Uso de Mapeamento Conceitual para Redução da Descontinuidade Semântica na Recuperação de Imagens Microscópicas de Carcinoma Tireoidiano[1]

Hugo Cesar Pessotti

DEFESA APRESENTADA AO
PROGRAMA INTERUNIDADES EM BIOINFORMÁTICA DA
UNIVERSIDADE DE SÃO PAULO

Programa Interunidades de Pós-Graduação em Bioinformática

Orientador: Profa. Dra. Alessandra Alaniz Macedo

Ribeirão Preto, Março de 2012

[1]Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da FAPESP (2008/08098-1)

Uso de Mapeamento Conceitual para Redução da Descontinuidade Semântica na Recuperação de Imagens Microscópicas de Carcinoma Tireoidiano

Esta versão corresponde à Defesa
de Hugo Cesar Pessotti a ser
apresentada para a Comissão Julgadora
e defendida em XX/03/2012.

Comissão Julgadora:

- Prof. Dr. Nome Completo - Instituição
- Prof. Dr. Nome Completo - Instituição
- Prof. Dr. Nome Completo - Instituição

Agradecimentos

Agradeço aos meus pais e minhas irmãs por todos esses anos de boa convivência, essenciais para minha formação e educação, e por constituir um Lar acima de tudo.

A minha grande amiga, Lariza, companheira de todas as horas, por todos anos de boa convivência, confiança, carinho e dedicação.

Aos amigos do Laboratório de Informática em Saúde (LIS), por sempre criarem um ambiente ideal para trabalho e pesquisa, com seus momentos de descontração e estudo.

A minha orientadora, Alessandra, pela paciência que teve comigo durante esses anos, pela seriedade que sempre levou nossos projetos, sua insistência em pontos importantes do trabalho e por sempre ter sido exigente com a qualidade como um todo.

Ao Prof. Dr. Edson Garcia Soares, pelas discussões muito produtivas e por sempre ter me recebido bem e doado um pouco de seu tempo para a execução deste projeto.

Ao Serviço de Patologia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP-USP), por ter fornecido as imagens utilizadas neste projeto.

Ao Antonio Renato Meirelles e Silva, do Laboratório de Neurologia Aplicada e Experimental da FMRP-USP, por ter auxiliado na captura das imagens utilizadas neste projeto.

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa de Mestrado concedida (Processo FAPESP 2008/08098-1).

Por fim, a todos que passaram e que irão passar pela minha vida, por deixarem em mim uma marca, cada um a sua maneira.

Resumo

O uso de ferramentas computacionais na área de imagens médicas tem crescido nos últimos anos. Esses sistemas normalmente utilizam mecanismos de Processamento de Imagens e Recuperação de Imagens Baseada em Conteúdo como técnicas de suporte para extração de atributos de imagem, desprezando informações semânticas relacionadas aos exames. Técnicas de Recuperação de Informação, artefatos de Processamento de Linguagem Natural e Morfometria podem ser aplicados para encontrar mapeamentos entre o conteúdo visual de uma imagem médica e informações semânticas úteis ao usuário. Este mapeamento pode resultar na geração de conhecimento que, posteriormente, poderá ser utilizado para reconhecer padrões de exames, refinar buscas por similaridade em uma base de imagens e fornecer sugestões de diagnóstico suportadas por exames similares. O objetivo do trabalho é realizar um estudo teórico-prático de técnicas das áreas de Recuperação de Imagens Baseada em Conteúdo e Recuperação de Informação para a diminuição da descontinuidade semântica existente entre a recuperação computadorizada de imagens médicas e a interpretação humana de seu conteúdo. Com este estudo, foi abstraído o *Framework* para Redução da Descontinuidade Semântica em Imagens Médicas (FREDS). Os resultados iniciais no contexto de câncer de tireoide em imagens de microscopia óptica sinalizaram positivamente o uso da informação semântica extraída a partir de imagem.

Palavras-chave: processamento de imagens, recuperação de informação, recuperação de imagens baseada em conteúdo, morfometria, câncer de tireoide, carcinoma tireoidiano.

