Neste sentido, muitos sistemas de Diagnóstico Auxiliado por Computador (CAD) têm sido desenvolvidos.

Sistema CAD tem como função principal auxiliar o diagnóstico médico por meio do fornecimento de informações oriundas da análise de dados quantitativos extraídos dos exames, que forneçam uma segunda opinião ao especilista com o objetivo de guiá-lo na tomada de decisão acerca do diagnóstico de doenças. Vale ressaltar que o computador não deve ser entendido como um substituto do especialista, mas como uma ferramenta de auxílio e, portanto, deve prover meios para que a precisão diagnóstica aumente através do efeito sinérgico das competências do médico com a capacidade de processamento de dados do computador (DOI, 2007; AZEVEDO-MARQUES, 2001).

Levando-se em consideração as características que envolvem este contexto como a grande quantidade de imagens de exames de TC geradas todos os dias e a tendência para o desenvolvimento de sistemas que forneçam um auxílio à tomada de decisão com base nos dados extraídos a partir de imagens e exames, os sistemas de Recuperação de Imagens Baseada em Conteúdo (CBIR) surgem com grande potencial.

CBIR pode ser sucintamente definido como uma categoria de sistemas que é voltada para recuperação de imagens baseadas em informações textuais (que fazem referências às propriedades dos objetos de interesse); ou extraídas a partir da borda, forma e textura dos objetos; ou a partir de outras formas de representação de objetos; ou ainda a partir da combinação de todas as informações anteriores. A similaridade entre os objetos é medida segundo critérios específicos para cada domínio como, por exemplo, uso de funções de distância para modelos cuja representação está definida no espaço vetorial. Em sentido amplo, CBIR ajuda os usuários a encontrarem imagens com conteúdos similares (AKGÜL et al., 2011; MüLLER et al., 2004; PONCIANO-SILVA et al., 2013) para que sirvam como guia na tomada de decisão.

Aplicado ao contexto deste trabalho, um sistema CBIR pode ser utilizado para recuperar exames cujos nódulos pulmonares são semelhantes a um nódulo de referência. O conjunto de nódulos semelhantes recuperados, previamente diagnosticados em avaliações médicas anteriores, pode servir como um guia para que o especialista decida pelo diagnóstico do nódulo cuja malignidade ainda é uma incógnita, tendo em vista que ele pode tomar por base as características semelhantes que correlacionam os nódulos encontrados pelo sistema e o nódulo que está sendo avaliado.

Contudo, as formas de representação e de medição da similaridade entre os objetos em sistemas CBIR ainda é considerada uma limitação, dado que não existem formas definitivas para tal (AKGÜL et al., 2011). Isso faz com que o tema seja foco para mais pesquisas. Autores como Xu et al. (2012), Ferreira Junior (2015), Dhara et al. (2012), Seitz Jr et al. (2012) e Kuruvilla e Gunavathi (2014) apresentam diferentes formas de tratar

estas limitações dos sistemas CBIR, mas ainda é possível alcançar melhores resultados.

Dentre as formas de representação de objetos, Atributos de Textura (AT) e Atributos de Nitidez de Borda (ANB) têm sido usados em trabalhos como os apresentados por Han et al. (2014), Xu et al. (2012) e Ferreira Junior (2015). Estes atributos têm a capacidade de descrever características estatísticas a respeito da distribuição dos pixels das imagens dos nódulos. A partir deles, é possível comparar nódulos por meio de funções de distância, como a Distância Euclidiana (DE), para mensurar a sua similaridade.

Kuruvilla e Gunavathi (2014), além da análise de diferentes atributos, comparam diferentes métricas para identificar aquela que apresenta melhores resultados com relação a precisão. Eles demonstram que diferentes métricas de similaridade com o mesmo conjunto de atributos podem alcançar diferentes resultados.

Desta forma, comparar atributos, como AT e ANB, através de uma Distância Euclidiana Ponderada (DEP) onde sejam utilizados pesos ajustados conforme os dados extraídos dos nódulos, com o intuito enfatizar alguns atributos e desenfatizar outros, pode alcançar maior precisão na recuperação de nódulos pulmonares similares.

1.2 Motivação

Tendo em vista o alto índice de mortalidade e as dificuldades que envolvem o diagnóstico do câncer pulmonar, construir sistemas computacionais para auxiliar aos especialistas na detecção desta doença pode proporcionar aumento na precisão do diagnóstico, além de fornecer meios para que ele seja feito cada vez mais precoce. Isso aumenta a qualidade de vida dos pacientes e leva a maior probabilidade de êxito no tratamento da doença.

O desenvolvimento de sistemas CBIR pode proporcionar ferramentas computacionais úteis para que os especialistas consigam desempenhar com maior precisão e menor desgaste suas atividades profissionais. Contudo, faz-se necessário encontrar melhores formas de representar e comparar os nódulos pulmonares. O objetivo disso é aumentar a precisão na recuperação de exames diagnósticados para que os nódulos sejam os mais similares.

1.3 Objetivos

O objetivo principal deste trabalho é apresentar um algoritmo de ajuste local de pesos para construir modelos resultantes de aproximação de função por meio do ajuste dos pesos da Distância Euclidiana Ponderada (DEP) em sistemas de CBIR, com o intuito de contribuir para a literatura relacionada ao câncer pulmonar através da apresentação de uma solução que propicie aumento da precisão no diagnóstico e permita que o câncer seja diagnosticado cada vez mais precoce.

1.3.1 Objetivo secundário

Como objetivo secundário será avaliada a precisão do algoritmo na recuperação de nódulos pulmonares semelhantes, sendo os nódulos pulmonares representados vetorialmente por meio dos AT 3D e dos ANB 3D. A análise destes atributos permitirá identificar qual destes conjuntos de dados proporcionam uma melhor precisão na recuperação de nódulos similares.

1.3.2 Hipótese

Por fim, será testada a hipótese de que a utilização da DEP, com os pesos ajustados pelo processo que será apresentado, é mais precisa do que a utilização da DE, quando utilizada como métrica de similaridade em sistemas CBIR cujos objetos são representados por meio de vetores de atributos.

1.4 Organização da dissertação

Neste capítulo, foram apresentadas as considerações iniciais e os objetivos a serem alcançados com o algoritmo proposto. O restante do texto segue a seguinte organização:

- **Capítulo 2 - Fundamentação teórica:** traz os principais conceitos que envolvem a problemática do câncer pulmonar, seu diagnóstico e a respeito da tecnologia para fornecer uma base teórica necessária para o entendimento do algoritmo proposto;
- **Capítulo 3 - Materiais e métodos:** descreve o algoritmo proposto, os recursos necessários para sua construção e a forma pela qual será avaliado;
- **Capítulo 4 - Resultados e discussão:** apresenta os resultados obtidos através da aplicação do algoritmo proposto no contexto do câncer pulmonar e discute os resultados alcançados fazendo comparações com outros trabalhos encontrados na literatura corrente;
- **Capítulo 5 - Conclusão:** apresenta as conclusões do trabalho, as limitações e os trabalhos futuros envolvendo o algoritmo proposto.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, será feita uma revisão da literatura com o objetivo de contextualizar a problemática do câncer pulmonar no mundo e fornecer uma base teórica suficiente para a compreensão do trabalho proposto. Ele está organizado da seguinte forma: a Seção 2.1 apresenta a problemática envolvendo o câncer de pulmão no Brasil e no mundo; a Seção 2.2 traz conceitos fundamentais necessários para o entendimento dos algoritmos envolvendo o processamento de imagens digitais; a Seção 2.3 apresenta uma visão geral a respeito do conceito de CAD; a Seção 2.4 apresenta uma breve explanação acerca do conceito de CBIR; a Seção 2.5 define o que são algoritmos de extração de atributos e especifica quais atributos serão utilizados neste trabalho para representar os nódulos pulmonares; a Seção 2.7 apresenta detalhes da construção da base de imagens utilizada neste trabalho; a Seção 2.8 define o que é uma métrica de similaridade e como é feita a comparação entre os nódulos neste trabalho; a Seção 2.9 apresenta os conceitos relacionados construção de modelos de regressão locais e globais; a Seção 2.10 traz conceitos relacionados a qualidade da recuperação dos sistemas de recuperação de imagens e define qual método será utilizado para avaliação e comparação com outros trabalhos da literatura; por fim, a Seção 2.11 apresenta uma breve revisão da literatura referenciando alguns trabalhos relacionados ao tema aqui tratado.

2.1 Câncer de pulmão

O câncer é um crescimento celular anormal e incontrolado que invade os tecidos vizinhos. Ele se inicia com uma única célula maligna e cresce a partir da duplicação de células no processo de divisão celular a uma razão constante (UEHARA; JAMNIK; SANTORO, 1998).

Ele é terceiro tipo de câncer mais diagnosticado no mundo, ficando atrás apenas do câncer de próstata nos homens e de mama nas mulheres, entretanto, o câncer pulmonar é mais letal do que os dois primeiros da lista (WENDER et al., 2013).

Os tipos mais comuns da neoplasia pulmonar maligna são o carcinoma espinocelular e o adenocarcinoma, representando 40 e 30%, respectivamente, o carcinoma indiferenciado de pequenas células varia de 15 a 20%, e o carcinoma indiferenciado de grandes células, aproximadamente 10% das patologias identificadas (UEHARA; JAMNIK; SANTORO, 1998).

Zamboni (2002) classifica os fatores que ocasionam o surgimento desta doença em externos e internos ao homem. O autor aponta como fatores externos: tabagismo ativo (fumante propriamente dito) e passivo (fumante indireto, aquele que apenas inala a famuça do cigarro); poluição atmosférica; problemas relacionados com a saúde ocupacional como,

por exemplo, exposição ao radônio, asbestos e outras fibras minerais, sílica, cromo, níquel, arsênico e hidrocabornetos aromáticos policíclicos. Já como fatores interno são citados por ele: doenças que estão associadas ao risco de câncer pulmonar como tumores da cabeça e pescoço; síndrome da imunodeficiência adquirida e outras doenças pulmonares não malignas; além desses, o fator genético também pode ser responsável pelo surgimento da doença, embora Zamboni (2002) evidencie que não esteja comprovada esta ligação, o autor afirma que já existem estudos que apontam para isso.

Além das consequências fisiológicas da doença, os pacientes diagnósticados com câncer pulmonar apresentam altos níveis de sofrimento psíquico e físico comparados a outros tipos de câncer. Eles, geralmente, são acometidos do estigma[1] do câncer e da auto-culpa, afetando diretamente o tempo pela procura médica. O estigma também é uma barreira na comunicação entre o paciente e o profissional da saúde (LIMA; PIMENTA, 2015).

2.2 A imagem de Tomografia Computadorizada

A TC é um dos inventos mais importantes da aplicação de processamento de imagens no diagnóstico médico (GONZALEZ; WOODS, 2008). Ela foi desenvolvida em meados da década de 1970 e ainda hoje é inestimável quando utilizada como ferramenta para o diagnóstico em muitas aplicações clínicas (desde o diagnóstico do câncer, a vizualizações de traumas e osteoporoses) (BUSHBERG; BOONE, 2011). Diciotti et al. (2010) reafirmam a importância da TC apontando-a como a principal técnica de imagem para detecção de nódulos pulmonares atualmente. O seu advento abriu uma nova perspectiva para o diagnóstico de nódulos pulmonares, principalmente de nódulos em estágios iniciais (quando têm menos de 3mm de diâmetro). Os avanços tecnológicos destas máquinas têm permitido melhorar a sensibilidade (medida que reflete a eficácia na identificação de indivíduos com uma determinada característica, neste caso, indivíduos com câncer pulmonar) e especificidade (medida que reflete a eficácia na identificação de indivíduos sem uma determinada característica, neste caso, indivíduos sem câncer pulmonar), provendo meios para a detecção de nódulos pequenos (estágio inicial) sem a necessidade do uso de contraste intravenoso (PASTORINO, 2010; TEAM, 2011).

Wender et al. (2013) afirmam que a taxa de sobrevida média de um paciente diagnosticado com câncer pulmonar é de cinco anos. Isso evidencia a necessidade do diagnóstico precoce da doença. Para tanto, Uehara, Jamnik e Santoro (1998) apresentam algumas técnicas diagnósticas utilizadas nos dias de hoje para auxiliar ao especialista, dentre elas: Radiografia (RD), TC, ressonância magnética, citologia de escarro, broncofibroscopia, biópsia por agulha transcutânea, mediastinoscópio, biópsia de céu aberto e toracocentese.

Das técnicas citadas acima, a TC e a RD são as técnicas mais comuns hoje em dia. Entretanto, a TC é a que tem possibilitado melhores resultados no auxílio ao diagnóstico

[1] O estigma é definido como um atributo negativo, que rotula uma pessoa como diferente ou com uma deficiência percebida negativamente pela sociedade.

do câncer de pulmão. Em uma comparação feita por Wender et al. (2013), ficou demonstrado que a TC proporcionou uma redução de 20% das mortes por câncer pulmonar nos casos de acompanhamento de pacientes com os sintomas do câncer em comparação com o acompanhamento feito com RD. Isso é possível devido às possibilidades trazidas pela tecnologia da TC como, por exemplo, possibilidade de reconstrução em 3D a partir das imagens dos exames através de algoritmos computacionais; possibilidade de visualização das fatias dos exames individualmente; capacidade de visualização dos exames sem que haja superposição dos órgãos nas imagens; permite identificar pequenas variações nos tecidos devido ao melhor contraste das imagens; e, por fim, possibilita a manipulação e otimização das imagens por meio de ferramentas computacionais. Por isso é possível obter informações muito mais precisas sobre os nódulos e aumentar o índice de detecção precoce de nódulos.

Gonzalez e Woods (2008) descrevem a TC como um processo no qual um anel de detectores envolvem um objeto (ou paciente), transpassando uma fonte de raios-X concêntricos com o anel detector, girando ao redor do objeto (Figura 1 a). Os raios-X passam através do objeto e são coletados no lado oposto ao da emissão pelos detectores correspondentes no anel. Conforme o emissor gira, este processo é repetido. A partir disso, algoritmos são aplicados sobre os dados capturados para construir uma imagem que representa uma fatia do objeto (Figura 1 b). O movimento do objeto na direção perpendicular ao anel de detectores produz um conjunto das tais fatias.

Assim, de forma sucinta, podemos afirmar que a TC é um exame cujo resultado é um *array* de imagens ordenadas temporalmente, capturadas a uma distância aproximadamente uniforme (1 a 10mm), formando um volume de imagens.

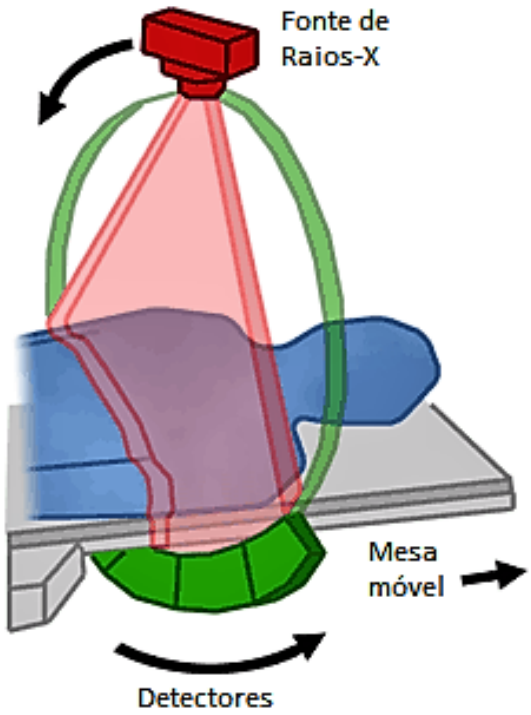
Uma imagem pode ser definida formalmente como uma função bi-dimensional $f(x, y)$, onde x e y representam as coordenadas espaciais (plano), e f é a amplitude em um dado ponto de coordenadas (x, y), que é chamada de intensidade que pode ser representada por cores ou níveis de cinza, da imagem naquele ponto. Quando x, y e f são todos finitos, com quantidades discretas, chamamos imagem digital (GONZALEZ; WOODS, 2008). Uma forma matricial de representação da imagem pode ser vista na Tabela 1. Neste trabalho, o foco será voltado para imagens com representação de intensidades através dos níveis ou escalas de cinza. Este tipo de imagem é conhecida também como imagem monocromática. Sendo assim, a partir de agora, sempre que for feita mensão ao termo imagem ele deverá ser associado a este tipo de imagem.

A unidade elementar de uma imagem bi-dimensional é chamada de *pixels (picture elements)*. Ele está representado como cada um dos índices na Tabela 1 e contém a intensidade de nível de cinza (l) correspondente àquela posição da imagem (Equação 2.1).

$$l = f(x, y) \tag{2.1}$$

Figura 1 – Ilustração do exame com TC.

(a) Representação de um aparelho de TC. Fonte: Imagem adaptada a partir de.



(b) Imagem gerada a partir da captura de uma fatia.

Fonte: FDA (2015).

Tabela 1 – Representação de uma imagem através de uma matriz $X \times Y$, onde X corresponde ao número de colunas e Y ao número de linhas

$$\begin{vmatrix} f(0,0) & f(1,0) & \dots & f(X-1,0) \\ f(0,1) & f(1,1) & \dots & f(X-1,1) \\ . & & & \\ . & . & & \\ . & & . & \\ f(0,Y-1) & f(1,Y-1) & \dots & f(X-1,Y-1) \end{vmatrix}$$

Fonte: Elaborada pelo autor.

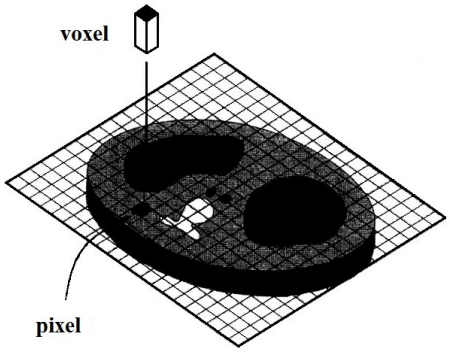
Onde l está em um intervalo específico (Equação 2.2)

$$L_{min} \leq l \leq L_{max} \tag{2.2}$$

O número de *pixels* em uma imagem corresponde ao número de *voxels (volume elements)*, mas esses estão relacionados a volumes de imagens em 3D. Os *voxels* têm a mesma dimensão dos *pixels* no plano, mas eles também incluem a espessura da fatia (dimensão) (Figura 2).

Segundo Gonzalez e Woods (2008), L_{min} e L_{max} são valores positivos e finitos, e o

Figura 2 – Um *pixel* é o elemento básico de uma imagem bi-dimensional. Cada *pixel* corresponde a um *voxel* no paciente (tri-dimensional). O *voxel* tem as duas dimensões de uma imagem e ainda a terceira dimensão, que é representada pela espessura da fatia de um exame de TC.



Fonte: Bushberg e Boone (2011).

intervalo $[L_{min}, L_{max}]$ é a escala de cinza. Na prática, este intervalo fica definido em $[0, L-1]$, sendo $l = 0$ considerado preto e $l = L-1$ considerado branco na escala de cinza. Todos os valores intermediários são tons de cinza variando do preto ao branco.

Ainda segundo os autores, o processo de digitalização requer a definição dos valores de X, Y e L. Não existem restrições para os valores de X e Y desde que $\{X, Y \in \mathbb{N}_+\}$. Entretanto, devido ao processo de armazenamento e restrições de *hardware*, o valor de L tipicamente é uma potência de dois, onde k é o número de *bits* necessários para armazenar os valores de cinza determinados (Equação 2.3).

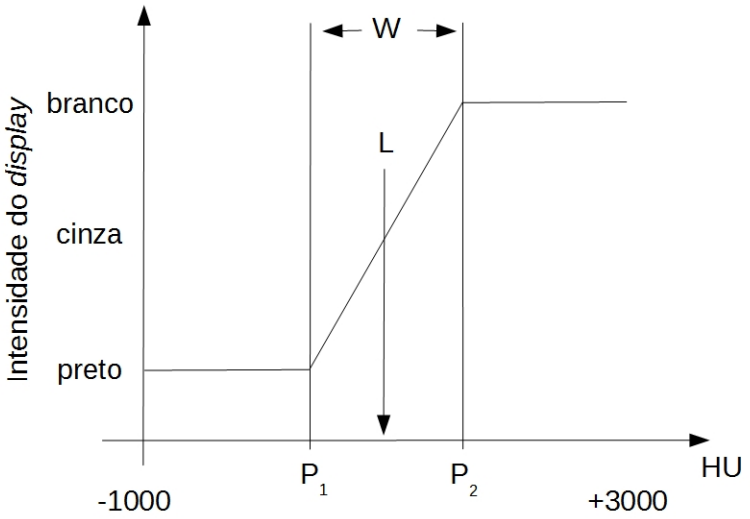
$$L = 2^k \tag{2.3}$$

Imagens de TC tipicamente possuem 12 *bits* para representar os níveis da escala de cinza, perfazendo um total de 4.096 (2^{12}) níveis. Entretanto, o olho humano tem a capacidade limitada, conseguindo ver de 30 a 90 tons de cinza, então 6 ou 8 *bits* são suficientes para apresentar a imagem. Assim, as imagens de 12 *bits* resultantes da captura pelos aparelhos de tomografia são reduzidas para 8 *bits* para se acomodar aos *hardwares* de visualização por meio da técnica Janelamento e Nível (BUSHBERG; BOONE, 2011).

Segundo Bushberg e Boone (2011) e Allisy-Roberts e Williams (2008), a largura da janela (W) determina o contraste da imagem. Enquanto o nível (L) é o número associado a intensidade no centro da janela. A seleção dos valores de W e L determinam dois pontos de inflecção P_1 e P_2, onde $P_1 = L - {}^{W}/_{2}$ e $P_2 = L + {}^{W}/_{2}$. Todos os valores abaixo de P_1 devem ser saturados para o preto e todos acima de P_2 devem ser saturados para o branco. Desta forma, as informações contidas nas áreas saturadas são perdidas (Figura 3).

A escolha dos valores de L e W deve ter como referência o valor do coeficiente de atenuação (μ) dos tecidos que são de interesse nos exames fazendo com que apenas eles sejam apresentados como níveis de cinza, como afirma Allisy-Roberts e Williams (2008).

Figura 3 – Conceitos de como a janela e o nível são usados para manipular o contraste das imagens de TC. O nível (L) corresponde ao centro da janela. A janela (W) determina o contraste da imagem. P_1 e P_2 são os pontos de inflecção.



Fonte: Imagem adaptada de Bushberg e Boone (2011, p. 359).

Os referidos autores apontam como menor valor de referência o ar (-1000 HU) e como maior valor de referência o osso (+1000 HU), por exemplo. O número de *Hounsfield* (HU) é a unidade de medida utilizada para mensurar o coeficiente de atenuação dos tecidos. Ele é calculado pela Equação 2.4, onde μ_w é o coeficiente de atenuação da água e μ_t é o coeficiente de atenuação do tecido de interesse.

$$HU = 1000 \times \frac{(\mu_t - \mu_w)}{\mu_w} \tag{2.4}$$

2.3 Diagnóstico Auxiliado por Computador

CAD pode ser definido como uma categoria de sistema que permite que o diagnóstico médico seja feito com o auxílio de informações de análises quantitativas determinadas por computador a partir de informações coletadas de exames. Ele serve como uma segunda opinião que objetiva guiar o profissional da saúde na tomada de decisão acerca do diagnóstico de doenças. Vale ressaltar que o computador deve ser entendido como uma ferramenta de auxílio e, portanto, não deve ter seu desempenho comparado ao do ser humano, mas deve promover um aumento da precisão através do efeito sinérgico das competências do médico com a capacidade de processamento de informações do computador (DOI, 2007; AZEVEDO-MARQUES, 2001).