Abstract

The use of computational tools in the field of medical imaging has increased in the recent years. These systems typically employ Image Processing and Content-Based Image Retrieval mechanisms as support techniques for extraction of image attributes, ignoring semantic information related to the exams. Information Retrieval techniques, artifacts of Natural Language Processing and Morphometry can be applied to find mappings between the visual content of a medical image and the semantic information found in the exams. This mapping can result in the generation of knowledge, which subsequently could be used to recognize exam patterns, refine searches by similarity in an image database and provide diagnosis suggestions supported by similar exams. The present work aims to perform a theoretical-pratical study in the areas of Content-Based Image Retrieval and Information Retrieval to reduce the semantic discontinuity that exists between the computerized medical image retrieval and the human interpretation of its content. From this study, we abstracted FREDS - *Framework para Redução da Descontinuidade Semântica em Imagens Médicas* (Framework for Semantic Discontinuity Reduction in Medical Imaging). Initial results in the context of thyroid cancer in microscopic images signaled positively the use of semantic information extracted from the image.

Keywords: image processing, information retrieval, content-based image retrieval, morphometry, thyroid cancer, thyroid carcinoma.

Sumário

1 Introdução **1**

2 Fundamentos Teóricos **3**
2.1 Câncer de Tireoide . . . 3
2.2 Morfometria . . . 4
2.3 Recuperação de Imagens Baseada em Conteúdo . . . 5
2.3.1 Segmentação e Extração de Objetos . . . 6
2.3.2 Matrizes de Coocorrência e Descritores de Haralick . . . 6
2.3.3 Transformada Discreta de Wavelets . . . 8
2.3.4 K-Vizinhos Próximos . . . 9
2.3.5 Redes Neurais Multicamada . . . 11
2.4 Processamento de Linguagem Natural . . . 13
2.4.1 Normalização Linguística . . . 13
2.4.2 Redução da Ambiguidade . . . 14
2.5 Recuperação de Informação . . . 14
2.5.1 Modelo Vetorial . . . 15
2.5.2 Similaridade por Cosseno . . . 15
2.6 Trabalhos Relacionados . . . 15

3 O *Framework* FREDS **17**
3.1 Rotulação de Imagens . . . 20
3.1.1 Rotulação Automatizada . . . 21
3.1.2 Rotulação Auxiliada pelo Usuário . . . 21
3.2 Seleção de Laudos . . . 22
3.3 Validação do FREDS . . . 23

4 Experimentação e Resultados **25**
4.1 Coleção de Imagens e Laudos . . . 25
4.2 Experimentação e Resultados . . . 27

5 Conclusões e Trabalhos Futuros **35**

Referências Bibliográficas **39**

Capítulo 1

Introdução

Os nódulos tireoidianos estão entre as doenças mais comuns envolvendo a tireoide. Nódulos palpáveis são encontrados entre 4 e 7% da população adulta (com maior incidência entre mulheres e idosos), aumentando para o intervalo de 19 a 67% em exames de ultrassom, devido à sua maior precisão [MAIA, 2007]. Essas estatísticas indicam a necessidade de um diagnóstico mais acurado para início do tratamento mais adequado ao paciente com câncer. Para pacientes com câncer de tiroide, a sobrevivência estimada em 10 anos é de 95% para o câncer papilífero e 75% para o folicular [COELI, 2005]. Esses índices são considerados altos quando o diagnóstico é fornecido precocemente. Dentre os carcinomas tireoidianos, os mais frequentes são os carcinomas papilífero e folicular.

O tipo de exame mais utilizado para distinguir nódulos benignos e malignos é a Punção Aspirativa com Agulha Fina, por ser um procedimento de fácil execução, seguro e de baixo custo [MAIA, 2007]. No entanto, a análise visual do material colhido, realizada em microscópios ópticos, é uma tarefa trabalhosa e exaustiva mesmo para profissionais experientes, devido a diferenças sutis de textura e de morfologia do tecido tireoidiano[DASKALAKIS, 2008]. Um sistema de apoio ao diagnóstico capaz de diferenciar entre os tipos de nódulos pode ajudar o patologista a concluir diagnósticos com o fornecimento de informação extra, evitando procedimentos ou exames desnecessários em pacientes com nódulos benignos e aumentando a sobrevida dos pacientes no caso de câncer.