Existem dois tipos de sistemas CAD apontados pela literatura: um tipo voltado para o auxílio à detecção de lesões (CADe) e outro voltado para o auxílio ao diagnóstico (CADx). O primeiro (CADe) tem por objetivo localizar padrões anormais na imagem. Já o segundo (CADx) tem por objetivo determinar a classe das lesões detectadas automaticamente ou manualmente, por meio de informações estatísticas extraídas das lesões (EADIE; TAYLOR; GIBSON, 2012). Este trabalho segue a linha dos sistemas CADx, logo, a partir de agora, este tipo será referenciado indiscriminadamente como CAD para facilitar o entendimento do texto.

A importância do uso de sistemas CAD na rotina médica está relacionada a capacidade de auxiliar na detecção e quantificação das lesões pelos radiologistas em meio a grande quantidade de informações geradas diariamente. A redução dos custos para obtenção de equipamentos de geração de imagem e armazenamento possibilita a massificação do uso de tecnologias avançadas como, por exemplo, a TC nos centros de radiologia, aumentando, assim, a quantidade de imagens de exames geradas todos os dias em serviços de radiologia ao redor do mundo (com taxas de produção que chegam a *Terabytes* por ano). Este número crescente de exames faz com que os radiologistas tenham cada vez mais exames para analisar diariamente, levando à problemas de ordem inter e intrapessoais como distrações, fadiga e limitações de memória. Além disso, a grande quantidade de imagens geradas implica dificuldades para indexação e acesso, o que diminue a capacidade de análise e tratamento dos dados e consequente geração de informação (OLIVEIRA; CIRNE; AZEVEDO-MARQUES, 2007). Este panorama exige o emprego de técnicas que permitam efetivamente manipular os dados e colher benefícios deles.

Outro fator que evidência a relevância dos sistemas CAD é que estes sistemas possibilitam que haja uma padronização no cálculo e interpretação dos dados evitando que os diagnósticos sejam afetados por tendências e pré-concepções. Além disso, sistemas CAD podem identificar mudanças sutis, mas importantes, que os especialistas podem não identificar. Tudo isso leva a crer que CAD tem o potencial de reduzir a taxa de erro nos diagnósticos de exames (EADIE; TAYLOR; GIBSON, 2012).

Segundo Akgül et al. (2011), a interpretação de imagens médicas possui três atividades chaves: (1) percepção de achados nas imagens (anormalidade detectada através da imagem); (2) interpretação dos achados para definir o dignóstico; e (3) recomendações para o gerenciamento clínico (biópsia, acompanhamento e etc.) ou mais exames de imagem para firmar o diagnóstico ainda não estabelecido. O potencial da interpretação assistida e da tomada de decisão é motivado não apenas por restrições de tempo para leitura dos exames, mas também pelo reconhecimento de variações entre os especialistas que estão relacionadas a fatores como erro, falta de treinamento e fadiga.

2.4 Recuperação de Imagens Baseada em Conteúdo

A dificuldade no diagnóstico de lesões pulmonares com base em análises visuais dos exames de TC, juntamente com o grande volume gerado diariamente pelos hospitais, tem levado vários grupos de pesquisa a buscarem soluções para estes problemas (AZEVEDO-MARQUES, 2001). Uma proposta de solução de sistema que vêm emergindo com grande força nas pesquisas aplicadas à medicina devido ao seu potencial é a CBIR (MüLLER et al., 2004). Em sentido amplo, CBIR ajuda os usuários a encontrarem imagens com conteúdos similares (AKGÜL et al., 2011).

Estes sistemas possuem a capacidade de buscar e recuperar imagens (ou lesões) de exames, já diagnosticados, similares a um caso novo, ainda não diagnosticado, segundo a similaridade entre as suas características. Com base nos exames similares recuperados, o especialista pode se sentir mais confiante em decidir o diagnóstico do novo caso. Isso faz com os sistemas CBIR tenham grande potencial para se tornar sistemas CAD (DOI, 2007).

Oliveira, Cirne e Azevedo-Marques (2007) afirmam que o objetivo do CBIR como sistema de informação em radiologia é fornecer a informação correta para o especialista no tempo apropriado, com o intuito de melhorar a qualidade e a eficiência dos diagnósticos. Os benefícios provenientes do uso da CBIR permitem recuperar imagens semelhantes com base na região anatômica e na patologia.

Além do diagnóstico, CBIR possibilita melhorias no ensino e pesquisa em áreas da medicina. Müller et al. (2004) mostram que estas aplicações também pode fornecem meios para a análise de aspectos visuais de casos específicos com características visuais similares e diagnósticos diferentes, e também permitem uniformizar o processo de aprendizagem durante as aulas por meio da análise visual de casos similares recuperados, que é o uso mais comum.

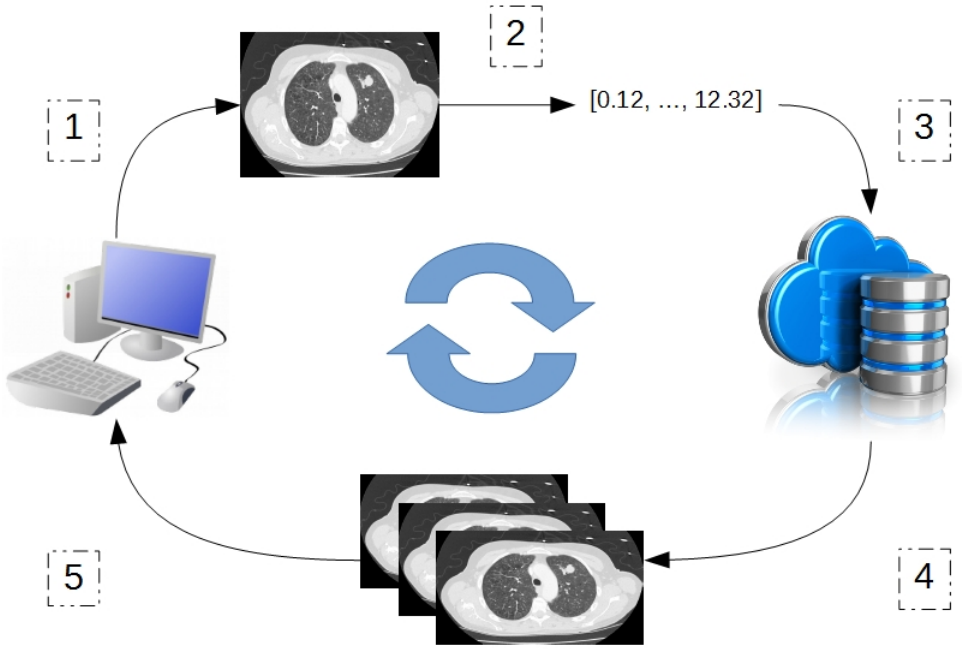
Segundo Akgül et al. (2011), um sistema CBIR genérico tem pelo menos dois componentes. O primeiro componente representa as informações contidas nos *pixels* das imagens através de atributos/descritores e tem por objetivo ser uma ponte entre o conteúdo visual e sua representação numérica em sistemas computacionais. O segundo componente provê um modelo para avaliação da similaridade entre os atributos da imagem baseados em análises matemáticas.

Por outro lado, Müller et al. (2004) apontam que a maioria dos sistemas CBIR têm uma arquitetura semelhante para busca e indexação de imagens (Figura 4). Segundo esses autores, geralmente, os sistemas são compostos por módulos de: armazenamento e acesso à bases de dados; motores de recuperação; módulos de extração de atributos; módulos de cálculos de distância; e interfaces para interação homem-máquina.

Vale salientar que mesmo havendo divergência na quantidade de módulos básicos entre os autores citados percebe-se que a essência do CBIR é a mesma para os autores.

Em sistemas CBIR, a recuperação de imagens é executada segundo a similaridade

Figura 4 – Modelo simplificado de um sistema CBIR. Em (1), é fornecida uma *interface* para que o usuário forneça a imagem que ele deseja passar como critério de busca. Em (2), a região de interesse, que é a região da imagem daqual se deseja extrair as informações, é convertida em atributos que a descrevem através de algoritmos extratores de atributos. Neste exemplo, a região de interesse está representada por um vetor n-dimensional. Em (3), o vetor extraído é utilizado para, através de uma determinada métrica de similaridade, buscar as imagens mais similares. Em (4), as imagens mais semelhantes são retornadas em ordem de similaridade (as mais similares vêm primeiro) segundo o critério estabalecido. Em (5), as mais semelhantes são exibidas e servem como base para a tomada de decisão pelo especialista.



Fonte: Elaborada pelo autor.

entre elas. Estes sistemas recuperam imagens similares a uma imagem de referência dado um critério de busca. A similaridade entre imagens pode ser calculada, por exemplo, comparando-se vetores de atributos extraídos a partir das imagens. Os vetores são construídos a partir de dados estatísticos extraídos automaticamente, semi-automaticamente ou manualmente do conteúdo das imagens através de algoritmos extratores de atributos (Seção 2.5) relacionados a domínios como textura, forma e cor. Com este tipo de representação dos objetos, a comparação, geralmente, é feita através de funções de distância (Seção 2.8) para medir o quão perto dois vetores estão no espaço multidimensional (PONCIANO-SILVA et al., 2013).

2.5 Algoritmos extratores de atributos

São várias as áreas de processamento de imagens digitais e não há um limite claro que as separe. Entretanto, Gonzalez e Woods (2008) apresentam um paridgma que considera três tipos de processos computadorizados, que são os processos de baixo, intermediário e alto nível. Os processos de baixo nível envolvem operações primitivas como redução de ruídos, melhoramente de contraste e outros. As entradas e saídas dos processos de baixo nível são imagens digitais. Os processos de nível intermediário envolvem atividades de segmentação (particionamento de uma imagem em regiões ou objetos), descrição dos objetos das imagens de forma reduzida que possibilite o processamento computacional. Os processos de nível intermediário são caracterizados por terem como entrada uma imagem e como saída atributos extraídos das imagens (bordas, contornos e textura que identificam os objetos). Por fim, processos de alto nível envolvem a ação de dar sentido aos objetos de uma imagem através da execução de funções cognitivas e análise de imagens. Os algoritmos de extração de atributos residem exatamente nos processos de nível intermediário.

Existem pelo menos duas abordagens para buscar por imagens em uma base: buscar por descritores textuais relacionados às imagens; ou buscar por meio de características inerentes às imagens obtidas através de algoritmos específicos para tal fim. A primeira abordagem não apresenta bons resultados devido a alguns fatores como, por exemplo, a falta de padronização nas palavras utilizadas para descrever os objetos de interesse, já que diferentes profissionais podem descrevê-los usando diferentes palavras ou usando as mesmas palavras, mas com diferentes conceitos; as anotações manuais são resultantes de descrições subjetivas com alto grau de dependência das capacidades, treinamento e experiência do especialista responsável por elas; outro fato que deve ser evidenciado é que a grande quantidade de exames gerados diariamente torna as anotações manuais inviáveis na rotina médica. A segunda abordagem é baseada na extração de características através de algoritmos computacionais, que é independente da subjetividade humana e possibilita a automação do processo de cálculo das características. Nesta última abordagem, a imagem passa a ser representada pelo conjunto de características que representam a essência da imagem, permitindo a comparação automática (KUMAR et al., 2013; KULKARNI; KULKARNI; STRANIERI, 2014).

A representação das características visuais de uma imagem pode ser feita através da quantificação de valores estatísticos calculados a partir das intensidades dos *pixels* e podem ser vistas como pontos em um espaço multidimensional. Os principais algoritmos utilizados com este propósito medem valores associados à cor, à textura e à forma da região de interesse (OLIVEIRA; AZEVEDO-MARQUES; FILHO, 2007). Textura e cor podem ser aplicadas tanto para a imagem de forma global (imagem inteira), como em regiões específicas da imagem. Já a forma é aplicada apenas para o último caso. A principal diferença entre analisar de forma global ou específica é que características extraídas

localmente contêm mais informações sobre o objeto ou estrutura de interesse (MüLLER et al., 2004).

Atributos, ou descritores, de imagem são derivados da interpretação visual de dados contidos na imagem. Eles são representados como dados alfa-numéricos em diferentes formatos como vetores ou grafos, os quais se apresentam como substitutos para o conteúdo visual. Pode-se distinguir pelo menos dois tipos de características visuais (AKGÜL et al., 2011):

- **Características fotométricas** - exploram a cor e a textura, sendo derivadas diretamente das informações contidas nos *pixels* das imagens;
- **Características geométricas** - fazem uso das informações visuais de forma.

Um algoritmo extrator de atributos pode ser definido como uma função do tipo

$$f : \mathbb{I} \rightarrow \mathbb{R}^n \tag{2.5}$$

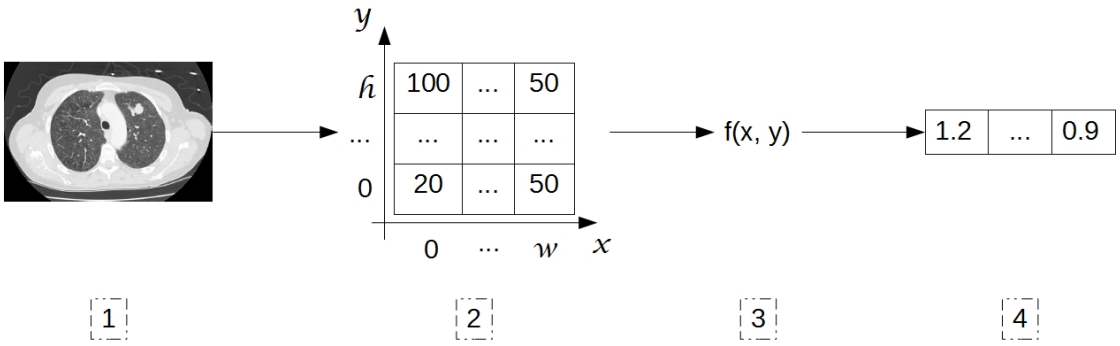
onde:

- $\mathbb{I}$ é a imagem representada por uma matriz $M_{X \times Y}$, na qual X corresponde ao número de colunas e Y corresponde ao número de linhas, sendo $X, Y \in \mathbb{N}^*$, e cada posição $(x, y) = l$ em que $l \in \mathbb{N}$ corresponde a uma intensidade de *pixel*;
- $\mathbb{R}^n$ é um vetor do espaço n-dimensional, e n corresponde a dimensão do vetor.

Desta forma, algoritmos de extração de atributos quantificam características estatísticas calculadas a partir da distribuição dos *pixels* de uma imagem e as utilizam para representar a imagem no espaço vetorial, ou através de histogramas, ou outras formas de representação existentes (Figura 5). Traina et al. (2003) define um vetor de atributos como sendo uma representação concisa de uma imagem que, segundo um critério específico, fornece a essência dela.

Embora existam diversas formas de representação, ainda não se tem uma técnica ouro, aquela que possibilita alcançar 100% de precisão no auxílio ao diagnóstico. Isso se deve ao fato de que os especialistas utilizam informações visuais, textuais, conhecimentos teóricos, consultas a outros especialistas e a sua experiência para a tomada de decisão, enquanto a máquina faz uso de representações matemáticas para modelar os objetos de interesse e, a partir delas, tomar decisões acerca dos nódulos apresentados. Como pode ser percebido, existem diferenças nos parâmetros utilizados pelo homem e pela máquina. Esta disparidade é conhecida como *gap* semântico na literatura de sistemas CAD e CBIR. Este é um problema que ainda está aberto, instigando os especialistas na busca por modelos computacionais mais precisos (BEDO et al., 2015).

Figura 5 – Demonstração do processo de extração de atributos com representação através de vetores. Em (1) é apresentada uma imagem oriunda de um exame de TC de um pulmão. (2) representa a extração da matriz de intensidade de *pixels* que representa a imagem em formato digital, onde w é a largura da imagem e h é a sua altura. (3) representação da aplicação de funções de extração de atributos. E (4) apresenta o resultado da extração de atributos em formato vetorial.



Fonte: Elaborada pelo autor.

2.5.1 Atributos de Textura 3D

Buscar formas de descrever as informações visuais de uma imagem através de representações matemáticas é um caminho natural quando se busca definir os tipos de características que o ser humano usa para interpretar a imagem (HARALICK; SHANMUGAM; DINSTEIN, 1973).

A textura é uma propriedade inata de qualquer superfície, pois ela contém informações sobre o arranjo estrutural da superfície e do ambiente ao seu redor. AT codificam a organização espacial dos valores dos *pixels* de uma imagem (AKGÜL et al., 2011). Embora seja fácil de reconhecer e descrever em termos empíricos por um observador humano, textura tem sido um conceito difícil para se definir precisamente em ambiente computacional. Mas como as texturas das imagens carregam informações importantes para propósitos de discriminação, desenvolver características texturiais se tornou um tema muito importante (HARALICK; SHANMUGAM; DINSTEIN, 1973).

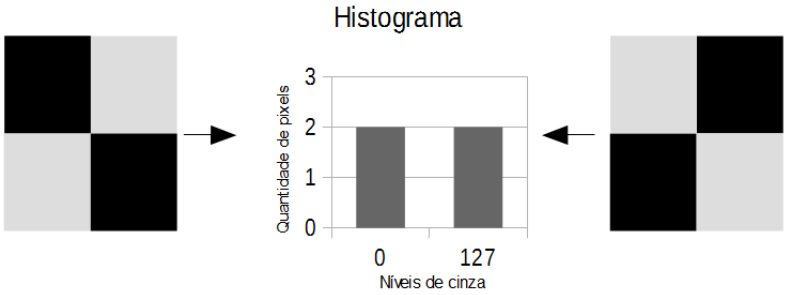
No domínio médico, descritores baseados em textura se tornam particularmente importantes porque eles podem refletir os detalhes finos contidos dentro da estrutura das imagens. Por exemplo, cistos e nódulos sólidos, geralmente, têm densidade interna uniforme, enquanto que lesões mais complexas têm características heterogêneas (AKGÜL et al., 2011).

As características visuais de uma imagem podem ser representadas de três formas: estrutural, espectral e estatística. A primeira representa as imagens através da identificação de estruturas primitivas e pela forma como estão dispostas. A segunda representa as imagens por meio do domínio da frequência como, por exemplo, propriedades do espectro de

Fourier. Por fim, a terceira representa as imagens através de descritores de características calculados a partir dos valores e disposição dos *pixels* das imagens. Estes descritores são comumente classificados em primeira e segunda ordens (HARALICK; SHANMUGAM; DINSTEIN, 1973).

Os descritores de primeira ordem são extraídos a partir de informações dos histogramas das imagens. Esses têm limitações devido a sua incapacidade de capturar informações sobre a distribuição espacial dos *pixels*, além de apresentar resultados ambiguos, porque diferentes imagens podem gerar o mesmo histograma (OLIVEIRA; AZEVEDO-MARQUES; FILHO, 2007) (Figura 6). Já os descritores de segunda ordem permitem obter informações sobre a distribuição dos *pixels* e são aplicados sobre a Matriz de Coocorrência (MCO) para produzir os descritores de textura (HARALICK; SHANMUGAM; DINSTEIN, 1973).

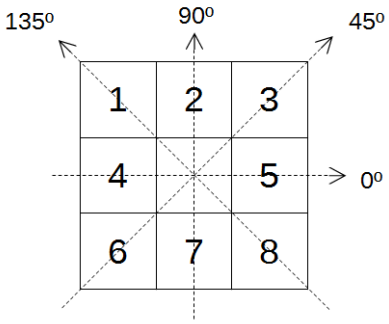
Figura 6 – Duas imagens diferentes com o mesmo histograma.



Fonte: Elaboraada pelo autor.

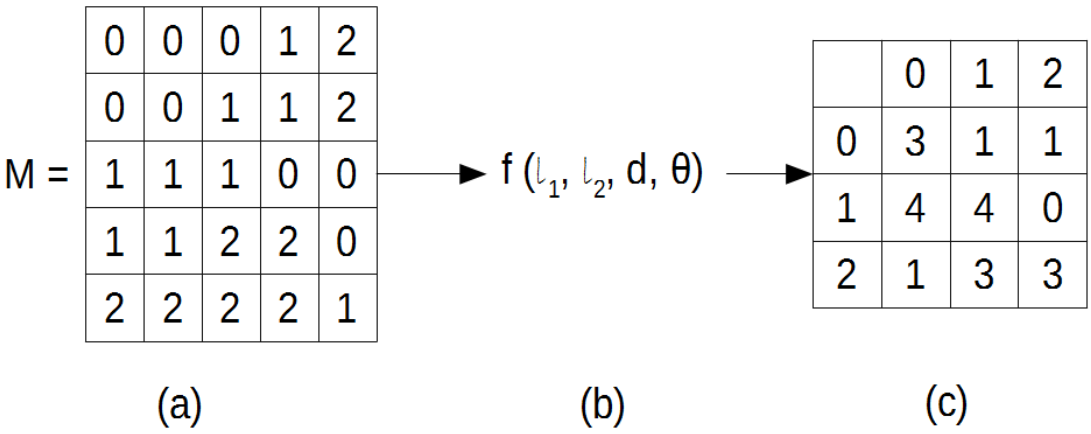
Os descritores de textura (descritores de segunda ordem) são especificados por matrizes de dependência espacial de tons de cinza, mais conhecidas como MCO, que são computadas para vários ângulos e distâncias entre pares de células vizinhas em uma imagem (Figura 7). MCO é um método estático para a quantização das características inerentes à textura. Ela é uma tabulação de quantas combinações diferentes de valores de intensidade dos *pixels* ocorrem em uma imagem. A principal função da MCO é caracterizar texturas em uma imagem através de um conjunto de estatísticas para as ocorrências de cada nível de cinza em pixels diferentes ao longo de diferentes direções. Mais especificamente, é uma matriz de frequências relativas P_{l_1,l_2} na qual duas células separadas por uma distância d ocorrem em uma imagem, uma com intensidade l_1 e outra com intensidade l_2. Ela pode ainda ser definida formalmente como uma função $F(l_1, l_2, d, \theta)$, onde l_1 e l_2 são intensidades de tons de cinza vizinhos com coordenadas $(x, y) \in \mathbb{N}^2$ segundo uma distância d e um ângulo θ (Figura 8).

Figura 7 – Distribuição dos ângulos ao redor de um *pixel* de referência com distância $d = 1$.



Fonte: Elaborada pelo autor.

Figura 8 – Exemplo da construção de uma MCO. Em (a) é apresentada uma imagem de exemplo composta por três níveis de cinza. (b) representa a aplicação da função que define a MCO tendo como parâmetros $d = 1$ e $\theta = 90^{\mathbf{o}}$. Em (c) é mostrada a MCO resultante. Observe que a MCO resultante é uma matriz $L \times L$, onde L corresponde à quantidade de níveis de cinza da imagem.



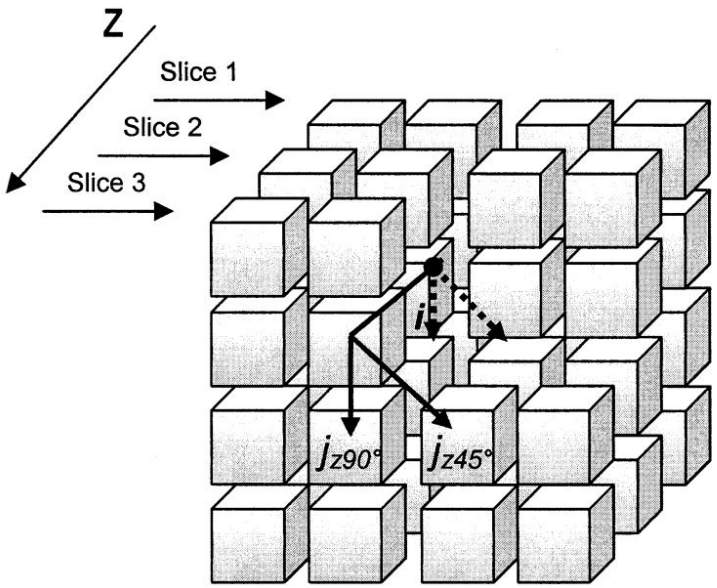
Fonte: Elaborada pelo autor.