O uso de sistemas computacionais pode auxiliar o estabelecimento do diagnóstico pela análise de imagens, uma vez que fornecem informações complementares de modo a diminuir o tempo e o esforço necessários para analisar um exame. Tradicionalmente, esses sistemas utilizam principalmente técnicas de Processamento de Imagens e Recuperação de Imagens Baseada em Conteúdo (*Content-Based Image Retrieval* - CBIR), que se concentram nos atributos das imagens [DOI, 2007]. A extração de atributos em CBIR pode não oferecer ajuda substancial durante o processo de recuperação, pois o profissional de saúde muitas vezes não está familiarizado com os atributos escolhidos, especialmente os de baixo nível como, por exemplo, energia e entropia de uma imagem [MUEEN, 2008]. O uso de informações em nível semântico pode aprimorar a recuperação de imagens, tornando possível ao usuário especificar mais facilmente sua intenção de busca e também auxiliar a interpretação dos resultados.

Exames patológicos são formados por imagens microscópicas e acompanhados por informações textuais inseridas pelo patologista. As informações textuais podem ser utilizadas pelo patologista em conjunto com a análise das imagens para a elaboração de diagnóstico, visando eliminar hipóteses e fundamentar o laudo. Nesse sentido, técnicas de Recuperação de Informação (*Information Retrieval* - IR) e artefatos linguísticos para Processamento de Linguagem Natural (*Natural Language Processing* - NLP) podem ser aplicados para buscar e definir mapeamentos entre o conteúdo de uma imagem médica e as informações

textuais presentes no exame, agregando conhecimento ao sistema computacional. Esse conhecimento pode ser posteriormente utilizado para reconhecer padrões de laudo, refinar buscas por similaridade em uma base de imagens, fornecer sugestões de diagnóstico de acordo com exames similares já realizados, entre outras aplicações.

Em sistema de apoio ao diagnóstico, espera-se que o uso de um mapeamento semântico possa auxiliar a tarefa de análise de imagens e diagnósticos, retornando para o usuário sugestões baseadas em laudos de imagens similares já diagnosticadas por outros patologistas, com um aumento de precisão e confiabilidade do sistema proporcional à experiência do grupo de patologistas.

O presente trabalho tem como objetivo contribuir com a diminuição da descontinuidade semântica existente entre a recuperação computadorizada de imagens médicas e a interpretação humana de seu conteúdo, realizando mapeamento semântico para geração de conhecimento. Este mapeamento está sendo aplicado no contexto de sistemas de apoio ao diagnóstico médico, com o objetivo de fornecer informações complementares aos exames para auxiliar os patologistas na elaboração do diagnóstico final. Estas informações complementares são obtidas por meio de dados morfométricos extraídos dos núcleos celulares presentes nas imagens microscópicas dos exames.

Realizou-se um estudo teórico-prático de técnicas das áreas de CBIR, IR e NLP para a definição de mapeamentos conceituais entre o conteúdo de uma imagem médica e as informações textuais presentes no exame. Aplicou-se os resultados no contexto de câncer de tireoide em imagens de microscopia óptica. Foram utilizadas imagens de exames de Punção Aspirativa com Agulha Fina, juntamento com seus laudos, a fim de investigar o potencial de uso do núcleo celular como classificador de casos de carcinoma papilífero, bócio coloide e lesões indeterminadas.

O restante desta dissertação está organizada da seguinte maneira: o Capítulo 2 apresenta os fundamentos teóricos; o Capítulo 3 apresenta o *framework* FREDS; o Capítulo 4 apresenta os resultados do estudo e o Capítulo 5, as conclusões obtidas e trabalhos futuros.

Capítulo 2

Fundamentos Teóricos

Durante a execução do presente trabalho, foram estudados conceitos relacionados a processamento de imagens e de informação textual. Neste capítulo serão descritos conceitos de câncer de tireoide e as técnicas empregadas no trabalho. A Seção 2.1 contém uma discussão sobre o tema central do trabalho, Câncer de Tireoide; a Seção 2.2 apresenta a Morfometria, uma fonte de informação semântica utilizada neste trabalho; a Seção 2.3 apresenta as técnicas de Recuperação de Imagens Baseada em Conteúdo; a Seção 2.4, Processamento de Linguagem Natural; a Seção 2.5, Recuperação de Informação; por fim, na Seção 2.6 são apresentados alguns trabalhos relacionados.

2.1 Câncer de Tireoide

Câncer é uma denominação utilizada para representar centenas de doenças que possuem algumas características em comum entre elas, principalmente, o crescimento desordenado e espalhamento de células para outros tecidos e regiões do corpo [GABRIEL; EBRARY, 2007].