A MCO apresentada por Haralick, Shanmugam e Dinstein (1973) é bidimensional por ser calculada a partir de uma imagem digital 2D. Mahmoud-Ghoneim et al. (2003) propuseram uma nova abordagem que estende as capacidades da MCO 2D para que possa trabalhar com volumes de imagem, tornando a MCO 3D. Segundo esses autores, esta nova abordagem proporciona melhores resultados comparados com a versão 2D por permitir que sejam feitos cálculos que levam em consideração os eixos X, Y e Z com o intuito de avaliar,

além das relações já existentes entre os *pixels* de X e Y, a relação entre esses e o eixo Z formado pelo conjunto de imagens capturadas. A avaliação dos atributos de textura sobre o volume 3D melhora a representação das características dominantes da textura permitindo, maior capacidade de discriminação de objetos (MAHMOUD-GHONEIM et al., 2003).

A MCO 3D é definida formalmente pela função $F(l_1, l_2, d, \theta, d_z)$, onde l_1 e l_2 são intensidades de tons de cinza de *voxels* vizinhos com coordenadas $(x, y) \in \mathbb{N}^2$ segundo uma distância d, um ângulo θ e d_z é a distância do *voxel* na direção Z (Figura 9). Esta nova matriz tem as mesmas características da MCO 2D. Entretanto, ela contém informações sobre a junção dos *voxels* dentro do volume de imagens.

Figura 9 – Construção da MCO 3D a partir de um volume de imagens contendo 3 fatias. A junção entre as fatias tem 1 *pixel* de distância em X, Y e Z.



Fonte: Mahmoud-Ghoneim et al. (2003).

Haralick, Shanmugam e Dinstein (1973) definiram quatorze atributos que podem ser extraídos a partir da MCO com o propósito de discriminação da textura. Oliveira, Cirne e Azevedo-Marques (2007) afirmam que existem aproximadamente vinte funções estatísticas propostas pela literatura para obtenção de informação a partir da MCO, mas nove delas são tidas como as mais importantes por produzirem satisfatória classificação de textura: entropia, inercia, energia, matiz, momento da diferença inverso, proeminência, correlação, variância e homogeneidade (Tabela 2).

Tabela 2 – AT utilizados, suas fórmulas e principais características. Oliveira (2006) e Ferreira Junior (2015) apresentam as principais características de cada um deles, assim como trazem as funções que definem os AT, onde μ_x, μ_y, σ_x e σ_y são a média e o desvio padrão de p_x e p_y, respectivamente, e $P(i,j)$ é a intensidade do pixel em escala de cinza.

Atributo de Textura	Função	Característica
Energia	$\sum_i \sum_j P^2(i,j)$	Mede a uniformidade da tonalidade dos pixels em toda região
Entropia	$-\sum_i \sum_j P(i,j) \log P(i,j)$	Mede o grau de dispersão de ocorrências de níveis de cinza em uma imagem
Contraste	$\sum_i \sum_j (i-j)^2 P(i,j)$	Mede as variações de intensidade dos pixels e resulta em valores maiores quando existem grandes diferenças entre os níveis de cinza
Momento da diferença inverso	$\sum_i \sum_j \frac{1}{1+(i-j)^2} P(i,j)$	Considera a concentração das ocorrências de níveis de cinza na diagonal da matriz de coocorrência
Matiz	$\sum_i \sum_j (i+j-\mu_x-\mu_y)^3 P(i,j)$	Resulta em altos valores quando a imagem não é simétrica
Proeminência	$\sum_i \sum_j (i+j-\mu_x-\mu_y)^4 P(i,j)$	Assim como Matiz, também resulta em altos valores quando a imagem não é simétrica
Correlação	$\sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)}{\sqrt{\sigma_x \sigma_y}} P(i,j)$	Calcula a dependência linear da tonalidade de níveis de cinza
Variância	$\sum_i \sum_j (i-\mu)^2 P(i,j)$	Mede a dispersão dos níveis de cinza em relação à intensidade média da imagem
Homogeneidade	$\sum_i \sum_j \frac{P(i,j)}{(1+\lvert i-j \rvert)}$	Homogeneidade é o inverso do contraste e resulta em grandes valores para níveis de cinza similares

Fonte: Elaborada pelo autor.

2.5.2 Atributos de Nitidez de Borda 3D

Outra forma de representar os nódulos pulmonares é através de atributos de forma. As duas principais maneiras de representar a forma são através de dois grupos de descritores que são baseados no contorno do objeto ou na informação contida na sua região. Os descritores baseados em contorno obtêm as características analisando os contornos dos objetos, enquanto os descritores baseados em região analisam a região interna delimitada pelo contorno (OLIVEIRA, 2006). Akgül et al. (2011) apresentam o termo forma referindo-se à informação que pode ser deduzida diretamente das imagens e que não pode ser representada pela cor ou textura, ou seja, define forma como um espaço complementar à cor e à textura. Além disso, afirmam que uma poderosa maneira de representar a forma é através de características geométricas como bordas, contornos, junções e regiões poligonais extraídas de uma imagem.

Entretanto, a aplicação destes atributos em sistemas CBIR ainda é muito limitada, devido à alta complexidade da segmentação de imagens, principalmente em cenas complexas onde estão presentes ruídos e oclusão de objetos (OLIVEIRA, 2006; AKGÜL et al., 2011).

No contexto médico, a forma é um dos fatores mais importantes na detecção de certas lesões e no entendimento acerca do seu quadro de evolução. Assim, descritores baseados na forma dos objetos são muito úteis por fornecerem detalhes finos ao processo de recuperação de imagem (XU et al., 2012; AKGÜL et al., 2011).

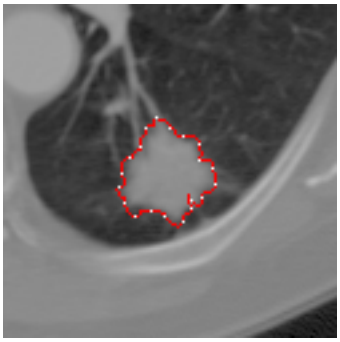
A nitidez da borda de uma lesão pulmonar identificada através de imagens de exames radiológicos é uma das categorias dos atributos de forma. Ela é um importante aspecto utilizado pelos especialistas para diferenciar lesões em termos de sua malignidade, porque ela varia dependendo do tipo de lesão (XU et al., 2012).

Neste trabalho, será utilizada uma análise tri-dimensional da borda apresentada por Ferreira Junior (2015), que é parcialmente baseada na metodologia apresentada por Xu et al. (2012). A implementação do algoritmo apresentado por Ferreira Junior (2015) parte de linhas traçadas nas bordas de todas as fatias dos nódulos dos exames de TC. Estas linhas são traçadas em pontos de controle automaticamente selecionados nas marcações das bordas da lesão (Figura 10(a)). Os pontos de controle são marcados a distância $d = \frac{p}{c}$, onde p é o número de *pixels* da borda e c é a quantidade de pontos de controle, que Xu et al. (2012) e Ferreira Junior (2015) definiram como 20 em seus respectivos trabalhos. Contudo, nem Xu et al. (2012), tampouco Ferreira Junior (2015), em seus respectivos trabalhos, demonstraram porque foi utilizada esta quantidade de pontos de controle. Em seguida, linhas perpendiculares à borda são, então, desenhadas em cada um dos pontos de controle colocados na borda dos nódulos (Figura 10(b)).

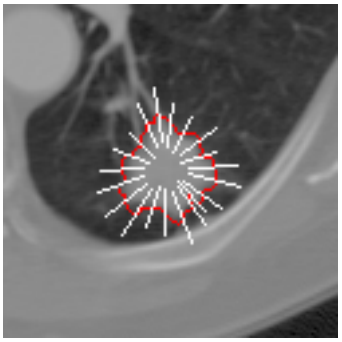
Xu et al. (2012) afirmam que a partir destas linhas torna-se possível capturar dois atributos que definem a nitidez da borda: diferença de intensidade, que é a diferença entre a intensidade do tecido do órgão ao redor da lesão e o tecido de dentro do lesão; e

Figura 10 – Determinação dos pontos de controle e demarcação dos segmentos de retas ao redos dos nódulos pulmonares apresentado inicialmente em.

(a) Pontos de controle demarcados na borda dos nódulos.



(b) Linhas perpendiculares traçadas sobre os pontos de controles ao redor da borda dos nódulos.



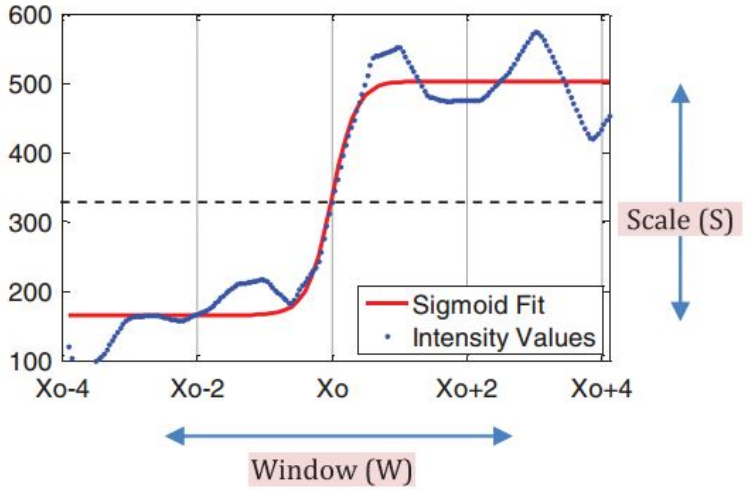
Fonte: Xu et al. (2012).

o *blur* da borda, que mede a mudança abrupta na transição entre a intensidade do tecido da lesão e o tecido do órgão ao redor da lesão.

Após a extração dessas linhas, Xu et al. (2012) gravam os valores de intensidades ao dos segmentos de reta usando interpolação bilinear. Em seguida, eles aplicam uma função sigmoide para ajustar os valores usando uma função de regressão não-linear ponderada (Figura 11). A partir da função sigmoide são calculados dois valores, que são utilizados para caracterizar cada segmento de reta: *window* e *scale*. O primeiro caracteriza o *blur* da borda pela medida da transição do tecido da lesão para o tecido ao redor da lesão. O segundo mede a diferença de intensidades dentro e fora da lesão. A nitidez da borda é então representada por um vetor de características composto por dois histogramas de 30 *bin* cada um. *Bin* corresponde à quantidade de valores contida em um intervalo de números, essa é uma forma de agrupar os valores e apresentá-los em um histograma. Em seguida, são removidas todas as linhas que ultrapassam as bordas do órgão, pois essas carregam informações que não satisfazem propriedades dos atributos escolhidos para representar os nódulos. Este método teve seus resultados avaliados utilizando o índice *Normalized Discounted Cumulative Gain* (NDCG) alcançado 84% de precisão.

A metodologia proposta por Ferreira Junior (2015) difere apresentada por Xu et al. (2012), porque ao invés de calcular os descritores apontados por estes autores, aquele autor faz análises estatísticas a partir das intensidades dos pixels dos segmentos de reta demarcados ao redor das bordas das lesões. Assim, o vetor de características resultante deste processo é composto pelos atributos listados nas Equações 2.6 - 2.17, onde A é o vetor de intensidades dos *pixels* de tamanho n, A_1 é a valor de intensidade de um pixel fora do nódulo e A_n é a intensidade do *pixel* dentro do nódulo. Portanto, cada nódulo é

Figura 11 – Exemplo da aplicação da função sigmoide sobre as intensidades dos pixels mostrando os atributos *window* e *scale* (XU et al., 2012). As intensidades dos *pixels* contidas no segmento de reta de um ponto de controle estão representadas em azul. Já o resultado da aproximação de função da sigmoide está representado em vermelho.



Fonte: Xu et al. (2012).

caracterizado por um vetor de características de nitidez de borda de doze dimensões.

$$\text{Diferença dos valores extremos} = A_n - A_1, \tag{2.6}$$

$$\text{Soma dos valores} = \sum_{i=1}^{n} A_i, \tag{2.7}$$

$$\text{Soma dos quadrados} = \sum_{i=1}^{n} A_i^2, \tag{2.8}$$

$$\text{Soma dos logarítmos} = \sum_{i=1}^{n} \log A_i, \tag{2.9}$$

$$\text{Média aritmética } (\mu) = \frac{1}{n} \sum_{i=1}^{n} A_i, \tag{2.10}$$

$$\text{Média Geométrica} = \sqrt[n]{\prod_{i=1}^{n} A_i}, \tag{2.11}$$

$$\text{Variância da população} = \frac{1}{n} \sum_{i=1}^{n} (A_i - \mu)^2, \tag{2.12}$$

$$\text{Variância da amostra } (v) = \frac{1}{n-1} \sum_{i=1}^{n} (A_i - \mu)^2, \tag{2.13}$$

$$\text{Desvio padrão } (s) = \sqrt{v}, \tag{2.14}$$

$$\text{Medida Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(A_i - \mu)^4}{s^4}, \tag{2.15}$$

$$\text{Medida de assimetria} = \frac{\frac{1}{n}\sum_{i=1}^{n}(A_i - \mu)^3}{s^3}, \tag{2.16}$$

$$\text{Segundo momento central} = \frac{\frac{1}{n}\sum_{i=1}^{n}(A_i - \mu)^2}{s^2}. \tag{2.17}$$

2.6 Normalização de valores

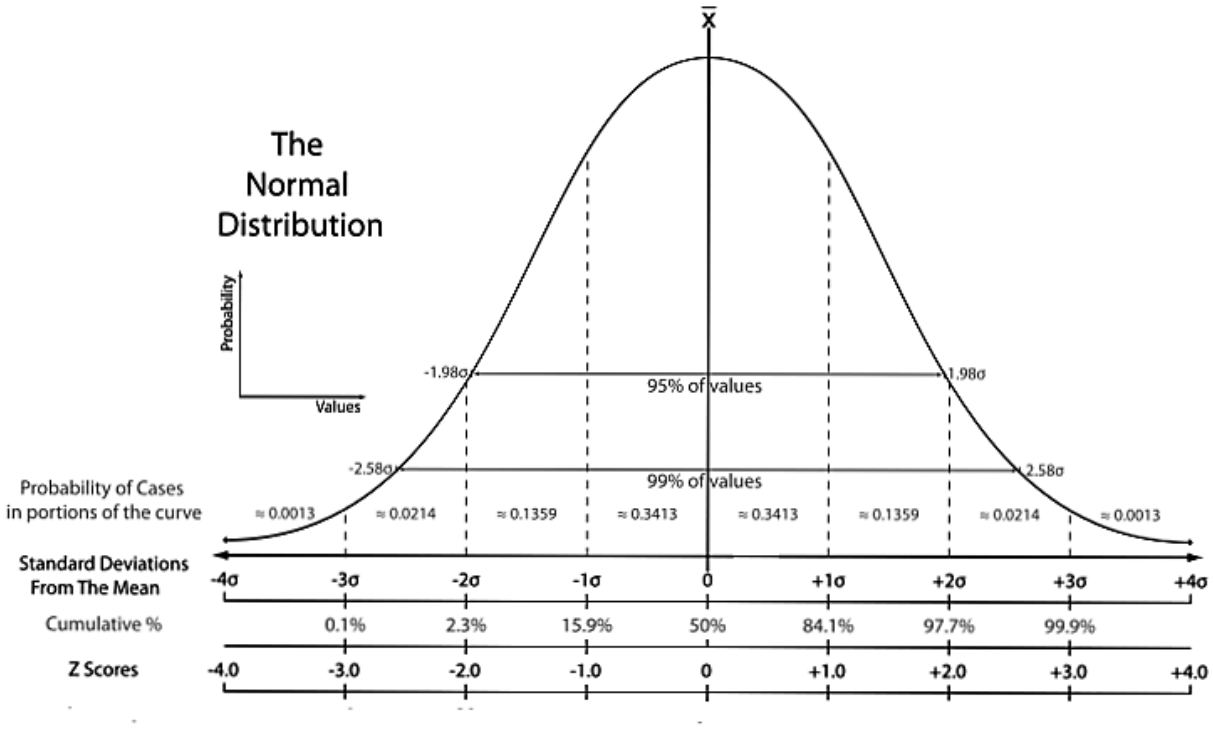
Cada atributo extraído tem seu próprio intervalo de valores (escala), os quais não são necessariamente coincidentes. Para utilizar uma métrica de similaridade baseada em distância, é preciso normalizar os dados para colocá-los em uma escala específica, visto que métricas de similaridades são sensíveis a diferenças em escalas (VISALAKSHI; THANGAVEL, 2009). Como neste trabalho é utilizada a DEP para comparar nódulos representados por vetores formados por diferentes atributos com diferentes escalas de valores, é necessária a normalização dos valores para que seja possível uma comparação justa entre eles (MILLIGAN; COOPER, 1988).

AL SHALABI, Shaaban e Kasasbeh (2006) afirmam que existem muitos métodos para a normalização de dados, dentre eles: normalização Min-Max, que executa uma transformação linear nos dados originais; normalização *Z-score*, que normaliza os valores iniciais baseado na média e no Desvio Padrão (DP) da amostra; e normalização por escala decimal, que normaliza alterando a escala movimentando o ponto decimal dos valores da amostra.

Z-score é um dos métodos mais utilizados na literatura para normalização de valores. A sua escala de valores está delimitada no intervalo $[-3, +3]$, e os valores são determinados segundo o DP e a média dos valores da amostra. O valor *Z-score* permite identificar onde um valor em particular está (acima ou abaixo) com relação a média na curva de distribuição normal da amostra normalizada (Figura 12). (VISALAKSHI; THANGAVEL, 2009; KENROSE, 2015).

A normalização *Z-score* é definida pela Equação 2.18, sendo A' o vetor com a distribuição normal e valores no intervalo; A é o vetor com os valores originais dos atributos;

Figura 12 – Distribuição normal da escala da normalização. Nela é possível verificar que a distribuição dos valores a partir da média (ponto zero) se dá em quantidades positivas e negativas do DP (σ) *Z-score*.



Fonte: Kenrose (2015).

$\overline{x}$ é a média dos valores; e σ é o DP da amostra.

$$A' = \frac{A - \overline{x}}{\sigma}, \tag{2.18}$$

2.7 Descrição da base de imagens

A base de nódulos utilizada neste trabalho é uma base não relacional orientada a documentos que extende o conteúdo da base de imagens do projeto *Lung Image Database Consortium* (LIDC) (ARMATO et al., 2011) através da adição dos AT 3D e ANB 3D apresentada por Ferreira Junior (2015, p. 43) (melhor descritos nas Seções 2.5.1 e 2.5.2).

O projeto LIDC foi iniciado pelo *National Cancer Institute* (NCI) dos Estados Unidos da América, continuado pelo *Foundation for the National Institute of Health* (FNIH) e acompanhado pelo *Food and Drug Administration* (FDA). É a maior fonte de nódulos pulmonares do mundo contendo mais de 1 mil casos de pacientes, cada caso inclui imagens de TC e informações associadas à localização das bordas dos nódulos em cada fatia do volume de imagens, além de nove características dos nódulos descritas por, pelo menos, quatro experientes radiologistas. Em todas as imagens, as lesões foram segmentadas utilizando as marcações feitas pelos especialistas. O LIDC tem se tornado um recurso essencial para pesquisas de imagens médicas por incentivar o desenvolvimento de sistemas

CADe, CADx e servir para validação e disseminação de diversas metodologias (HAN et al., 2015).

De acordo com as regras que definiram a construção da base de imagens, os exames de TC estão disponíveis no padrão *Digital Imaging and Communications in Medicine* (DICOM), com informações a respeito das posições dos *pixels* das bordas dos nódulos em cada fatia do volume, características dos nódulos determinadas pelo especialista (calcificação, estrutura interna, lobulação, margem, esfericidade, espiculação, sutileza, textura), e a probabilidade de malignidade dos nódulos que varia de 1 a 5 (do benigno ao maligno) (ARMATO et al., 2011), como podemos observar:

- Malignidade 1 : probabilidade alta de ser benigno;
- Malignidade 2 : probabilidade moderada de ser benigno;
- Malignidade 3 : malignidade indeterminada;
- Malignidade 4 : probabilidade moderada de ser maligno;
- Malignidade 5 : probabilidade alta de ser maligno;

As lesões resultantes do processo de seleção definido no projeto LIDC foram classificadas em três categorias:

- Nódulos $<$ 3mm: lesões que possuem tamanho inferior a 3 milímetros;
- Nódulos $\geq$ 3mm: lesões que possuem tamanho igual ou superior a 3 milímetros, classificadas independente de histologia, desta forma, elas podem ser câncer de pulmão primário, metástase, processo não cancerígeno ou de natureza indeterminada;
- Não-nódulos $\geq$ 3mm: lesões que possuem tamanho igual ou superior a 3 milímetros, mas que não possuem características de nódulo pulmonar. Não-nódulos menores que 3 milímetros foram descartados.

Apenas os nódulos maiores que 3mm são de interesse neste trabalho, cujo objetivo é utilizar informações extraídas deles para auxiliar na tomada de decisão. A partir da segmentação destes nódulos, Ferreira Junior (2015) extraiu os AT 3D e os ANB 3D para caracterizar os nódulos pulmonares estendendo, assim, o conteúdo disponibilizado pelo projeto LIDC.

A tecnologia utilizada para disponibilizar a base foi a *Not only Structured Query Language* (NoSQL), ou não relacional orientada a documentos. Foi utilizado a base de dados MongoDB na versão 2.4.6 para a disponibilização em duas formas: uma para acesso local e outra para acesso remoto em uma nuvem pública. O Apêndice A possui mais informações de acesso e utilização da base de imagem aqui utilizada.

2.8 Métrica de similaridade

Um dos maiores desafios para os sistemas CBIR é como definir apropriadamente a forma de avaliação de similaridade usada para indexar a base de dados e/ou fazer o *ranking* baseado na similaridade de imagens recuperadas segundo um determinado critério de busca (AKGÜL et al., 2011). Isso porque a precisão na recuperação de imagens é fortemente influenciada não apenas pelos atributos escolhidos para representar os objetos, mas também pela medida de similaridade utilizada (SILVA, 2009). O que leva a necessidade de definir qual a função de distância permite recuperar as imagens mais similares segundo o domínio do espaço de busca (BEDO et al., 2015).

Um método comum é empregar distância vetorial em espaço multidimensional, comumente um espaço euclidiano, no qual cada imagem é representada através de vetores de descritores/atributos (AKGÜL et al., 2011). Neste contexto, basicamente, todos os sistemas usam a suposição de que há equivalência entre a imagem e sua representação vetorial através dos atributos. Estes sistemas, muitas vezes, usam métricas que são facilmente entendíveis para medir a distância entre a imagem de referência e as possíveis imagens similares encontradas como resultado das buscas. Todas elas representadas por vetores de características em um espaço n-dimensional (MüLLER et al., 2004).

Uma função de distância $d()$ avalia a distância, ou dissimilaridade, entre um par de elementos e ela deve atender as seguintes propriedades (BEDO et al., 2015; SILVA, 2009), onde $\mathbb{V}$ é um vetor n-dimensional definido no espaço euclidiano:

1. Simetria: $\{\forall v \in \mathbb{V} | d(v_1, v_2) = d(v_2, v_1)\}$;
2. Não negatividade: $\{\forall v \in \mathbb{V} | 0 \leq d(v_1, v_2) < \infty\}$;
3. Desigualdade triangular: $\{\forall v \in \mathbb{V} | d(v_1, v_2) \leq d(v_1, v_3) + d(v_3, v_2)\}$.

Intuitivamente, menores distâncias correspondem à maior similaridade. Sendo assim, quanto mais próximo de zero o valor da distância, maior será a similaridade dos objetos segundo os critérios utilizados pelos descritores da imagem. Inversamente, quanto maior o valor, menor será a similaridade (TRAINA et al., 2003).

Silva (2009) apresenta a Família de Distâncias *Minkowski* (FDM) que é muito utilizada como métrica de comparação no domínio $\mathbb{V}$. Esta família é definida como L_p:

$$L_p(v_1, v_k) = \sqrt[p]{|v_1 - v_k|^p} \tag{2.19}$$

Da família L_p a mais conhecida é a L_2, que é a DE:

$$L_2(v_1, v_k) = \sqrt[2]{|v_1 - v_k|^2} \tag{2.20}$$

Silva (2009) e Atkeson, Moore e Schaal (1997, p. 25) apresentam uma variação da DE que permite a manipulação dos pesos dos atributos que compõem um vertor $v \in \mathbb{V}$. Essa é a DEP.

A DEP é calculada segundo a Equação 2.21, onde $\vec{v_1}$ é o vetor de atributos de referência, $\vec{v_2}$ é de comparação, e $\vec{w}$ é o conjunto de pesos associados a cada um dos atributos que compõem $\vec{v_1}$ e $\vec{v_2}$.

$$d(\vec{v_1},\ \vec{v_2}) = \sqrt{\vec{w}\ (\vec{v_1} - \vec{v_2})^2}, \tag{2.21}$$

Os pesos representam a influência dos atributos no processo de recuperação do sistema CBIR e identificam quais atributos carregam a informação mais relevante para a classificação das lesões. Isso permite alcançar melhores resultados na precisão do diagnóstico médico por prover resultados mais precisos em algoritmos de recuperação de imagens (FACELI et al., 2011).

2.9 Atualização local e global de pesos

Na maioria dos métodos de aprendizagem, um único modelo global é utilizado para ajustar toda a base de treinamento, enquanto modelos locais tentam ajustar a base de treinamento apenas na região ao redor do ponto de referência da busca. Alguns exemplos de algoritmos de aprendizagem local são *K-Nearest Neighbor* (KNN), *Weighted Average* (WA) e *Locally Weighted Regression* (LWR). Cada um destes modelos combinam objetos próximos do objeto de referência para estimar a saída apropriada. Modelos KNN utilizam os objetos mais próximos do objeto de referência para determinar o valor de saída. WA atribui pesos para os objetos próximos ao objeto de referência que são inversamente proporcionais à distância entre eles no espaço n-dimensional. Já LWR ajustam os objetos próximos por meio de uma regressão de distância ponderada (ATKESON; MOORE; SCHAAL, 1997).

A proposta aqui apresentada possui características apontadas por Atkeson, Moore e Schaal (1997) as quais se referem a *Locally Weighted Learning* (LWL). De forma conscisa, LWL é um conceito que se refere a sistemas que aprendem tardiamente (*lazy learning*) com o objetivo de construir modelos resultantes de aproximação de função por meio de ajuste de pesos em funções polinomiais. O objetivo do LWL é enfatizar dados que são similares ao objeto de referência, e desenfatizar os dados que são dissimilares, ao invés de tratar todos os dados de forma igual. Os requisitos citados por Atkeson, Moore e Schaal (1997) para que um sistema seja do tipo LWL são:

- **Função de distância** - sistemas LWL requerem uma medida de relevância. A principal suposição feita pelo LWL é que a relevância pode ser mensurada usando uma medida de distância;

- **Critério de separação** - sistemas LWL calculam pesos a partir de cada objeto de treinamento;
- **Objetos classificados** - cada objeto precisa ter associado a si uma saída apropriada. Para modelos de classe, a saída deve ser uma classificação (*label*). Para modelos de regressão, a saída deve ser um valor esperado;
- **Representação** - cada objeto deve ser representado por um vetor de valores de tamanho fixo (simbólico ou numérico) para uma lista específica de características.

2.10 Métodos de avaliação

Muitos métodos diferentes para avaliação de sistemas CBIR foram definidos e utilizados por pesquisadores até o momento (VOGEL; SCHIELE, 2006; AKGÜL et al., 2011). Müller et al. (2001) cita alguns métodos de avaliação de recuperação que são: *User comparison*, *Rank of the best match*, *Average rank of relevant images*, Precisão × Revocação (PR), *Target testing*, *Error rate*, *Retrieval eficiency* e *Correct and incorrect detection*. Além desses, o refirido autor apresenta algumas formas gráficas para avaliação de recuperação que são: o gráfico PR; Precisão *versus* o número de imagens recuperadas e Revocação *versus* o número de imagens recuperadas; *Correctly retrieved versus* todas as imagens recuperadas; e, por fim, *Retrieval accuracy versus* o ruído na recuperação. Já Järvelin e Kekäläinen (2002) apresentam outros métodos de avaliação baseadas em medidas de ganho acumulado como, por exemplo: *Direct Cumulated Gain*; *Discounted Cumulated Gain*; *Average Search Length*; *Expected Search Length*; *Normalized Recall Measure*; *Ranked Half-Life Measures*; *Relative Relevance*; e outros.

Além das várias formas de avaliar uma coleção de objetos recuperados, as pesquisas feitas, embora para áreas afins como medicina, utilizam bases de dados diferentes, o que aumenta ainda mais o leque de problemas para definir um padrão de comparação entre os trabalhos. A necessidade de padronização da avaliação é clara dado que muitas métricas são variações pequenas de uma mesma definição (MÜLLER et al., 2001; VOGEL; SCHIELE, 2006; AKGÜL et al., 2011).

Segundo Vogel e Schiele (2006), a avaliação da *performance* é indispensável por permitir a comparação entre diferentes sistemas e a análise de como está relacionada a *performance* dos sistemas ao contexto no qual ele é aplicado. A avaliação de diferentes algoritmos e sistemas também permite a integração de abordagens com o intuito de construir ferramentas mais poderosas. Para eles, o objetivo da avaliação da *performance* se refere à análise da qualidade, e não à velocidade de um algoritmo.

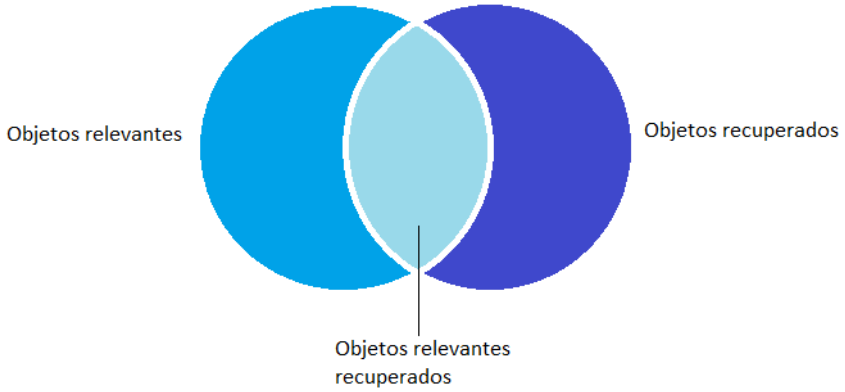
Dentre os tipos de avaliação existentes, o que é mais comumente utilizado para avaliar sistemas de recuperação de informação é o método PR e sua representação gráfica. Os pesquisadores estão familiarizados com o modelo gráfico do PR e conseguem extrair informações dele sem muitos problemas de interpretação (MÜLLER et al., 2001).

Vogel e Schiele (2006) definem Precisão (Equação 2.22) como a porcentagem dos objetos recuperados que são relevantes, e Revocação (Equação 2.23) como a porcentagem de objetos relevantes recuperados. Objeto relevante é aquele que é de interesse em um determinado contexto (Figura 13).

$$\text{Precisão} = \frac{\text{Nº de objetos relevantes recuperados}}{\text{Nº total de objetos recuperados}} \tag{2.22}$$

$$\text{Revocação} = \frac{\text{Nº de objetos relevantes recuperados}}{\text{Nº total de objetos relevantes}} \tag{2.23}$$

Figura 13 – Representação do PR na recuperação de objetos.



Fonte: Elaborada pelo autor.

Müller et al. (2001) afirmam que Precisão e Revocação sozinhos não contêm informações suficientes. Sendo assim, eles devem ser utilizados em conjunto. Também os gráficos do PR podem não conter todas as informações desejadas, assim, outras medidas podem ser utilizadas como, por exemplo, o Precisão(n) (PN) que mede a precisão após a recuperação de n objetos, a *Mean Average Precision*, *Recall at 0.5 precision*, $R(n)$ que mede a Revocação após a recuperação de n objetos, e *Rank first relevant*. Outro fator negativo dos gráficos do PR apontado por Müller et al. (2001) é que o resultado do gráfico é dependente do número de objetos relevantes para um dado critério de busca.

2.11 Trabalhos relacionados

Xu et al. (2012) descrevem um algoritmo que extrai atributos a partir das informações da nitidez da borda dos nódulos. Como explicado na Seção 2.5.2, estes autores representam os nódulos por meio de vetores construídos a partir de histogramas dos atributos *window* e *scale* calculados a partir da função sigmoide.

Ferreira Junior (2015) utilizou a base de nódulos LIDC 1.171 nódulos para apresentar um sistema CBIR cujos nódulos foram descritos vetorialmente pelos AT 3D e ANB 3D e como métrica de similaridade a DE. A forma de avaliação se deu por meio do cálculo

da Precisão. Ele fez testes utilizando os AT 3D, ANB 3D e AI (Seção 2.5) para 10 recuperações com 10 nódulos cada uma ($P(10)$), e aplicou o conceito de classe binária (nódulos ou benignos ou malignos) resultando em 745 nódulos benignos e 426 nódulos malignos.

Dhara et al. (2012) apresentam uma proposta de sistema CBIR para recuperar nódulos pulmonares sólidos com tamanho entre 3mm e 30mm. Foram utilizados 30 exames selecionados do projeto LIDC e 25 exames da base de nódulos do PGIMER *Chandigarh* onde cada conjunto de exames contém 40 nódulos pulmonares sólidos, em um total de 80 nódulos, para avaliação do desempenho do sistema. Para representar os nódulos foram calculados atributos de forma (esfericidade, índice de lobulação, índice de espiculação, distância radial média, índice de calcificação e 3D *accutance* da superfície do nódulo) e textura (contraste, entropia e tendência do *cluster*, homogeneidade e a classificação da textura do tecido interno) a partir dos nódulos reconstruídos em 3D. Também foram utilizadas as nove características associadas aos nódulos do projeto LIDC (textura, sutileza, espiculação, lobulação, esferecidade, borda, malignidade, estrutura interna e calcificação). O mesmo protocolo utilizado pelos radiologistas para extrair as características dos nódulos LIDC foi aplicado à base PGIMER *Chandigarh* com o objetivo de extrair as mesmas informações. Segundo estes autores, algumas características são úteis para representar as imagens e outras não. Sendo assim, eles usaram regressão logística para encontrar o subconjunto de atributos que permitiam maior discriminação usando o critério de relevância máxima e redundância mínima (os autores não apontaram os atributos que proporcionaram maior discriminação). Por fim, a métrica de similaridade utilizada para recuperar e construir o *ranking* de nódulos recuperados foi a DE.

Seitz Jr et al. (2012) descrevem um sistema CBIR em combinação com Algoritmos Genéticos para determinar a combinação ótima de atributos de imagem para aumentar a precisão na recuperação de nódulos similares. Eles utilizaram 399 exames da base de nódulos do LIDC contendo 914 nódulos distribuídos entre as malignidades 1, 2, 4 e 5. Foram utilizados 63 atributos extraídos a partir da textura (usando filtro *Gabor*, *Markov Random Fields* e os atributos propostos por Haralick a partir da MCO), tamanho, forma e intensidade para representar vetorialmente os nódulos. Os atributos foram normalizados uilizando o método *Z-score*. A métrica de similaridade usada foi a DE. As malignidades dos nódulos foram agrupadas da seguinte forma: malignidades 1 e 2 foram classificados como benignos; e malignidades 4 e 5 foram classificados como malignos.

Kuruvilla e Gunavathi (2014) apresentam outro trabalho onde é utilizado um sistema CBIR para recuperar exames, com nódulos pulmonares semelhantes, com o objetivo de encontrar o conjunto de atributos que melhor descrevem os nódulos, segundo os parâmetros usados para calcular a acurácia no algoritmo de Rede Neural. Além disso, os autores avaliaram diferentes métricas de similaridade para identificar aquela que apresenta maior precisão na recuperação dos nódulos. Para tal, usaram imagens de 180 exames conti-

dos no projeto LIDC. Dois conjuntos de atributos foram calculados a partir dos nódulos: atributos da MCO e atributos estatísticos. Os atributos calculados com a MCO foram energia, entropia, dissimilaridade, contraste, diferença inversa, correlação, homogeneidade, autocorrelação, *cluster shade*, *cluster proiminence*, probabilidade máxima, soma dos quadrados, *sum average*, soma da variância, soma da entropia, diferença da variância, diferença da entropia, informação da medida de correlação, coeficiente de correlação máxima, diferença inversa normalizada e momento de diferença inversa normalizada. Os atributos estatísticos calculados foram a média, o DP, assimetria e *kurtosis*. Dentre todos os atributos calculados, os que foram selecionados como mais relevantes pela Rede Neural foram: autocorrelação, contraste, correlação, *cluster shade*, *cluster proiminence*, dissimilaridade, energia, entropia, homogeneidade, soma da variância e assimetria. As métricas de similaridades utilizadas foram a DE, Distância *Manhattan*, Distância *City Block*, Distância *Chebychev*, Distância *Tversky*, Distância *Canberra*, *Bray-Curtis*, Distância *Chi Squared* e Distância *Squared Chord*.

Outra comparação pode ser feita com o trabalho apresentado por Han et al. (2015), apesar destes autores, assim como Xu et al. (2012), utilizarem outra forma de avaliação. A comparação dos resultados avaliados com outro método se dá devido ao uso dos atributos de textura como uma das formas de representação dos nódulos pulmonares.

Han et al. (2015) apresentaram uma avaliação de classificação por um modelo preditivo construído a partir do algoritmo *Support Vector Machine* (SVM) com um *kernel Radial Basis Function* (RBF) utilizando a Área Sob a Curva (AUC). A base de imagens utilizada foi a do LIDC. Os nódulos tiveram suas malignidades (de 1 a 5) mapeadas em um conjunto binário de classes (maligno e benigno) em dois cenários diferentes. O primeiro cenário considerou as malignade 1, 2 e 3 como benigno. Já o segundo cenário considerou as malignades 3, 4 e 5 como maligno. A avaliação do modelo foi feita através do cálculo da sensibilidade e da especificidade e apresentada por meio da AUC. Os atributos por eles utilizados foram classificados em três tipos: Atributos de Textura; Atributos de Gabor; e Padrões Locais Binários. No caso dos Atributos de Textura, eles foram calculados a partir da MCO 2D e 3D. No primeiro cenário, os Atributos de Textura, combinando 2D e 3D, alcançaram uma AUC média de 91,24%; os Atributos de Gabor, AUC média de 91%; e Padrões Locais Binários, AUC média de 90,12%. Já no segundo cenário, os Atributos de Textura, combinando 2D e 3D, alcançaram uma AUC média de 78,10%; Atributos de Gabor, AUC média de 80,45%; e Padrões Locais Binários, AUC média 77,99%. Vale ressaltar que esta análise não levou em consideração um sistema CBIR, e sim um modelo de classificação.

Os trabalhos apresentados nesta Seção fazem um breve resumo dos trabalhos voltados para construção de sistemas CBIR que utilizam o LIDC como base de nódulos e fazem a representação dos nódulos através de vetores de atributos. Embora existam trabalhos no sentido de medir pesos, ou determinar a influência, dos atributos e outros que fazem

análises voltadas para a forma de medição da similaridade dos nódulos, nenhum deles apresenta uma proposta onde a métrica de similaridade seja utilizada juntamente com os conceitos de pesos dos atributos. É neste ponto onde o trabalho aqui proposto se localiza, por trazer um método de ajuste de pesos que está diretamente ligado à métrica de similaridade para buscar aumentar a precisão na recuperação de nódulos pulmonares.

3 MATERIAIS E MÉTODOS

Este capítulo está organizado da seguinte forma: a Seção 3.1 apresenta como a base de imagens foi utilizada; a Seção 3.2 define a forma como os atributos foram normalizados para que pudessem ser utilizados em funções de distância; a Seção 3.4 apresenta uma visão holística do processo de atualização local de pesos composto por duas fases (Avaliação e Treinamento), que é o cerne deste trabalho; a Seção 3.4.1 explica em detalhes a Fase de Avaliação; e, por fim, a Seção 3.4.2 traz uma explicação detalhada da Fase de Treinamento.

3.1 Base de imagens

A base de imagens usada neste trabalho possui 752 exames e 1.944 nódulos pulmonares. Ela foi descrita em detalhes na Seção 2.7. Dentre os nódulos contidos na base, aqueles que possuem malignidade 3 foram descartados por estarem associados a uma malignidade indeterminada. Sendo assim, dos 1.944 nódulos iniciais, foram escolhidos os 1.171 nódulos com as malignidades 1, 2, 4 e 5 (Tabela 3).

Tabela 3 – Quantidade de nódulos por malignidade.

Probabilidade de malignidade	1	2	4	5	Total
Quantidade de nódulos	273	472	266	160	1.171

Fonte: Elaborada pelo autor.

A partir destes 1.171 nódulos, Ferreira Junior (2015) extraiu 48 atributos para representar nódulos. Eles estão distribuídos em 12 ANB 3D e 36 AT 3D:

- **AT 3D** - energia, entropia, matiz, momento da diferença inverso, contraste, proeminência, correlação, variância e homogeneidade em 4 direções diferentes da MCO 3D (0^o, 45^o, 90^o e 135^o);
- **ANB 3D** - diferença entre os extremos, soma dos valores, soma dos quadrados, soma dos logarítmos, média aritmética, média geométrica, variância da população, variância da amostra, desvio padrão, medida de Kurtosis, medida de assimetria e segundo momento central.

Estes atributos foram extraídos tendo como referência as regiões dos nódulos segmentadas manualmente pelos especialistas envolvidos no projeto LIDC. A região de interesse nas imagens dos exames é a região demarcada nas bordas dos nódulos segmentados. Desta forma, foram utilizados os tamanhos originais dos nódulos contidos nos exames.

Para as Fases de Avaliação (Seção 3.4.1), Treinamento (Seção 3.4.2) e Validação a base de nódulos foi dividida da seguinte maneira:

- **Avaliação** - composta de 65 nódulos de cada malignidade resultando em 260 nódulos;
- **Treinamento** - composta de 65 nódulos de cada uma das 4 malignidades resultando em 260 nódulos;
- **Validação** - composta de 30 nódulos de cada malignidade resultando em 120 nódulos.

A determinação do tamanho das bases citadas anteriormente está relacionada a menor quantidade de nódulos dentre as malignidades utilizadas, ou seja, tomou-se como referência a quantidade de nódulos da malignidade 5 que é de 160 (Tabela 3) e dividiu-se este valor em três partes como explicado acima. Como as outras malignidades possuem um número maior de nódulos, delas foram selecionadas as mesmas quantidades de nódulos de forma aleatória sendo descartados os demais.

As duas primeiras bases foram utilizadas durante a Avaliação e o Treinamento para definir os pesos, respectivamente, enquanto a terceira base foi utilizada para validar os melhores pesos encontrados. Todos os nódulos foram escolhidos de forma aleatória e as bases são disjuntas, ou seja, não existe nódulo que pertença a mais de uma base ao mesmo tempo.

3.2 Normalização da base de atributos

Neste trabalho, foi adotado o método de normalização *Z-score* (Função 2.18) apresentado na Seção 2.6, devido ao seu uso recorrente como método de normalização; e por ela ser indicada em casos onde se conhece os valores mínimo e máximo da amostra e que seja possível calcular a média e o DP da amostra (AL SHALABI; SHAABAN; KASASBEH, 2006).

A normalização foi feita sobre todos os valores dos atributos contidos nas bases de Avaliação, Treinamento e Validação de forma conjunta, ou seja, a normalização foi aplicada a partir de todos os atributos dos nódulos como se fizessem parte de uma única base. Desta forma, todos os nódulos envolvidos no processo de atualização de pesos são utilizados para determinar os valores de média e DP da amostra.

3.3 Métricas de similaridade

Neste trabalho foi adotada a DEP (Seção 2.8) para realizar a comparação entre os nódulos pulmonares representados pelos vetores de atributos (AT 3D e ANB 3D). Ela foi escolhida devido ao uso na literatura voltada para CBIR relacionada à métricas de similaridade (OLIVEIRA; CIRNE; AZEVEDO-MARQUES, 2007; AKGÜL et al., 2011) e por permitir inserir de forma clara o conceito da atualização de pesos aqui proposta.

Os pesos da DEP são determinados pelo Inverso do Desvio Padrão (IDP) dos valores dos atributos sobre a recuperação de nódulos similares (este método será descrito detalhadamente na Seção 3.4.2). Isso leva à seguinte indução: valores maiores são atribuídos à atributos que tem valores homogêneos e valores menores são atribuídos à atributos cujos valores são heterogêneos. Por exemplo, se *energia* tem baixa variabilidade nos valores dos nódulos recuperados de uma malignidade específica, a ela será atribuído um peso maior do que outro que tem alta variabilidade. Isso é possível devido ao uso do IDP para calcular o valor do peso dos atributos.

3.4 Processo de atualização local de pesos

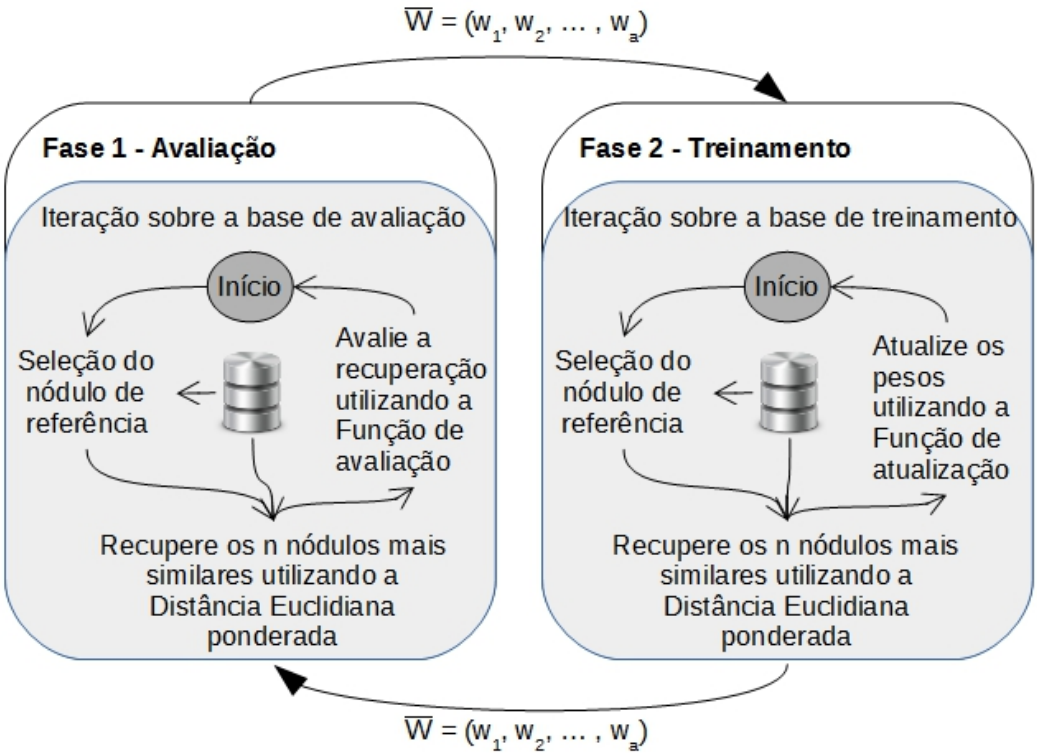
O processo apresentado neste trabalho é composto por duas fases que são executadas de forma sequencial e cíclica: Fase de Avaliação (Seção 3.4.1) e Fase de Treinamento (Seção 3.4.2).

As fases possuem estruturas semelhantes que consistem basicamente em iterações sobre as bases de dados correspondentes, selecionando cada um dos nódulos armazenados e utilizando-os como nódulo de referência para a recuperação dos nódulos mais semelhantes usando a DEP. Ao término do ciclo de execuções das fases, obtém-se um conjunto de pesos de atributos que permite melhores resultados na recuperação de nódulos semelhantes (Figura 14). O ciclo deve ser executado até que o critério de parada seja alcançado. O critério adotado foi que o processo parasse caso se passassem 100 iterações de treinamento e avaliação sem que houvesse melhora nos resultados da avaliação. Ou seja, após obter um valor máximo da Função de Avaliação após i iterações, caso este valor não seja aumentado nas 100 iterações subsequentes, o sistema para o processo de ajuste de pesos e aponta aquele que alcançou a melhor avaliação como sendo o conjunto de pesos ideal para o vetor de atributos.

O conjunto de pesos que obteve os melhores resultados é, então, validado em uma terceira base de nódulos (base de validação), que não foi utilizada durante o processo de ajuste de pesos, isso é realizado para avaliar se o conjunto de pesos está super-ajustado à base de validação, o que levaria a resultados tendenciosos.

O método de iteração na base de nódulos adotado remete ao método conhecido como *Leave-One-Out* (LOO), que é uma variação do *Cross-validation*, onde uma base é dividida em uma parte de treinamento e outra de validação (em casos de modelos de predição ou regressão). A parte de treinamento é utilizada para criar o modelo preditivo, enquanto a parte de validação é utilizada para avaliar o modelo criado. No caso específico do LOO, a parte de validação tem tamanho 1 e o restante da base é utilizada para treinamento do modelo, isso é repetido para todos os objetos na base (FACELI et al., 2011). O objetivo de utilizar o LOO é garantir que todos os nódulos serão utilizados para determinar os pesos e também para avaliá-los.

Figura 14 – *Workflow* do processo de atualização de pesos.



Fonte: Elaborada pelo autor.

O Algoritmo 1 é o pseudo-código do processo de atualização de pesos. Como resultado dele, são retornados os pesos que melhor ajustam a recuperação, os quais devem ser utilizados junto à métrica de similaridade utilizada para fazer a comparação entre os nódulos.

Algorithm 1: Pseudo-código do processo de atualização automática de pesos.

Data: baseAvaliacao:[ ][ ], baseTreinamento:[ ][ ], baseValidacao:[ ][ ], a:int
Result: W:[real]
$W[i] = 1/a$
$WCorrente[i] = 1/a$
while *criterioParada* **do**
  v_1 = avaliacaoPesos(baseAvaliacao, W)
  $WCorrente$ = treinamento(baseTreinamento, $WCorrente$)
  v_2 = avaliacaoPesos(baseAvalicao, $WCorrente$)
  if $v_1 \leq v_2$ **then**
    $W = WCorrente$
  end
end
validarPesos(baseValidacao, W)

Consideram-se nódulos semelhantes aqueles em que os valores de cada um dos atributos

são muito próximos, ou até mesmo iguais. Ou seja, por exemplo, os nódulos que têm valores de *entropia* próximos de e_1 são semelhantes entre si; e os nódulos que tiverem valores próximos de e_2 são também semelhantes entre si. Assim, caso esta regra seja respeitada para as duas classes, o atributo é considerado como bom para representar os nódulos e ele receberá um valor de peso alto. Por outro lado, caso esta regra não seja satisfeita, o atributo é considerado ruim para representar tais nódulos e ele receberá um valor de peso baixo.

3.4.1 Fase de Avaliação

O ciclo se inicia com a Fase de Avaliação (Fase 1 da Figura 14) tendo como pesos iniciais um conjunto $W = [w_1, w_2, \ldots, w_a]$, onde $\{w \in \mathbb{R}_+^* | w = 1/a\}$ e $a \in \mathbb{N}^*$ correspondente ao número de atributos utilizados para representar os nódulos. Cada um dos nódulos da base é usado como referência para recuperar os n mais similares. A cada recuperação feita é calculado o valor de avaliação da recuperação através da Função de Avaliação (Equação 3.1), que é uma função de decaimento exponencial, onde:

$$f(R_{n\times a}) = \sum_{i=1}^{n} \gamma^i s_i \tag{3.1}$$

- $R_{n\times a}$ é a matriz ordenada com os nódulos recuperados na qual a corresponde ao número de atributos que representam os nódulos, e n corresponde ao número de nódulos similares que serão recuperados. A ordem da tabela é determinada pela similaridade dos nódulos, os mais similares situam-se nas posições iniciais;

- s_i é o valor da recompensa associado à relevância do nódulo n da matriz $R_{n\times a}$ na posição i;

- $\{\gamma \in \mathbb{R} \mid 0 < \gamma \leq 1\}$ é o fator de desconto, que ajusta a relevância das recompensas s dadas ao longo do *ranking* de recuperação.

Esta Função de Avaliação foi adotada por ter a capacidade de representar a amortização das recompensas ao longo da ordem de recuperação, que é uma característica importante para a nossa proposta, pois, devido à grande quantidade de exames recuperados, os usuários tendem a avaliar os resultados melhores colocados e esses direcionarão os especialistas no diagnóstico (MüLLER et al., 2004; FARIA et al., 2010).

O Algoritmo 2 descreve sucintamente o processo de avaliação dos pesos dos atributos na recuperação dos n nódulos mais semelhantes.

Algorithm 2: Pseudo-código da Função de Avaliação.

Data: baseAvaliacao:[ ][ ], W:[ ], $n : int$
Result: $avaliacao : real$
$avaliacao : real = 0.0$
$\gamma : real$ = fator de desconto
$m : int$ = posição do vetor onde fica a malignidade do nódulo
$s : real$ = recompensa
while *baseAvaliacao.hasNext()* **do**
    $noduloReferencia$ = baseAvalicao.next()
    nSimiliares:[ ][ ] = recupereNSimilares(baseAvaliacao, $noduloReferencia$, W, n)
    for $i \leftarrow 1$ ***to*** n **do**
        $malignidade$ = nSimilares[i][m]
        s = calcularRecompensa(noduloReferencia[m], $malignidade$)
        $avaliacao \mathrel{+}= \gamma^i \times s$
    end
end

As recompensas aplicadas aos nódulos dependem da malignidade do nódulo de referência e da malignidade do nódulo recuperado. Os valores atribuídos são os seguintes:

- 4, se for altamente relevante;
- 2, se for moderadamente relevante;
- 0, se for altamente ou moderadamente irrelevante.

Foi definida uma política de recompensas que privilegia os nódulos relevantes, não recompensando os que não satisfazem esta condição. A relevância é determinada da seguinte forma: se o nódulo de referência tiver malignidade 5 ou 4, as malignidades dos nódulos recuperados serão altamente relevantes se tiverem malignidade 5, moderadamente relevantes se tiverem malignidade 4, moderadamente irrelevantes se tiverem malignidade 2 e altamente irrelevantes se tiverem malignidade 1 (Tabela 4); se o nódulo de referência tiver malignidade 1 ou 2, as malignidades dos nódulos recuperados serão altamente relevantes se tiverem malignidade 1, moderadamente relevantes se tiverem malignidade 2, moderadamente irrelevantes se tiverem malignidade 4 e altamente irrelevantes se tiverem malignidade 5 (Tabela 5). Aos nódulos altamente e moderadamente irrelevantes são atribuídos o valor de recompensa 0 (zero), porque esses podem induzir o especialista ao erro. Logo, a eles não foi atribuída recompensa positiva. Os nódulos de malignidade 3 foram descartados, como citado na Seção 2.7, porque não contribuem para o auxílio ao diagnóstico, já que não possuem um grau de malignidade definido.

Tabela 4 – Recompensas para nódulos de referência com malignidade 5 ou 4.

Recompensas	Malignidade do nódulo recuperado
4	5
2	4
0	2
0	1

Fonte: Elaborada pelo autor.

Tabela 5 – Recompensas para nódulos de referência com malignidade 1 ou 2.

Recompensas	Malignidade do nódulo recuperado
4	1
2	2
0	4
0	5

Fonte: Elaborada pelo autor.

3.4.2 Fase de Treinamento

A Fase de Treinamento tem por objetivo encontrar o conjunto W de pesos associados aos atributos que permitam uma recuperação na qual os nódulos sejam o mais semelhantes possível. Ela é iniciada recebendo como pesos iniciais os pesos utilizados na Fase de Avaliação imediatamente anterior e, a partir daí, se inicia o processo de atualização de pesos. O ajuste automático dos pesos se dá conforme o método de atualização de pesos e leva em consideração os nódulos mais similares recuperados para ajustar os valores de pesos. Ao seu término, um novo conjunto de pesos ($WCorrente$) é determinado e será passado para a Fase de Avaliação para determinar o seu valor de avaliação.

Os pesos associados aos atributos refletem as diferentes contribuições dos descritores na caracterização do objeto. Não existe um mapeamento direto entre os critérios de classificação utilizados pelo usuário e a forma de representação dos objetos pela máquina. O que se busca aqui é uma adequação dos pesos de tal forma que seja possível alcançar melhores resultados na recuperação de nódulos semelhantes por meio da minimização da influência de atributos que possuem alto índice de dispersão e aumento da influência dos atributos que possuem baixo índice de dispersão em nódulos de mesma classe.

A proposta de atualização de pesos é baseada no DP para atualização dos pesos. O DP (Equação 3.2) é uma medida de dispersão estatística, ou seja, ele mede a dispersão dos dados de uma amostra com relação a sua média, onde $\overline{x}$ é a média da amostra (Equação 3.3).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{t}(x_i - \overline{x})^2}{t-1}} \tag{3.2}$$

$$\overline{x} = \frac{1}{t}\sum_{i=1}^{t} x_i \tag{3.3}$$

A proposta é baseada nas seguintes premissas: se todos os nódulos semelhantes têm valores similares para um determinado conjunto de atributos, isso significa que esses são bons indicadores para representar estes nódulos. Por outro lado, se os valores de um conjunto de atributos são muito diferentes, ou seja, muito dispersos, então eles não são bons indicadores. Logo, o IDP (Equação 3.4) dos dados associados a um atributo pode ser considerado uma boa estimativa para o seu peso, porque quanto menor a variância, maior é o peso e vice-versa.

$$w(a) = \sigma^{-1} \tag{3.4}$$

A Figura 15 ilustra a atualização dos pesos dos atributos com base nos n nódulos recuperados considerando:

- n o número de nódulos recuperados;
- f o número de atributos usados para representar cada nódulo;
- Π_f representando a projeção do atributo f na matriz de nódulos recuperados;
- $\sigma^{-1}(\Pi_f)$ representando a aplicação do IDP sobre a amostra resultante da projeção Π_f;
- w_f como sendo o peso do atributo a_f.

Após a identificação de cada w_f associado aos atributos a_f é preciso aplicar a Equação 3.5 para normalizar os valores, porque $w_f \in \mathbb{R}_+^*$ e ele pode assumir valores muito grandes quando a amostra varia pouco, ou muito pequenos quando a amostra varia muito. Com isso, temos um novo peso $w'f$ com valores entre zero (0) e um (1).

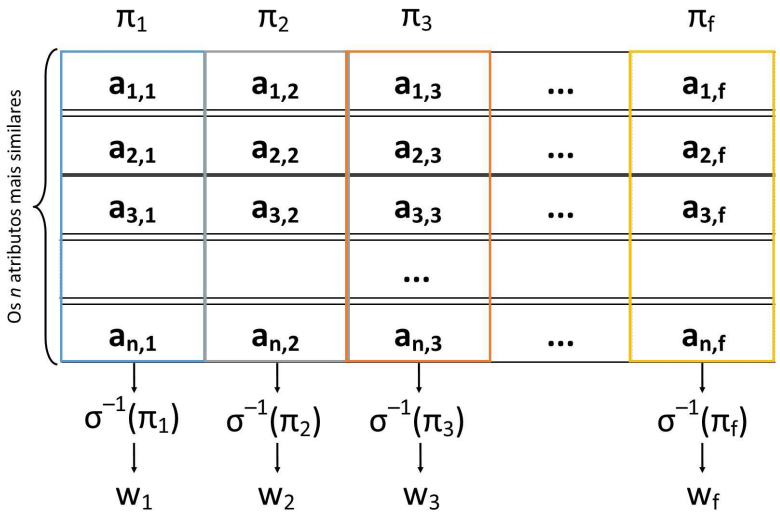
$$w'_f = \frac{w_f}{\sum_{i=1}^{f} w_i} \tag{3.5}$$

A cada iteração é encontrado um $WCorrente$, que é usado para ajustar o vetor de pesos W usando a Função de Ajuste (Equação 3.6), onde:

$$W^* = W + \alpha(W - WCorrente) \tag{3.6}$$

- W é o vetor com os melhores pesos até o momento;
- $WCorrente$ é o vetor de pesos da iteração corrente;
- α é a taxa de ajuste;

Figura 15 – Método de atualização de pesos.



Fonte: Elaborada pelo autor.

- W^* é o novo vetor de pesos ajustados.

Para melhor apreciação deste processo, veja o Algoritmo 3.

Algorithm 3: Pseudo-código da Função de Treinamento.

Data: baseTreinamento:[ ][ ], W:[ ], a:int
Result: WCorrente:[ ]
$WCorrente_i = \frac{w_i}{\sum w}$
while *baseTreinamento.hasNext()* **do**
    $noduloReferencia$ = baseTreinamento.next()
    nSimilares:[ ][ ] = recupereNSimilares(baseTreinamento, $noduloReferencia$, W, n)
    for $i \leftarrow 1$ ***to*** a **do**
        for $j \leftarrow 1$ ***to*** n **do**
            Π_i = nodulosSemelhantes[j][i]
            WCorrente[i] = $\sigma^{-1}(\Pi_i)$
        end
    end
end

A característica da aprendizagem tardia na proposta aqui apresentada está relacionada à taxa de ajuste dos pesos (α), a qual determina o quanto será aprendido em cada iteração na base de treinamento. Como pode ser visto na Equação 3.6, o vetor de pesos *WCorrente* determinado na iteração LOO sobre a base de treinamento é resultante do cálculo dos pesos naquela iteração, enquanto o vetor de pesos W é o vetor calculado durante todas as iterações anteriores. A aplicação de α na diferença entre W e *WCorrente* implica

no quanto que a diferença entre a memória de aprendizagem (W) e a aprendizagem da iteração atual ($WCorrente$) irá influenciar na memória de aprendizagem final (W^*).

Já a característica da LWL na proposta deste trabalho remete ao número de nódulos recuperados (n) durante a fase de treinamento que foram utilizados para calcular o peso dos atributos por meio do IDP. Devido a esta característica, apenas os nódulos mais próximos do nódulo de referência são utilizados para ajustar os pesos dos atributos com o propósito de enfatizar os nódulos de mesma classe e desenfatizar os nódulos com classes diferentes por meio da atribuição de pesos maiores àqueles atributos que proporcionam recuperação de nódulos com valores de atributos com menor índice de dispersão. Observe que a enfatização de nódulos semelhantes é feita indiretamente com base na premissa de que nódulos de mesma malignidade são semelhantes e, por isso, possuem vetores de atributos também semelhantes (com baixa variabilidade), ou seja, não são utilizados critérios baseados na classe para determinar os pesos dos objetos como em modelos KNN e WA.

LWL é criticamente dependente da função de distância utilizada e a função não necessariamente precisa satisfazer requisitos matemáticos formais para métricas de distância. A variação da DE com a introdução dos pesos nas dimensões dos objetos serve para que sejam atribuídos valores de influência das dimensões. Atribuir valor zero para uma dimensão é o mesmo que ignorá-la na função de distância. Devido a isso, é adotada a DEP. O valor máximo da função ponderada deve ser alcançado com a distância zero, e esse valor deve decair suavemente quando a distância aumentar. Além disso, funções de peso devem ser sempre não-negativas, porque valores negativos podem levar a um aumento da taxa de erro durante o treinamento. Já os pesos finais podem ser positivos ou negativos (ATKESON; MOORE; SCHAAL, 1997).

3.5 Avaliação dos resultados

Para a análise dos resultados foram utilizados os métodos PR e PN como citado na Seção 2.10. Eles foram calculados a partir da recuperação de todos os nódulos das bases de avaliação e validação por meio da técnica LOO.

Com o objetivo de mostrar as capacidades do algoritmo aqui proposto e comprovar a precisão dos resultados alcançados, foram definidas configurações de testes com base na variação dos principais parâmetros do algoritmo, que são a taxa de ajuste, o fator de desconto e o número de nódulos recuperados, aplicados à combinações dos vetores de atributos extraídos dos nódulos pulmonares. A partir destes testes, resultaram os gráficos que serviram de base para que fossem discutidos os resultados.

Foram utilizados 3 vetores de atributos calculados a partir dos nódulos da base: 1 com AT 3D (36 atributos); 1 com ANB 3D (12 atributos); e, por fim, 1 criado a partir da concatenação dos dois vetores citados anteriormente AI (48 atributos) (Seção 2.5). A escolha pela utilização destes 3 vetores tem como objetivo analisar a precisão de cada um deles para identificar qual alcança maior precisão.

Dada a grande quantidade de combinações envolvendo os parâmetros do algoritmo e os vetores de atributos, tendo em vista que $\{n \in \mathbb{N} \mid 0 < n \leq \text{quantidade de nódulos na base}\}$ e $\{\alpha, \gamma \in \mathbb{R} \mid 0 < \alpha, \gamma < 1\}$, foram definidos como valores referência para n, α e γ, respetivamente 15, 0.3 e 0.8, e mais dois conjuntos de valores sendo um maior e outro menor do que os valores de referência. Estes valores foram definidos para que reduzissem a quantidade de combinações possíveis dos vetores de atributos com os parâmetros do algoritmo, mas que possibilitassem as análises necessárias para medição da precisão. A determinação dos valores de referência se deu de forma empírica, porque eles proporcionaram melhores resultados, tanto de precisão, quanto de tempo de execução do algoritmo, durante os diversos testes executados no seu desenvolvimento. A Tabela 6 resume as configurações dos testes.

Tabela 6 – Resumo das configurações definidas para os testes, onde α corresponde à taxa de ajuste, γ ao fator de desconto e n ao número de nódulos recuperados.

AT			ANB			AI		
n	α	γ	n	α	γ	n	α	γ
10	0.2	0.7	10	0.2	0.7	10	0.2	0.7
15	**0.3**	**0.8**	**15**	**0.3**	**0.8**	**15**	**0.3**	**0.8**
20	0.4	0.9	20	0.4	0.9	20	0.4	0.9

Fonte: Elaborada pelo autor.

A Tabela 7 traz os identificadores e as descrições das configurações utilizada nos testes. Isso se faz necessário para facilitar o entendimento dos resultados.

Para avaliar a precisão com os métodos definidos (PR e PN), as malignidades associadas aos nódulos foram agrupadas para que eles sejam classificados em benignos ou malignos. Assim, nódulos com probabilidade de malignidade 1 e 2 foram agrupados como benignos, e nódulos com probabilidade de malignidade 4 e 5 foram agrupados como malignos. Vale ressaltar que os nódulos com malignidade 3 foram descartados, porque possuem classificação indefinida (Tabela 8).

A execução dos testes definidos na Tabela 7 permitiu avaliar o aumento na precisão a partir da utilização dos pesos calculados automaticamente segundo a metodologia proposta neste trabalho.

Tabela 7 – Identificadores e configurações utilizadas nos testes relacionados.

ID do teste	Descrição da configuração
1	$AT; n = 10; \alpha = 0.3; \gamma = 0.8$
2	$AT; n = 15; \alpha = 0.3; \gamma = 0.8$
3	$AT; n = 20; \alpha = 0.3; \gamma = 0.8$
4	$AT; n = 15; \alpha = 0.2; \gamma = 0.8$
5	$AT; n = 15; \alpha = 0.4; \gamma = 0.8$
6	$AT; n = 15; \alpha = 0.3; \gamma = 0.7$
7	$AT; n = 15; \alpha = 0.3; \gamma = 0.9$
8	$ANB; n = 10; \alpha = 0.3; \gamma = 0.8$
9	$ANB; n = 15; \alpha = 0.3; \gamma = 0.8$
10	$ANB; n = 20; \alpha = 0.3; \gamma = 0.8$
11	$ANB; n = 15; \alpha = 0.2; \gamma = 0.8$
12	$ANB; n = 15; \alpha = 0.4; \gamma = 0.8$
13	$ANB; n = 15; \alpha = 0.3; \gamma = 0.7$
14	$ANB; n = 15; \alpha = 0.3; \gamma = 0.9$
15	$AI; n = 10; \alpha = 0.3; \gamma = 0.8$
16	$AI; n = 15; \alpha = 0.3; \gamma = 0.8$
17	$AI; n = 20; \alpha = 0.3; \gamma = 0.8$
18	$AI; n = 15; \alpha = 0.2; \gamma = 0.8$
19	$AI; n = 15; \alpha = 0.4; \gamma = 0.8$
20	$AI; n = 15; \alpha = 0.3; \gamma = 0.7$
21	$AI; n = 15; \alpha = 0.3; \gamma = 0.9$

Fonte: Elaborada pelo autor.

Tabela 8 – Mapeamento das malignidades em classes.

Probablidade de malignidade	Classe
1 e 2	Benigno
2 e 4	Maligno

Fonte: Elaborada pelo autor.

4 RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os resultados obtidos com a avaliação do método de atualização de pesos seguindo as configurações de testes estabelecidas na Seção 3.5. Primeiramente, serão apresentados os resultados da avaliação PR e PN de cada vetor de atributos. Em seguida, serão feitas as comparações entre os vetores de atributos AT 3D, ANB 3D e AI para apontar aquele que obteve os melhores resultados com a atualização de pesos. Por fim, serão feitas comparações com outros resultados apresentados na literatura.

A Figura 16 apresenta os resultados do PR a partir do vetor AT 3D com os pesos iniciais, como apresentado na Seção 3.4, utilizando apenas a base de avaliação. Com ele foi alcançada precisão de 87% com revocação de 25%, precisão de 64% com revocação de 50% e precisão de 50% com revocação de 75% na recuperação de nódulos benignos (Figura 16 a). Já na recuperação de nódulos malignos foi obtida 91% de precisão com revocação até 25%, precisão de 86% com revocação de 50% e precisão de 74% com revocação de 75% (Figura 16 b). O PN alcançou precisão de 87% para nódulos benignos (Figura 16 c) e de 93% para nódulos malignos (Figura 16 d) para recuperação até o trigésimo nódulo.

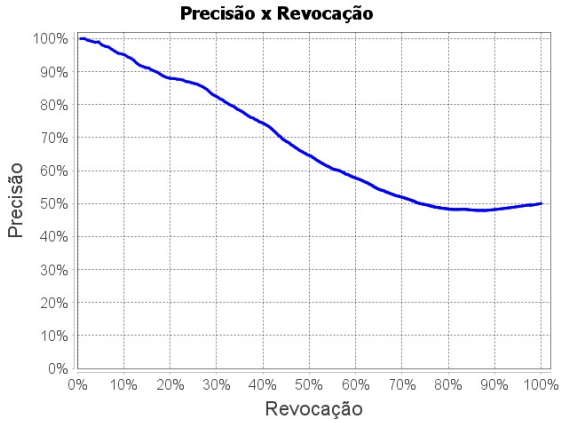
A execução do processo de atualização de pesos seguindo as configurações de teste definidas (testes de 1 a 7) proporcionou melhora na precisão nos casos 1, 2, 4, 5, 6 e 7 usando o vetor AT 3D em comparação com os pesos iniciais. Todos eles alcançaram o mesmo valor de precisão, apresentado nas Figuras 17 e 18. O único teste que não melhorou, comparado com os valores obtidos pelos pesos iniciais, após o treinamento foi o teste 3. É possível que neste teste 3, dada a maior quantidade de nódulos recuperados, o processo tenha chegado a uma situação de máximo local, convergindo para um valor menor que os demais testes.

Com o ajuste, obteve-se precisão de 98% com revocação de 25%, precisão de 97% com revocação de 50% e precisão de 97% com revocação de 75% na recuperação de nódulos benignos (Figura 17 a), já na recuperação de nódulos malignos obteve-se precisão de 99% com revocação de 25%, precisão de 99% com revocação de 50% e precisão de 96% com revocação de 75% na base de avaliação (Figura 17 b). Enquanto na base de validação, a precisão foi de 98% com revocação de 25%, precisão de 98% com revocação de 50% e precisão de 97% com revocação de 75% na recuperação de nódulos benignos (Figura 17 c), já na recuperação de nódulos malignos, obteve-se precisão de 100% com revocação de 25%, precisão de 100% com revocação de 50% e precisão de 97% com revocação 75% (Figura 17 d). O PN também melhorou, obtendo precisão de 98% na recuperação de nódulos benignos e malingos usando as duas bases (avaliação e validação) (Figura 18) para recuperação de 30 nódulos em cada base.

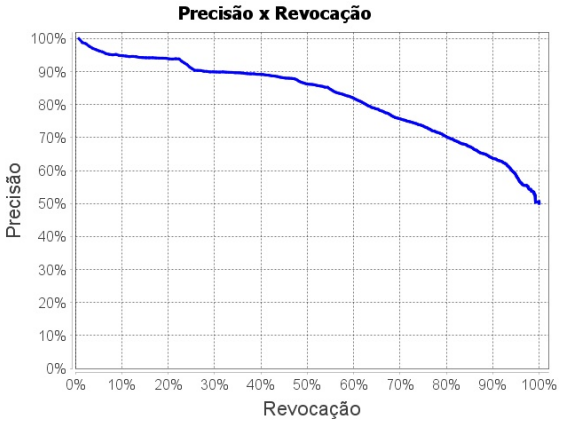
As Tabelas 9 e 10 apresentam de forma resumida os resultados da avaliação na base

Figura 16 – PR e PN da recuperação de nódulos benignos e malignos contidos nas bases de avaliação calculados com os pesos iniciais (Seção 3.4) representados por meio do vetor de Atributos de Textura 3D.

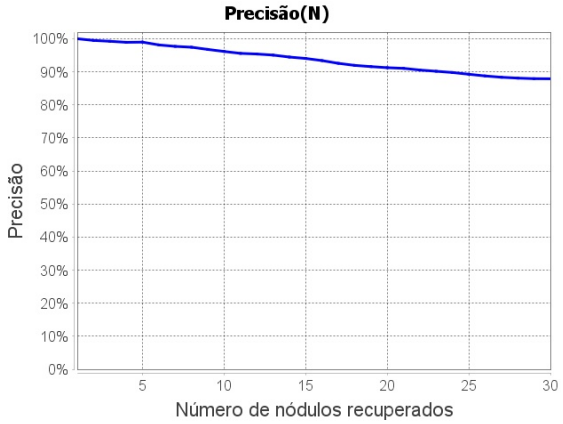
(a) PR calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



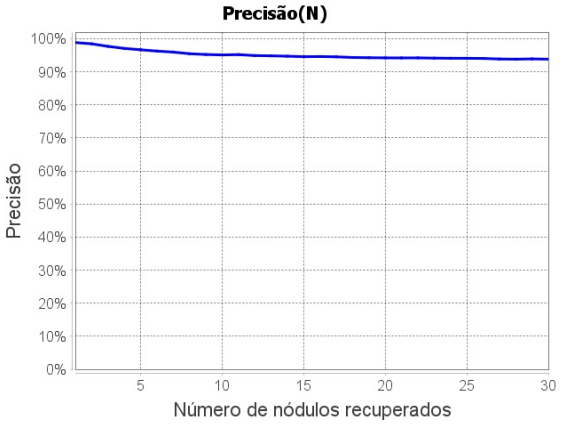
(b) PR calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PN calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



(d) PN calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



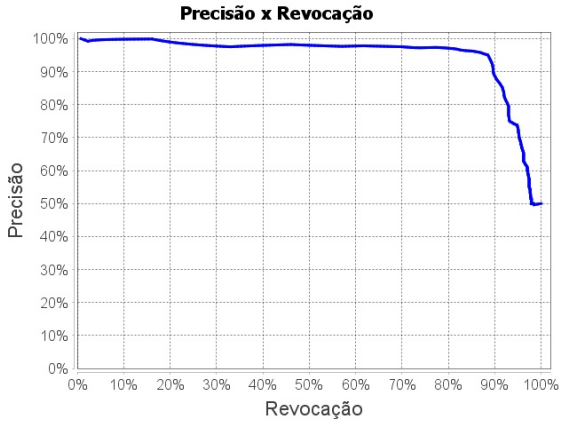
Fonte: Elaborada pelo autor.

de validação usando Precisão × Revocação e Precisão(n) do vetor AT 3D sem e com o ajuste de pesos respectivamente.

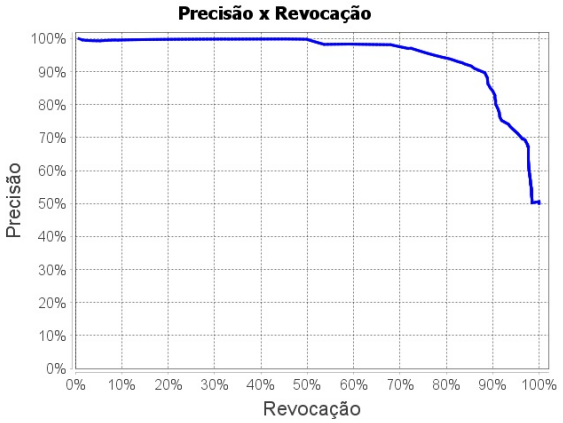
A Figura 19 apresenta os resultados do PR a partir do vetor ANB 3D com os pesos iniciais utilizando apenas a base de avaliação. Com ele foi alcançada precisão de 96% com revocação de 25%, precisão de 86% com revocação de 50% e precisão de 64% com revocação de 75% na recuperação de nódulos benignos (Figura 19 a). Enquanto com nódulos malignos, proporcionou uma precisão de 96% com revocação de 25%, precisão de 83% com revocação de 50% e precisão de 60% com revocação de 75% (Figura 19 b). O PN alcançou 96% de precisão na recuperação de nódulos benignos (Figura 19 c) e 95% de precisão com nódulos malingos (Figura 19 d) para recuperação de 30 nódulos de cada malignidade.

Figura 17 – PR da recuperação de nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados com o conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Textura 3D.

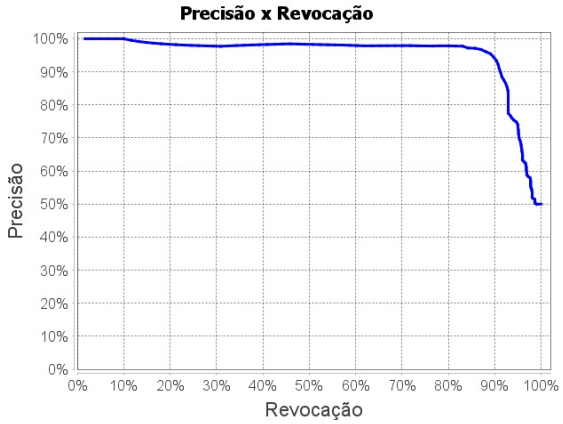
(a) PR calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



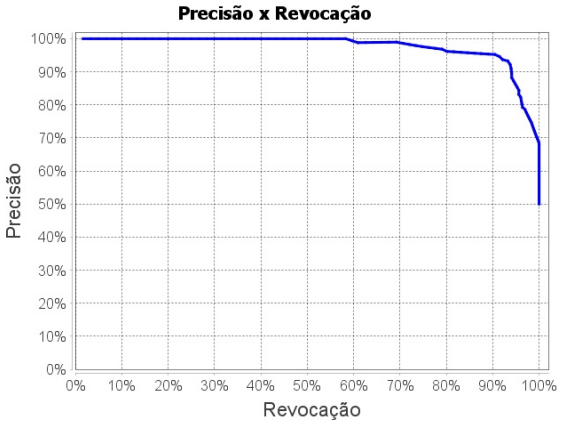
(b) PR calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PR calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PR calculado a partir da recuperação de todos os nódulos malignos da base de validação.



Fonte: Elaborada pelo autor.

Tabela 9 – Resumo dos resultados de Precisão × Revocação e Precisão(n) obtidos com o vetor AT 3D sem ajuste de pesos na base de validação

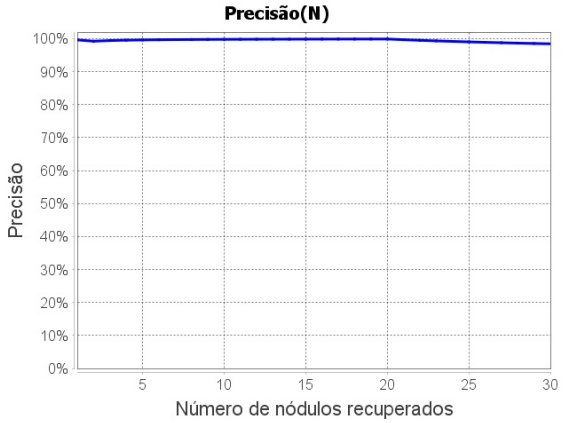
Precisão para benignos	87%	64%	50%
Precisão para malignos	91%	86%	74%
Revocação	25%	50%	75%

	Precisão($n = 30$)
Benigno	87%
Maligno	93%

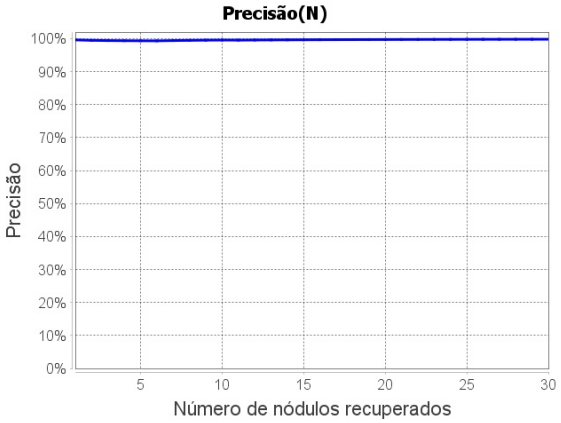
Fonte: Elaborada pelo autor.

Figura 18 – PN da recuperação de 30 nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados através do conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Textura 3D.

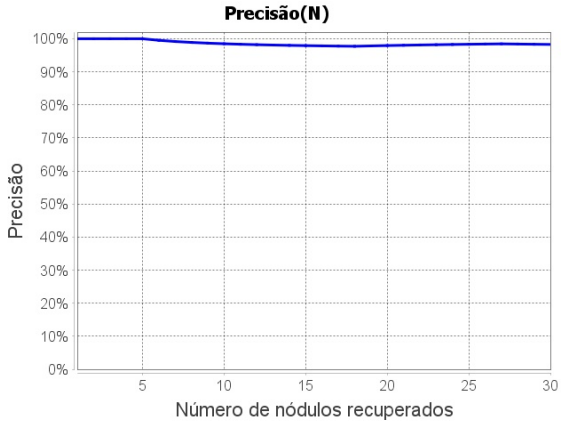
(a) PN calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



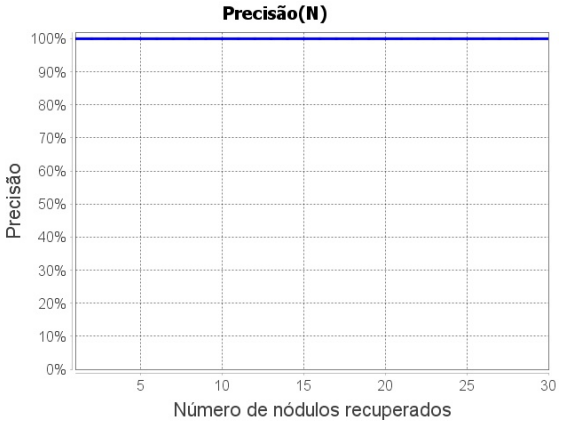
(b) PN calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PN calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PN calculado a partir da recuperação de todos os nódulos malignos da base de validação.



Fonte: Elaborada pelo autor.

Tabela 10 – Resumo dos resultados de Precisão × Revocação e Precisão(n) obtidos com o vetor AT 3D com ajuste de pesos na base de validação

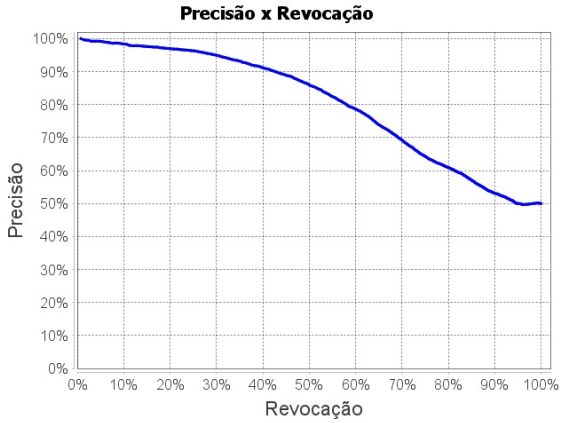
Precisão para benignos	98%	97%	97%
Precisão para malignos	99%	99%	96%
Revocação	25%	50%	75%

	Precisão($n = 30$)
Benigno	98%
Maligno	98%

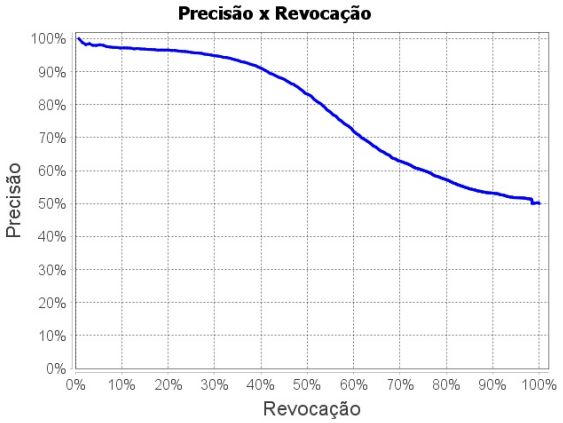
Fonte: Elaborada pelo autor.

Figura 19 – PR e PN da recuperação de nódulos benignos e malignos contidos nas bases de avaliação calculados com os pesos iniciais (Seção 3.4) representados por meio do vetor de Atributos de Nitidez de Borda 3D.

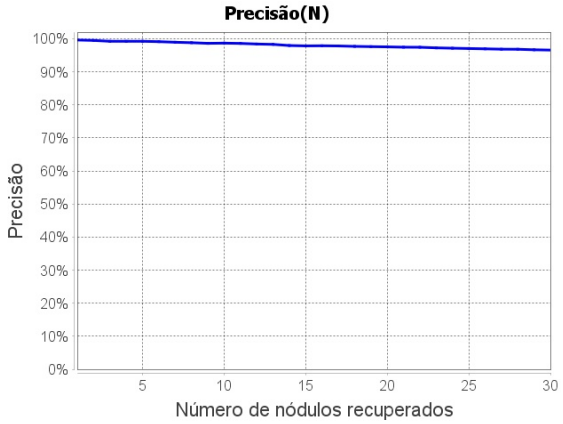
(a) PR calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



(b) PR calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PN calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



(d) PN calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.

Fonte: Elaborada pelo autor.

Após o processo de atualização de pesos definido no grupo de testes (testes de 8 a 14), o conjunto de pesos associado ao vetor ANB 3D apresentou melhora na precisão em todos os casos, mas alcançando valores diferentes.

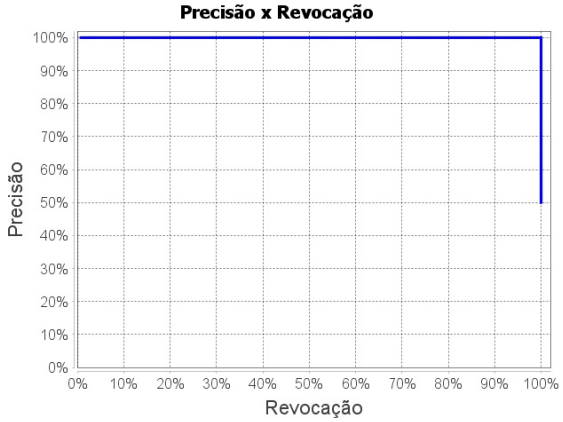
O teste 8 alcançou 100% de precisão com revocação até 98%, tanto para recuperação de nódulos benignos quanto para nódulos malingos, nas bases de avaliação e validação (Figuras 20), e o PN manteve-se em 100% para recuperação de 30 nódulos, benignos e malignos, nas bases de avaliação e validação (Figura 21).

Os demais testes do grupo (do 9 ao 14), nas bases de validação e avaliação, mantiveram a precisão de 100% com revocação até 93% para recuperação de nódulos benignos (Figuras 22 a e c), e precisão de 100% com a revocação até 98% para nódulos malignos (Figuras 22 b e d). O PN para recuperação de 30 nódulos alcançou 100% de precisão na

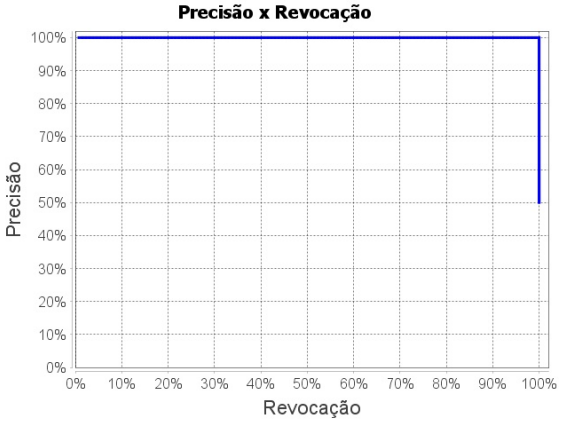
recuperação de nódulos benignos e malingos nas bases de avaliação e validação (Figura 23).

Figura 20 – PR da recuperação de nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados com o conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Nitidez de Borda 3D com a configuração definida no teste 8.

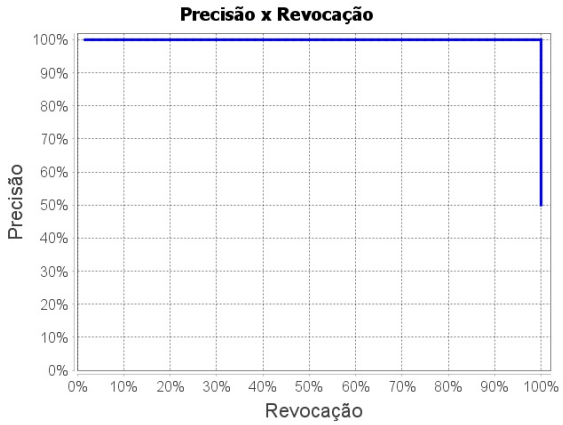
(a) PR calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



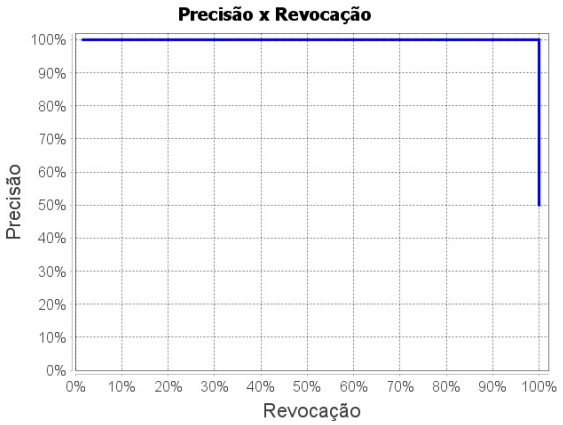
(b) PR calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PR calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PR calculado a partir da recuperação de todos os nódulos malignos da base de validação.



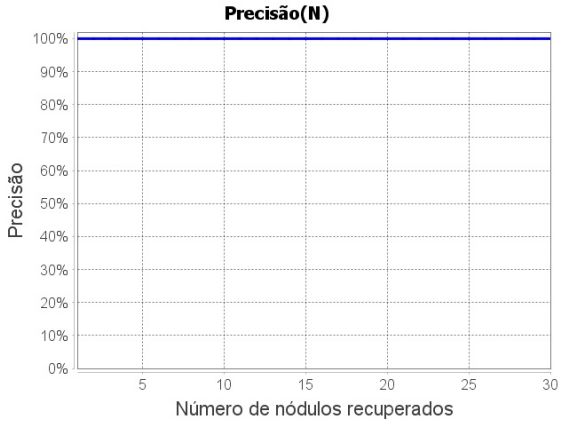
Fonte: Elaborada pelo autor.

As Tabelas 11 e 12 apresentam de forma resumida os resultados da avaliação na base de validação usando Precisão × Revocação e Precisão(n) do vetor ANB 3D sem e com o ajuste de pesos respectivamente.

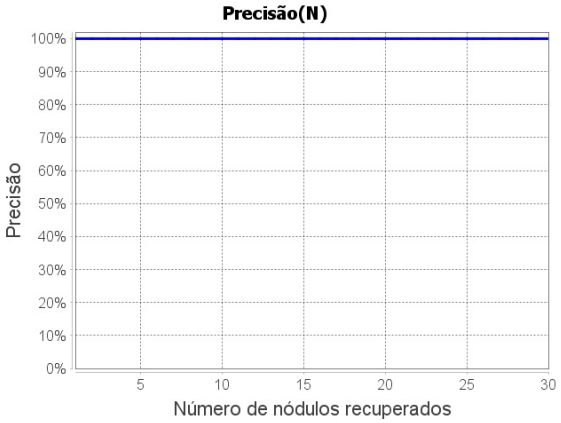
Por fim, a Figura 24 apresenta o resultado do PR a partir do vetor AI com os pesos iniciais utilizando apenas a base de avaliação. Com ele foi alcançada precisão de 82% com revocação de 25%, precisão de 66% com revocação de 50% e precisão de 52% com revocação de 75% na recuperação de nódulos benignos (Figura 24 a), enquanto na recuperação de nódulos malignos obteve precisão de 86% com revocação de 25%, precisão de 74% com revocação de 50% e precisão de 62% com revocação de 75% (Figura 24 b). O PN para

Figura 21 – PN da recuperação de 30 nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados com o conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Nitidez de Borda 3D com a configuração definida no teste 8.

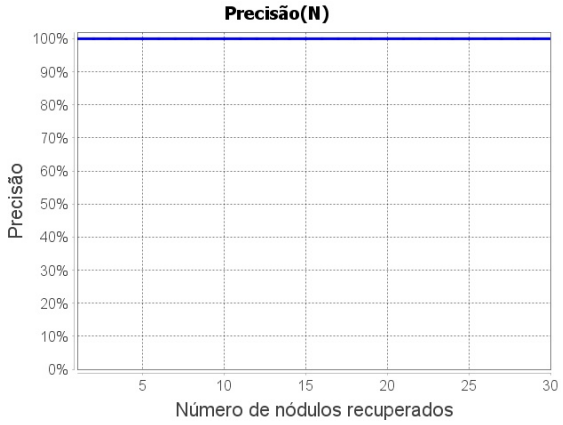
(a) PN calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



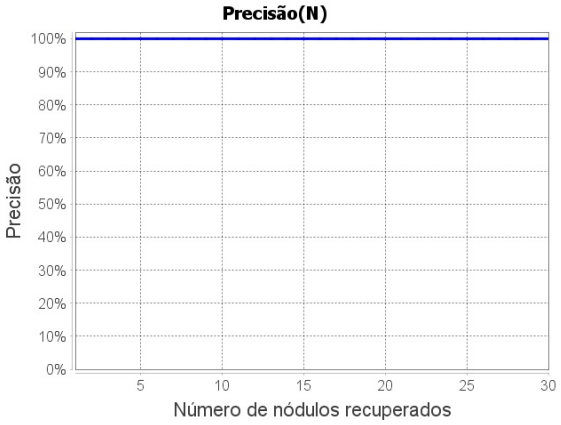
(b) PN calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PN calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PN calculado a partir da recuperação de todos os nódulos malignos da base de validação.



Fonte: Elaborada pelo autor.

Tabela 11 – Resumo dos resultados de Precisão × Revocação e Precisão(n) obtidos com o vetor ANB 3D sem ajuste de pesos na base de validação

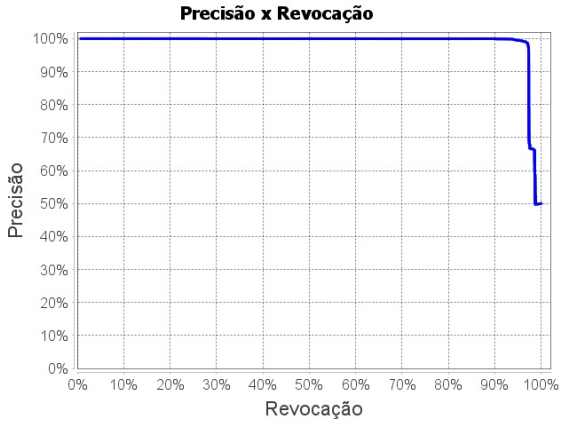
Precisão para benignos	96%	86%	64%
Precisão para malignos	96%	83%	60%
Revocação	25%	50%	75%

	Precisão($n = 30$)
Benigno	96%
Maligno	95%

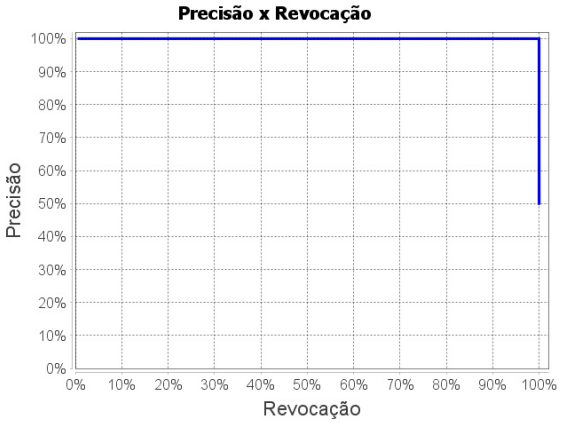
Fonte: Elaborada pelo autor.

Figura 22 – PR da recuperação de nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados com o conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Nitidez de Borda 3D com a configuração definida nos testes 9 ao 14.

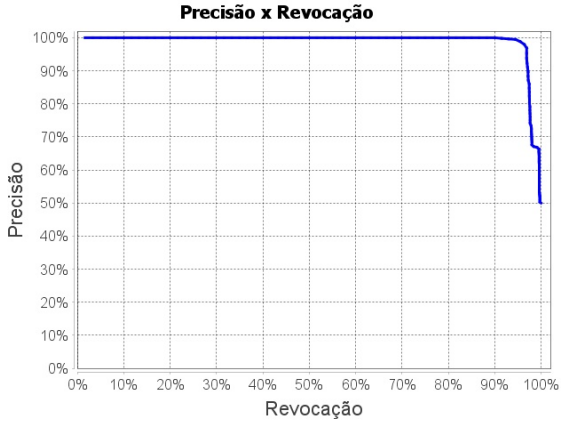
(a) PR calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



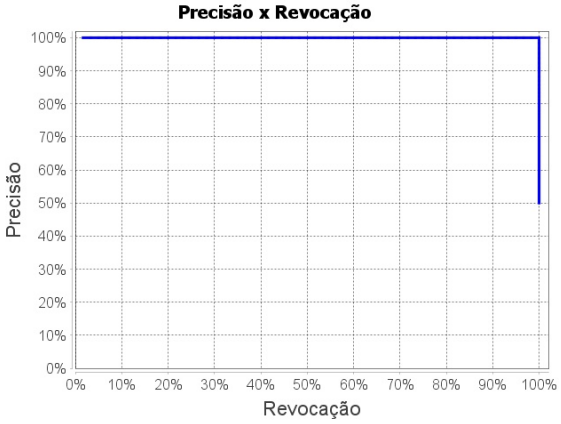
(b) PR calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PR calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PR calculado a partir da recuperação de todos os nódulos malignos da base de validação.



Fonte: Elaborada pelo autor.

Tabela 12 – Resumo dos resultados de Precisão × Revocação e Precisão(n) obtidos com o vetor ANB 3D sem ajuste de pesos na base de validação

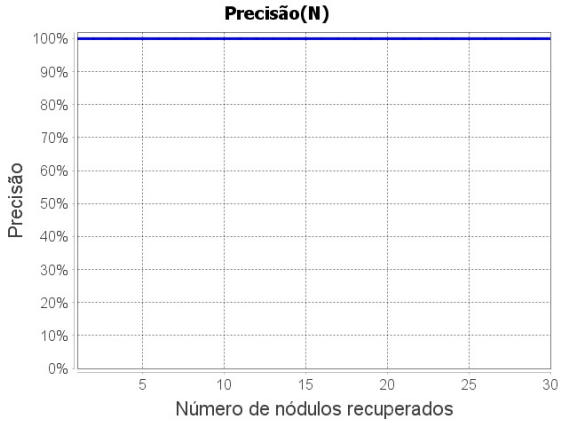
Precisão para benignos	100%	100%	100%
Precisão para malignos	100%	100%	100%
Revocação	25%	50%	75%

	Precisão($n = 30$)
Benigno	100%
Maligno	100%

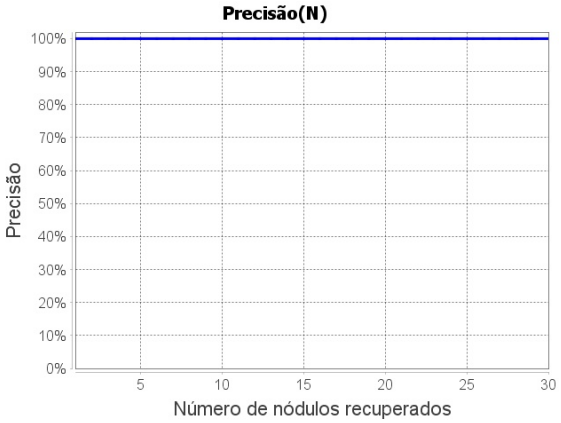
Fonte: Elaborada pelo autor.

Figura 23 – PN da recuperação de 30 nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados com o conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Nitidez de Borda 3D com a configuração definida nos testes 9 ao 14.

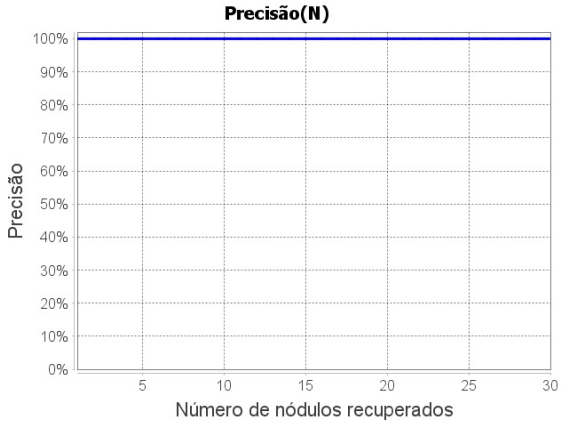
(a) PN calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



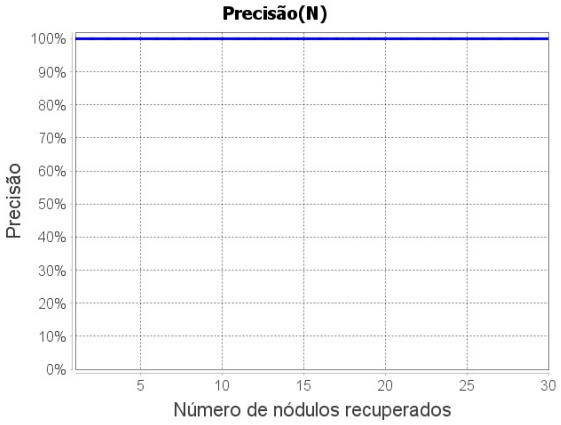
(b) PN calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PN calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PN calculado a partir da recuperação de todos os nódulos malignos da base de validação.



Fonte: Elaborada pelo autor.

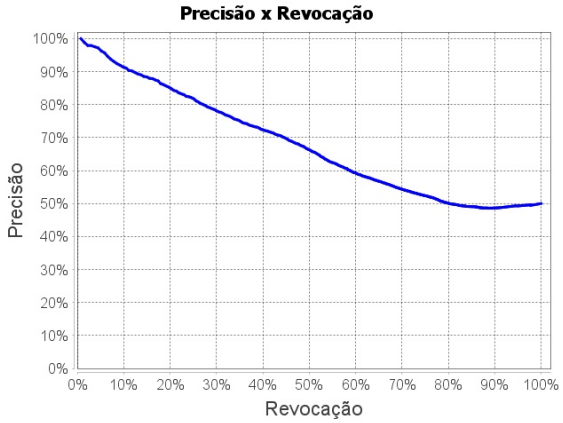
recuperação de 30 nódulos alcançou precisão de 85% para nódulos benignos (Figura 24 c) e de 87% para nódulos malignos (Figura 24 d).

Após o processo de atualização de pesos definido no último grupo de testes (testes de 15 a 21), cujo vetor de atributos é o AI, houve melhora na precisão apenas no teste 15, os demais não obtiveram resultado melhor do que o alcançado com os pesos iniciais.

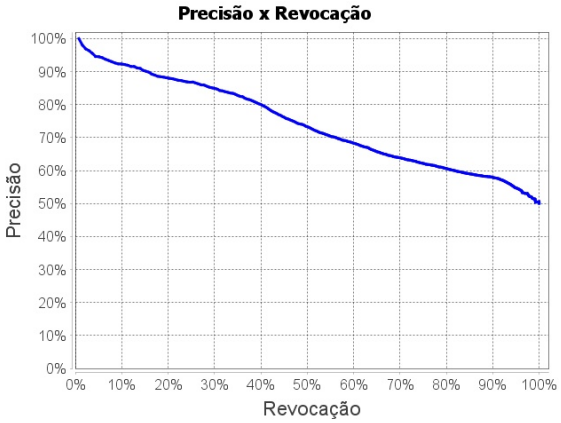
No teste 15 foi alcançada precisão de 95% com revocação de 25%, precisão de 90% com revocação de 50% e precisão de 70% com revocação de 75% para recuperação de nódulos benignos (Figuras 25 a); e precisão de 91% com revocação de 25%, precisão de 86% com revocação de 50% e precisão de 67% com revocação de 75% para recuperação de nódulos malingos (Figuras 25 b) na base de avaliação. Já na base de validação, o conjunto de

Figura 24 – PR e PN da recuperação de nódulos benignos e malignos contidos nas bases de avaliação calculados com os pesos iniciais (Seção 3.4) representados por meio do vetor resultante da junção dos Atributos de Textura 3D e dos Atributos de Nitidez de Borda 3D.

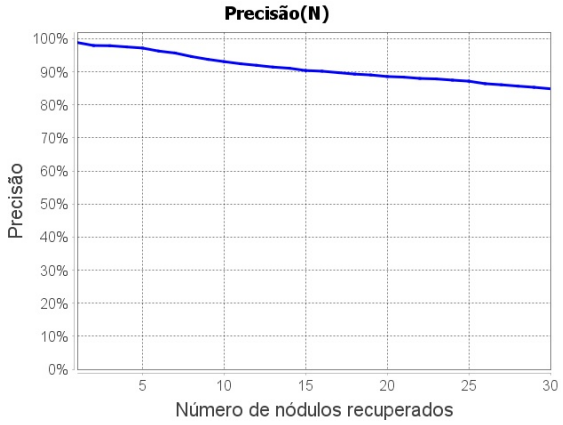
(a) PR calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



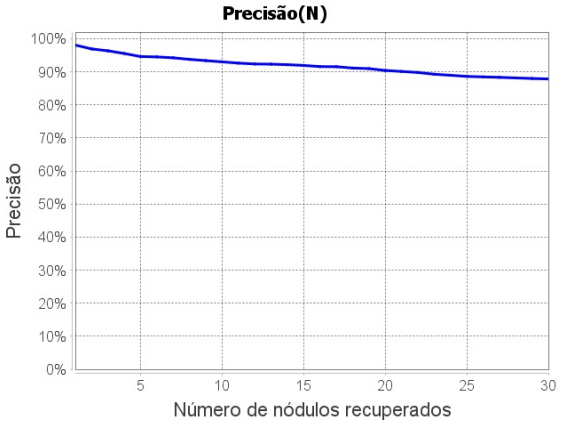
(b) PR calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PN calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



(d) PN calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



Fonte: Elaborada pelo autor.

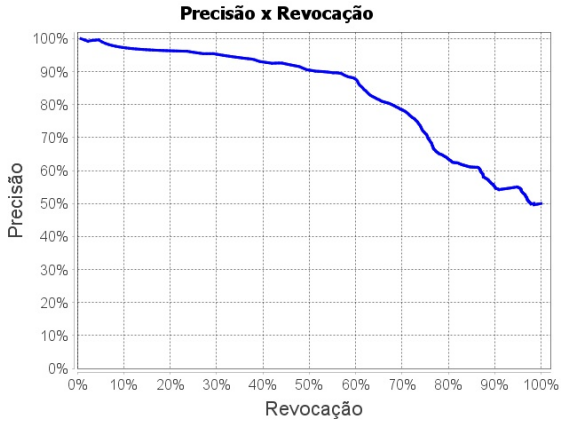
pesos resultantes deste teste obteve precisão de 96% com revocação de 25%, precisão de 91% com revocação de 50% e precisão de 84% com revocação de 75% na recuperação de nódulos benignos (Figuras 25 c); e precisão de 100% com revocação de 25%, precisão de 92% com revocação de 50% e precisão de 73% com revocação de 75% na recuperação de nódulos malignos (Figuras 25 d). No que se refere aos PN, os resultados apontam 95% de precisão para a recuperação de 30 nódulos para ambas as malignidades (Figuras 26).

As Tabelas 13 e 14 apresentam de forma resumida os resultados da avaliação na base de validação usando Precisão $\times$ Revocação e Precisão(n) do vetor AI sem e com o ajuste de pesos respectivamente.

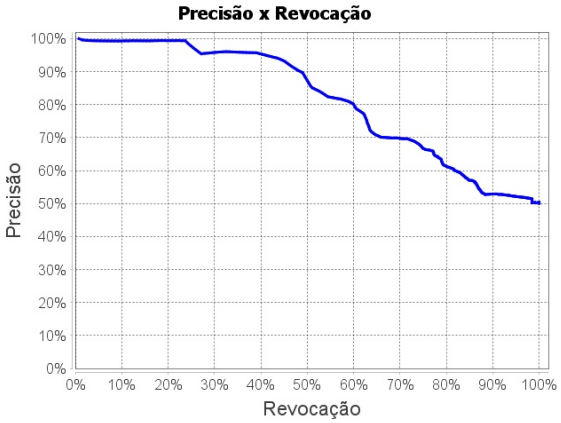
A partir da análise dos resultados encontrados, o melhor vetor para recuperação de nó-

Figura 25 – PR da recuperação de nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados com o conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Textura 3D e Atributos de Nitidez de Borda 3D conjuntamente com a configuração definida no teste 15.

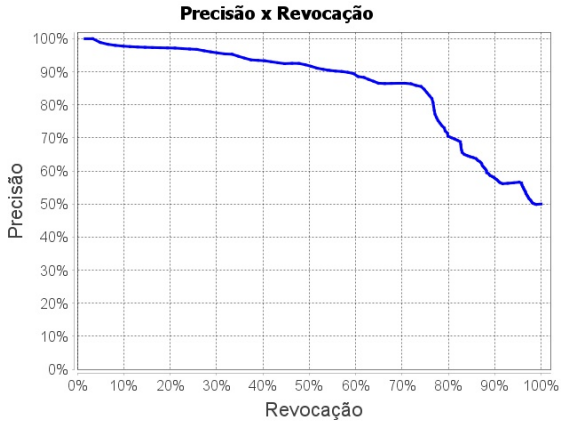
(a) PR calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



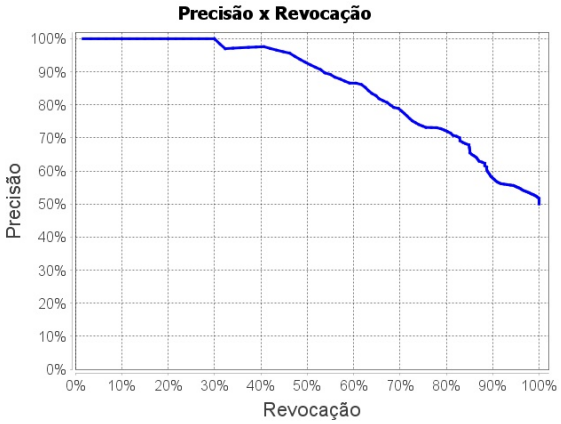
(b) PR calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PR calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PR calculado a partir da recuperação de todos os nódulos malignos da base de validação.



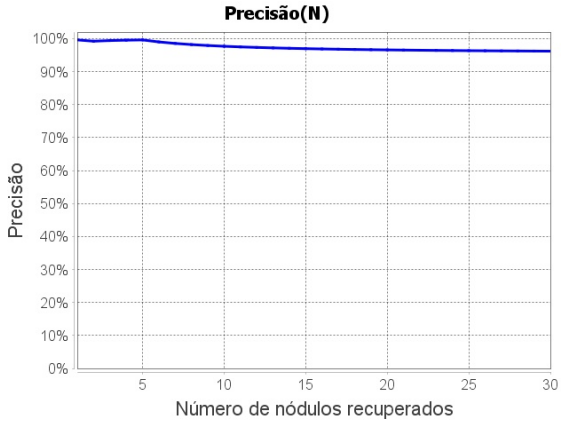
Fonte: Elaborada pelo autor.

dulos é aquele formado pelo ANB 3D com os pesos resultantes do processo de treinamento cujos valores de n, α e γ são 10, 0.3 e 0.8 respectivamente. Ele proporcionou valores que chegam a 100% de precisão na recuperação de nódulos pulmonares benignos e malignos com revocação acima de 98% tanto para a base de avaliação (Tabela 15), quanto para a base de validação (Tabela 16). A comparação entre os resultados alcançados utilizando os pesos iniciais e os pesos ajustados demonstram a eficácia do algoritmo e a sua capacidade para produzir bons resultados.

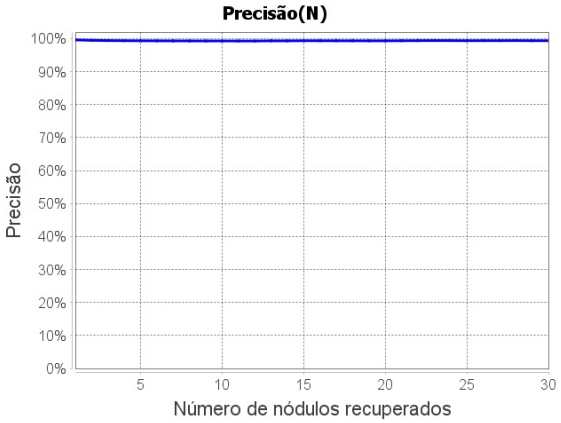
Embora tenha sido percebida uma influência dos valores dos parâmetros utilizados no processo de atualização, principalmente no que se refere ao número de iterações necessárias

Figura 26 – PN da recuperação de 30 nódulos benignos e malignos contidos nas bases de avaliação e validação recuperados com o conjunto de pesos que melhor ajustou a DEP para recuperação utilizando Atributos de Textura 3D e Atributos de Nitidez de Borda 3D conjuntamente com a configuração definida no teste 15.

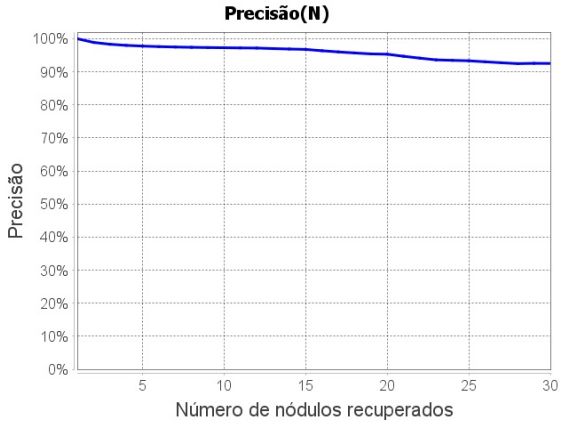
(a) PN calculado a partir da recuperação de todos os nódulos benignos da base de avaliação.



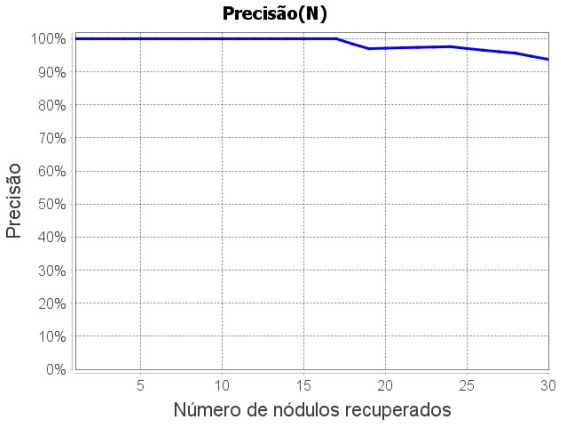
(b) PN calculado a partir da recuperação de todos os nódulos malignos da base de avaliação.



(c) PN calculado a partir da recuperação de todos os nódulos benignos da base de validação.



(d) PN calculado a partir da recuperação de todos os nódulos malignos da base de validação.



Fonte: Elaborada pelo autor.

para se chegar ao melhor resultado, ainda não é possível mensurar esta influência em um valor escalar, ou seja, ela é percebida empiricamente, mas não é possível descrevê-la quantitativamente.

Os resultados alcançados pelo ajuste dos pesos no vetor ANB 3D podem ser comparados com os resultados apresentados por Xu et al. (2012). A avaliação da recuperação utilizou o método NDCG alcançando um *score* de 85% na recuperação de nódulos pulmonares. Embora a métrica seja outra, nos resultados aqui apresentados foi possível alcançar até 100% de precisão com 98% de revocação nas bases de nódulos pulmonares.

Em comparação aos resultados apresentados por Ferreira Junior (2015), os ANB 3D

Tabela 13 – Resumo dos resultados de Precisão × Revocação e Precisão(n) obtidos com o vetor AI sem ajuste de pesos na base de validação.

Precisão para benignos	82%	66%	52%
Precisão para malignos	86%	74%	62%
Revocação	25%	50%	75%

	Precisão($n = 30$)
Benigno	85%
Maligno	87%

Fonte: Elaborada pelo autor.

Tabela 14 – Resumo dos resultados de Precisão × Revocação e Precisão(n) obtidos com o vetor AI com ajuste de pesos na base de validação.

Precisão para benignos	95%	90%	70%
Precisão para malignos	91%	86%	67%
Revocação	25%	50%	75%

	Precisão($n = 30$)
Benigno	95%
Maligno	95%

Fonte: Elaborada pelo autor.

Tabela 15 – Resultados obtidos através da execução do caso de teste 8 utilizando a base de avaliação. Aqui é apresentado o percentual da Precisão associado ao percentual de Revocação correspondente.

Precisão	100%	100%	100%	50%
Revocação	25%	50%	75%	100%

Fonte: Elaborada pelo autor.

isolados apresentaram maior precisão média para os 10 primeiros casos malignos recuperados em comparação com os AT 3D e o AI. A precisão média dos ANB 3D foi de aproximadamente 81%, enquanto para os AT 3D e AI a precisão média foi de aproximadamente 78% para recuperação de nódulos malignos. Já quando foram recuperados nódulos benignos, os resultados mostraram que os ANB 3D apresentaram precisão média para os 10 primeiros casos recuperados de 84%, enquanto os AT 3D e AI apresentaram uma precisão média de 82%. Nós resultados aqui apresentados, o ajuste de pesos proporcionou melhores resultados do PN para 30 nódulos em todos os casos.

O trabalho apresentado por Dhara et al. (2012), que utilizou regressão linear para reduzir a dimensionalidade do vetor formado por atributos de textura e forma, apresentou resultados que foram medidos através do cálculo da Precisão na recuperação de 5 nódulos similares. A precisão média alcançada por eles na recuperação dos 40 nódulos do LIDC foi de 72,18% e na base de nódulos PGIMER foi de 78,29%.

Tabela 16 – Resultados obtidos através da execução do caso de teste 8 utilizando a base de validação. Aqui é apresentado o percentual da Precisão associado ao percentual de Revocação correspondente.

Precisão	100%	100%	100%	50%
Revocação	25%	50%	75%	100%

Fonte: Elaborada pelo autor.

Seitz Jr et al. (2012) determinaram um vetor ótimo para representação a partir de um conjunto de 63 atributos extraídos a partir da textura, forma, tamnho e intensidade. Para tal, eles utilizaram Algoritmos genéticos para encontrar a melhor combinação de atributos. Os resultados foram avaliados através do cálculo da Precisão média para recuperação de 3, 5, 10, 20 e 50 imagens. O melhor resultado proporcinou precisão média de 86,91% para recuperação de 3 nódulos através do vetor ótimo formado por 29 atributos, dentre os 63 iniciais.

Já Kuruvilla e Gunavathi (2014) buscaram o vetor ótimo para representação dos nódulos em sistemas CBIR pelo cálculo da acurácia no algoritmo de Rede Neural. Além disso, eles avaliaram diferentes métricas de similaridade para identificar aquela que proporciona maior precisão na recuperação de nódulos similares. Os resultados foram avaliados através do cálculo da Precisão. O melhor resultado alcançou 95% de precisão média utilizando os parâmetros apontados pela Rede Neural e tendo a função *Bray-Curtis* como métrica de similaridade.

Nos resultados preliminares deste trabalho (LUCENA; Ferreira Junior; OLIVEIRA, 2014), nós obtivemos uma precisão média de aproximadamente 58% na recuperação de 10 nódulos de malignidade 1, e alcançou-se uma precisão média de aproximadamente 65% na recuperação de 10 nódulos de malignidade 5 em 10 recuperações feitas em cada avaliação. A precisão foi calculada pela Equação 4.1, onde P é um vetor de tamanho i com as precisões da recuperação. Cada posição contém a precisão na ordem n. VP é o número de verdadeiros-positivos obtidos até a ordem n. T é o número total de nódulos recuperados até a ordem n. n é o número da ordem na recuperação. E i é o número de nódulos recuperados. Nele, foi utilizado um sistema CBIR com a base de dados do LIDC (contendo 1.171 nódulos de 4 malignidades diferentes: 1, 2, 4 e 5), tendo como métrica de similaridade a DE, a representação dos nódulos pulmonares feitas através de vetores de AT 3D calculados da mesma forma que a apresentada neste trabalho e a recuperação dos nódulos com o objetivo de se encotrar nódulos de mesma malignidade.

$$P_i = \sum_{n=1}^{i} \frac{VP_n}{T_n} \tag{4.1}$$

Em trabalhos mais recentes (LUCENA et al., 2015a; LUCENA et al., 2015b), foi utilizado o mesmo sistema CBIR e base de imagens do trabalho (LUCENA; Ferreira

Junior; OLIVEIRA, 2014), divergindo pela utilização dos pesos na DEP, proposta aqui defendida. Com os pesos, a precisão foi aumentada, na média, em 17,3% em comparação com a recuperação usando a DE. Esta melhora foi observada levando-se em consideração os nódulos classificados com as 4 malignidades (1, 2, 4 e 5) agrupados em duas classes: 1 e 2 classificados como benignos; 4 e 5 classificados como malignos.

Como pôde ser visto, os trabalhos de Xu et al. (2012), Han et al. (2015) e Lucena, Ferreira Junior e Oliveira (2014) foram avaliados com métodos diferentes sendo o NDCG, AUC e Equação 4.1 respectivamente. Estes foram apresentados aqui porque, segundo nossas pesquisas, foram encontrados poucos trabalhos que tenham utilizado os mesmos atributos, base de nódulos e arquitetura de sistema CBIR para que fossem comparados com o nosso e isso nos levou a apresentar trabalhos da literatura com outros métodos de avaliação, que não fossem a Precisão $\times$ Revocação.

5 CONCLUSÃO

Este trabalhou apresentou um algoritmo de atualização automática local de pesos para a DEP com o objetivo de aumentar a precisão na recuperação de nódulos pulmonares em sistemas CBIR cuja métrica de similaridade está definida no espaço vetorial multidimensional.

Como base para o desenvolvimento de tal proposta, foi utilizada uma arquitetura genérica de sistema CBIR. Para representação dos nódulos foram utilizados vetores de AT 3D e ANB 3D. A métrica de similaridade escolhida foi a DEP e para medir os pesos foi usado IDP. Com isso, buscou-se alcançar melhores resultados do que os que já foram alcançados utilizando uma arquitetura similar, mas sem o conceito dos pesos associados aos atributos aqui apresentado.

Os resultados alcançados comprovaram a eficácia do método proposto aumentando a precisão na recuperação de nódulos benignos e malignos utilizando os AT 3D, ANB 3D e o vetor AI. Para a base de avaliação contendo 260 nódulos (130 benigos e 130 malignos) foram alcançados os seguintes índices de precisão: AT 3D sem o ajuste de pesos obteve 87% de precisão com 25% revocação na recuperação de nódulos benignos e 91% de precisão com 25% de revocação para recuperação de nódulos malignos, após o ajuste dos pesos foi alcançada precisão 98% com revocação de 25% para recuperação de nódulos benignos e precisão de 99% com revocação de 25% para nódulos malignos; ANB 3D sem ajuste obteve 96% de precisão com 25% de revocação para recuperação de nódulos benignos e malignos, após o ajuste dos pesos foi alcançada precisão de 100% com 25% de revocação para recuperação de ambos os nódulos; por fim, AI sem ajuste de pesos obteve precisão de 82% com revocação de 25% para recuperação de nódulos benignos e 86% com revocação de 25% para recuperação de nódulos malignos, após o ajuste de pesos foi alcançada precisão de 95% com revocação de 25% para recuperação de nódulos benignos e 91% de precisão com revocação de 25% para recuperação de nódulos malignos.

Com base nos resultados alcançados, a DEP aplicada aos três vetores de atributos utilizados (AT 3D, ANB 3D e AI), com os pesos ajustados pelo processo apresentado, sempre apresentou melhores resultados do que os encontrados com a DE, ou seja, sem os ajustes de pesos. Isso comprova a hipótese de que a DEP proporcionaria melhores resultados do que a DE.

Entretanto, acreditamos que ainda seja possível melhorar o algoritmo. Para diminuir a probabilidade de resultados em pontos de máximos locais, pode ser incluinda aleatoriedade na busca pelos pesos com a utilização de algoritmos de busca heurística. Com o propósito de incluir mais informações úteis relacionadas à probabilidade de malignidade dos nódulos ao processo de atualização de pesos, a função de ajuste pode ser alterada para introduzir a noção da malignidade do nódulo no cálculo da distância.

Por fim, a partir da análise dos pesos dos atributos não foi possível avaliar quais os atributos que são indicadores das classes (benigno ou maligno), ou seja, não é possível afirmar que determinado atributo é bom ou ruim para determinar a classificação dos nódulos, já que no método de atualização de pesos não existe correlação dos pesos com a classe do nódulo. A atualização de pesos busca apenas uniformizar a amostra recuperada através da enfatização dos atributos cujos valores são mais similares e desenfatização dos atributos cujos valores são muito dispersos. Esta ênfase não corresponde a indicação de malignidade.

5.1 Limitações

Embora a precisão na recuperação dos nódulos tenha alcançado 100%, nós acreditamos que eles podem ser melhorados com a inserção de dois fatores que ainda não foram aplicados à nossa solução: a inserção da aleatoriedade na seleção dos pesos para atualizá-los por meio do uso de algoritmos de busca, por exemplo, para minimizar a possibilidade da ocorrências de máximos locais nos resultados encontrados; e criação de uma correlação entre a atualização de pesos e a malignidade do nódulo no processo de atualização, o que poderia melhorar os resultados alcançados através da adição desta informação importante para o método de recuperação.

Faz-se necessário também um estudo mais aprofundado para mensurar, em termos escalares, a influência dos parâmetros n, α, γ e a política de recompensas. Com isso, a determinação dos respectivos valores pode levar a resultados ainda melhores.

A recuperação dos nódulos feita neste trabalho esteve restrita à análise da malignidade. Não houve uma análise visual feita por especialistas para verificar se os nódulos semelhantes recuperados são semelhantes tanto devido a similaridade dos vetores, quanto são semelhantes visualmente segundo as características visuais utilizadas pelos especialistas.

Outro fator que merece maior análise é a utilização de outras métricas de similaridade como, por exemplo, a Distância Manhattan, a Distância Mahalanohis, a Variância Média Ponderada, e outras métricas definidas no espaço vetorial com a aplicação do conceito de pesos apresentados neste trabalho.

Também vale ressaltar a importância de buscar outras funções que tenham a capacidade de medir a dispersão dos dados como, por exemplo, a Entropia para que seja aplicada na cálculo dos pesos.

Por fim, na proposta apresentada não foi possível a indicação da relação entre os pesos dos atributos e as classes dos nódulos, ou seja, não foi possível afirmar quais atributos são bons para determinar nódulos como benignos ou malginos. Entretanto, determinar esta correlação pode alavancar as pesquisas que utilizam atributos extraídos dos nódulos como modelo representacional dos objetos.

5.2 Trabalhos futuros

Aqui apresento algumas melhorias que podem ser aplicadas como trabalhos futuros para aumento na precisão do algoritmo proposto:

- aplicação de algoritmos busca heurística junto à atualização de pesos com o objetivo de minimizar a possibilidade de acontecerem resultados de máximo local;
- alteração da função de ajuste de peso para que leve em consideração a relação da classe do objeto de referência com as classes dos objetos recuperados para dar mais ênfase aos nódulos semelhantes, segundo os valores dos atributos, que têm a mesma malignidade;
- avaliar outras políticas de recompensa e mensurar o seu impacto no ajuste dos pesos;
- avaliar o impacto dos três parâmetros do processo de ajuste de pesos: o número de nódulos recuperados (n), a taxa de ajuste (α) e o fator de desconto (γ);
- aplicar o processo de atualização com outras métricas de similaridade;
- utilizar outras funções que quantifiquem dispersão em amostras de dados com o intuito de compará-las com o IDP no ajuste dos pesos;
- buscar formas de correlacionar o peso à sua importância na classificação dos nódulos pulmonares;
- analisar, junto com especialistas em radiologia, se a similaridade dos vetores de atributos corresponde à similaridade visual dos nódulos recuperados ao nódulo de referência.

5.3 Contribuições Científicas do trabalho

Este trabalho foi desenvolvido no Laboratório de Telemedicina e Informática Médica (LaTIM) que está vinculado à Universidade Federal de Alagoas (UFAL) em parceria com o Hospital Universitário Professor Alberto Antunes (HUPAA). Durante 6 meses o projeto esteve amparado pelo Programa de Bolsas de Pós-Graduação *stricto sensu* da UFAL.

Os resultados parciais deste trabalho foram aceitos em eventos de relevância internacional e nacional com Qualis CAPES durante todo o período de desenvolvimento da proposta. Os trabalhos científicos aceitos para apresentação foram:

- LUCENA, D. J. F.; JUNIOR, J. R. F.; OLIVEIRA, M. C. . Avaliação da Precisão de Atributos de Textura 3D Normalizados Aplicados à Recuperação de Nódulos Pulmonares Similares. Anais do XIV Congresso Brasileiro de Informática em Saúde (CBIS), 2014 - Brasil.

- LUCENA, D. J. F.; JUNIOR, J. R. F.; OLIVEIRA, M. C. Caracterização de nódulos pulmonares através de um vetor de atributos ideal. Apresentação na categoria pôster. No: XIV Congresso Brasileiro de Informática em Saúde (CBIS), 2014, Santos - Brasil.

- LUCENA, D. J. F.; JUNIOR, J. R. F.; OLIVEIRA, M. C.; Pamponet, A. M. *Proposal of Local Automatic Weighing Attribute in CBIR.* Apresentação na categoria pôster. No: *eHealth-enabled Health* (MEDINFO), 2015, São Paulo - Brasil.

- LUCENA, D. J. F.; JUNIOR, J. R. F.; OLIVEIRA, M. C.; Pamponet, A. M. Atualização local automática de pesos de atributos para recuperação de nódulos pulmonares similares. Anais do XXVIII *Conference on Graphics, Patterns and Images* (SIBGRAPI), 2015, Salvador - Brasil.

- LUCENA, D. J. F.; JUNIOR, J. R. F.; OLIVEIRA, M. C.; Pamponet, A. M. *Proposal of local automatic weighing attribute to retrieve similar lung cancer nodules.* Anais do XI *Workshop* de Visão Computacional (WVC), 2015, São Carlos - Brasil.

REFERÊNCIAS

AKGÜL, C. B. et al. Content-based image retrieval in radiology: current status and future directions. *Journal of Digital Imaging*, Springer, v. 24, n. 2, p. 208–222, 2011.

AL SHALABI, L.; SHAABAN, Z.; KASASBEH, B. Data mining: A preprocessing engine. *Journal of Computer Science*, v. 2, n. 9, p. 735–739, 2006.

ALLISY-ROBERTS, P.; WILLIAMS, J. R. *Farr's physics for medical imaging*. 2nd. ed. [S.l.]: Elsevier Health Sciences, 2008. ISBN 9780702028441.

ARMATO, S. G. et al. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, v. 38, p. 915–931, 2011.

ATKESON, C. G.; MOORE, A. W.; SCHAAL, S. Locally weighted learning. *Artificial Intelligence Review*, Kluwer Academic Publishers, v. 11, n. 1-5, p. 11–73, 1997. ISSN 0269-2821. Disponível em: <http://dx.doi.org/10.1023/A%3A1006559212014>.

AZEVEDO-MARQUES, P. M. Diagnóstico auxiliado por computador na radiologia. *Radiologia Brasileira*, SciELO Brasil, v. 34, 2001.

BEDO, M. V. N. et al. Endowing a content-based medical image retrieval system with perceptual similarity using ensemble strategy. *Journal of Digital Imaging*, Springer US, p. 1–16, 2015. ISSN 0897-1889. Disponível em: <http://dx.doi.org/10.1007/s10278-015-9809-1>.

BUSHBERG, J. T.; BOONE, J. M. *The essential physics of medical imaging*. [S.l.]: Lippincott Williams & Wilkins, 2011.

DHARA, A. K. et al. Content-based image retrieval system for differential diagnosis of lung cancer. *Indian Journal of Medical Informatics*, v. 6, n. 1, p. 1, 2012.

DICIOTTI, S. et al. The characteristic scale: A consistent measurement of lung nodule size in ct imaging. *Medical Imaging, IEEE Transactions on*, IEEE, v. 29, n. 2, p. 397–409, 2010.

DOI, K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, v. 31, n. 4-5, p. 198–211, 2007.

EADIE, L. H.; TAYLOR, P.; GIBSON, A. P. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European Journal of Radiology*, v. 81, n. 1, p. e70 – e76, 2012. ISSN 0720-048X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0720048X11001574>.

FACELI, K. et al. *Inteligência artificial - Uma abordagem de aprendizagem de máquina*. [S.l.]: LTC, 2011. ISBN 9788521618805.

FARIA, F. F. et al. Learning to rank for content-based image retrieval. In: *Proceedings of the International Conference on Multimedia Information Retrieval*. New York, NY, USA: ACM, 2010. (MIR '10), p. 285–294. ISBN 978-1-60558-815-5. Disponível em: <http://doi.acm.org/10.1145/1743384.1743434>.

FDA, U. F. . D. A. *Computed Tomography (CT)*. 2015. Disponível em: <http://www.fda.gov/Radiation-EmittingProducts/RadiationEmittingProductsandProcedures/MedicalImaging/MedicalX-Rays/ucm115317.htm>.

Ferreira Junior, J. R. *Auxílio Computadorizado ao Diagnóstico do Câncer de Pulmão Otimizado por GPU*. Dissertação (Mestrado) — Universidade Federal de Alagoas, Alagoas, 2015.

Ferreira Junior, J. R.; OLIVEIRA, M. C. Banco de dados nosql público de nódulos pulmonares para auxílio à pesquisa e diagnóstico do câncer de pulmão. In: *Anais do XXIV - Congresso Brasileiro de Engenharia Biomédica (CBEB)*. [S.l.: s.n.], 2014. p. 177–180.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. [S.l.]: Pearson/Prentice Hall, 2008. ISBN 9780131687288.

HAN, F. et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *Journal of Digital Imaging*, Springer, p. 1–17, 2014.

HAN, F. et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *Journal of digital imaging*, Springer, v. 28, n. 1, p. 99–115, 2015.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. H. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, IEEE, n. 6, p. 610–621, 1973.

INCA, I. N. do Câncer José de Alencar Gomes da S. *Tipos de câncer - Pulmão*. 2015. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao/definicao>.

JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, ACM, v. 20, n. 4, p. 422–446, 2002.

KENROSE, S. *Z-Score: Definition, Formula and Calculation*. 2015. Disponível em: <http://www.statisticshowto.com/how-to-calculate-a-z-score/>.

KULKARNI, P.; KULKARNI, S.; STRANIERI, A. A novel architecture and analysis of challenges for combining text and image for medical image retrieval. *International Journal for Infonomics (IJI)*, v. 7, 2014.

KUMAR, A. et al. Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data. *Journal of Digital Imaging*, Springer US, v. 26, n. 6, p. 1025–1039, 2013. ISSN 0897-1889. Disponível em: <http://dx.doi.org/10.1007/s10278-013-9619-2>.

KURUVILLA, I.; GUNAVATHI, K. Content based image retrieval for ct images of lungs. *International Journal of Soft Computing*, v. 9, n. 6, p. 386–390, 2014.

LIMA, I. C. P. C.; PIMENTA, C. A. M. Estigma do câncer de pulmão: conceito, fatores associados e avaliação. *Investigación en Enfermería: Imagen y Desarrollo*, Pontificia Universidad Javeriana, v. 17, p. 97–112, 2015.

LUCENA, D. J. F. d.; Ferreira Junior, J. R.; OLIVEIRA, M. C. Avaliação da precisão de atributos de textura 3d normalizados aplicados à recuperação de nódulos pulmonares similares. XIV Congresso Brasileiro em Informática em Saúde, 2014.

LUCENA, D. J. F. d. et al. Atualização local automática de pesos de atributos para recuperação de nódulos pulmonares similares. In: . [s.n.], 2015. Disponível em: <http://sibgrapi.sid.inpe.br/col/sid.inpe.br/sibgrapi/2015/07.13.22.57/doc/SIBGRAPI-VERSAO-APROVADA2.pdf>.

LUCENA, D. J. F. d. et al. Proposal of local automatic weighing attribute to retrieve similar lung cancer nodules. *XI Workshop de Visão Computacional - WVC*, 2015. Disponível em: <https://www.researchgate.net/profile/Jose_Ferreira_Junior/publication/284031503_Proposal_of_Local_Automatic_Weighing_Attribute_to_Retrieve_Similar_Lung_Cancer_Nodules/links/564b5e1108aeab8ed5e7463f.pdf>.

MAHMOUD-GHONEIM, D. et al. Three dimensional texture analysis in mri: a preliminary evaluation in gliomas. *Magnetic resonance imaging*, Elsevier, v. 21, n. 9, p. 983–987, 2003.

MILLIGAN, G. W.; COOPER, M. C. A study of standardization of variables in cluster analysis. *Journal of classification*, Springer, v. 5, n. 2, p. 181–204, 1988.

MüLLER, H. et al. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International Journal of Medical Informatics*, v. 73, n. 1, p. 1 – 23, 2004. ISSN 1386-5056. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1386505603002119>.

MÜLLER, H. et al. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters*, Elsevier, v. 22, n. 5, p. 593–601, 2001.

OLIVEIRA, M. C. *Grids Computacionais para recuperação de imagens médicas a partir de conteúdo: um estudo de viabilidade*. Tese (Doutorado) — Universidade de São Paulo, São Paulo, 2006.

OLIVEIRA, M. C.; AZEVEDO-MARQUES, P. M.; FILHO, W. C. C. Grades computacionais na otimização da recuperação de imagens médicas baseada em conteúdo. *Radiol Bras*, SciELO Brasil, v. 40, p. 255–61, 2007.

OLIVEIRA, M. C.; CIRNE, W.; AZEVEDO-MARQUES, P. M. Towards applying content-based image retrieval in the clinical routine. *Future Generation Computer Systems*, v. 23, n. 3, p. 466–474, 2007.

PASTORINO, U. Lung cancer screening. *British journal of cancer*, Nature Publishing Group, v. 102, n. 12, p. 1681–1686, 2010.

PONCIANO-SILVA, M. et al. Does a cbir system really impact decisions of physicians in a clinical environment? In: *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. [S.l.: s.n.], 2013. p. 41–46.

REEVES, A. P.; KOSTIS, W. J. Computer-aided diagnosis of small pulmonary nodules. *Seminars in Ultrasound, {CT} and {MRI}*, v. 21, n. 2, p. 116 – 128, 2000. ISSN 0887-2171. The Solitary Pulmonary Nodule. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0887217100900180>.

Seitz Jr, K. A. et al. Learning lung nodule similarity using a genetic algorithm. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *SPIE Medical Imaging*. [S.l.], 2012. p. 831537–831537.

SILVA, M. P. d. *Processamento de consultas por similaridade em imagens médicas visando à recuperação perceptual guiada pelo usuário*. Dissertação (Mestrado) — Universidade de São Paulo, São Paulo, 2009. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-15052009-110247/pt-br.php>.

TEAM, N. L. S. T. R. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine*, NIH Public Access, v. 365, n. 5, p. 395, 2011.

TRAINA, A. J. M. et al. Efficient content-based image retrieval through metric histograms. *World Wide Web*, Springer, v. 6, n. 2, p. 157–185, 2003.

UEHARA, C.; JAMNIK, S.; SANTORO, I. L. Câncer de pulmão. *Medicina (Ribeirao Preto. Online)*, v. 31, n. 2, p. 266–276, 1998.

VISALAKSHI, N. K.; THANGAVEL, K. Impact of normalization in distributed k-means clustering. *International Journal of Soft Computing*, v. 4, n. 4, p. 168–172, 2009.

VOGEL, J.; SCHIELE, B. Performance evaluation and optimization for content-based image retrieval. *Pattern Recognition*, Elsevier, v. 39, n. 5, p. 897–909, 2006.

WENDER, R. et al. American cancer society lung cancer screening guidelines. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 63, n. 2, p. 106–117, 2013.

XU, J. et al. Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval. *Medical Physics*, v. 39, p. 5405–5418, 2012.

ZAMBONI, M. Epidemiologia do câncer do pulmão. *J Pneumol*, SciELO Brasil, v. 28, n. 1, p. 41–7, 2002.

APÊNDICE A – ACESSO AO BANCO DE IMAGENS

Existem duas versões do banco de imagens: uma para acesso local e outra para acesso remoto através de uma nuvem pública.

Para acesso local, estão disponíveis arquivos de *backup* através do endereço <http://bit.ly/1NISgs9> (acessado em 12 de julho de 2015) e que para serem utilizados basta realizar a operação de *restore* do MongoDB.

Para acesso remoto, o banco de imagens foi implantado em uma estrutura de nuvem, para garantir a disponibilidade dos dados. A plataforma utilizada foi a Morpheus, disponível através do endereço <www.gomorpheus.com/public-cloud> (acessado em 12 de julho de 2015), que é um serviço que permite o desenvolvimento, armazenamento e distribuição de bancos de dados em arquitetura MongoDB. A leitura dos dados pode ser feito pelo MongoDB Shell, API ou por uma ferramenta de gerenciamento de bancos de dados em MongoDB. As configurações para acesso ao banco são: readonly é o nome do usuário com privilégios de somente-leitura, gH@h6NL38V é a senha do usuário, 162.252.108.127 é o IP, 12279 é a porta, e publicDB é o nome da base de dados.