O câncer tem sido diagnosticado há mais de 100 anos. Contudo, suas causas permaneceram desconhecidas até a descoberta do DNA recombinante na década de 70, quando o câncer foi considerado uma doença genética decorrente do acúmulo de alterações nas células [DEVITA, 1997; PANNO, 2004]. O desenvolvimento das técnicas de sequenciamento de DNA também melhorou significativamente o conhecimento que se tinha sobre o câncer. Mundialmente, o número de casos de câncer tem aumentado. Em 2008 foram estimados aproximadamente 12,7 milhões de casos e 7,6 milhões de mortes causadas pelo câncer, sendo que 56% dos casos e 64% das mortes ocorreram em países em desenvolvimento, como o Brasil [JEMAL, 2011].

Os genes responsáveis pelo câncer podem ser agrupados em dois tipos: os oncogenes e os supressores tumorais. Esses dois tipos de genes estão, como todos os demais, sujeitos a mutações. Quando mutados os oncogenes tendem a acelerar o processo de surgimento de um câncer. Já os supressores tumorais agem naturalmente contra o surgimento de cânceres porém, quando mutados, podem perder sua função supressora [DEVITA, 1997].

Os nódulos tireoidianos estão entre as doenças mais comuns envolvendo a tireoide. Nódulos palpáveis são encontrados em 4 a 7% da população adulta (com maior incidência entre mulheres e idosos), aumentando para 19 a 67% em exames de ultra-som devido a sua maior precisão [MAIA, 2007]. Embora sejam comuns, apenas 5% dos nódulos são malignos e representam câncer de tireoide, justificando a necessidade de um diagnóstico mais acurado para início do tratamento mais adequado ao paciente com câncer.

Estudos demonstram que nódulos solitários em geral têm maior probabilidade de serem neoplásicos, assim como os nódulos presentes em pacientes jovens e também pacientes do sexo masculino [ROBBINS,

2009]. Dentre os carcinomas tireoidianos os mais frequentes são os diferenciados, caracterizados como carcinoma papilífero (aproximadamente 85% dos casos) e carcinoma folicular (de 5 a 15% dos casos) [ROBBINS, 2009]. A sobrevivência estimada em 10 anos é de 95% para o carcinoma papilífero e 75% para o folicular [COELI, 2005].

Morfologicamente, o carcinoma papilífero é uma lesão solitária ou multifocal que pode apresentar limites bem definidos ou até mesmo ser parcialmente encapsulada [MELMED, 2008], mas também pode ser infiltrada no parênquima adjacente, resultando em bordas pouco definidas. Sua organização celular é predominantemente de forma papilar, embora esta não seja uma característica essencial para definir o diagnóstico.

O carcinoma papilífero pode apresentar variantes que confundem o patologista, imitando outros tipos de lesões de tireoide. Estas variantes podem representar até 20% de todos os casos de carcinoma papilífero [MELMED, 2008]. A variante mais comum é a folicular, cujas células apresentam todas características de um carcinoma papilífero, porém se organizam de forma a constituir uma arquitetura folicular. Esta variante pode levar o patologista a errar seu diagnóstico se não foram observadas as características das células.

Embora existam vários procedimentos para determinar a malignidade de um nódulo tireoidiano, apenas a análise do tecido pode dar o diagnóstico conclusivo de câncer [FELIG, 2001]. A modalidade de exame mais utilizada para distinguir nódulos benignos e malignos é a Punção Aspirativa com Agulha Fina (PAAF), por ser um procedimento de fácil execução, seguro e de baixo custo [MAIA, 2007], além de ser o exame citológico de maior sensibilidade e especificidade para determinar a malignidade de um nódulo [FELIG, 2001]. Em estudos realizados, a identificação das lesões tireoidianas por meio de PAAF reduziu em até 50% o número de cirurgias em pacientes cujos casos podiam ser acompanhados com segurança e aumentou em até duas vezes o número de tumores retirados por cirurgia [GHARIB, 1994].

No entanto, a análise do material colhido em exames de PAAF é realizada por microscópios ópticos, tornando-a uma tarefa trabalhosa e exaustiva mesmo para profissionais experientes devido a diferenças sutis de textura e morfologia do tecido tireoidiano[DASKALAKIS, 2008].

O diagnóstico do carcinoma papilífero é dado considerando as características nucleares das células de tireoide, constituindo os padrões de achados patológicos [ROBBINS, 2009]. O núcleo das células de um carcinoma papilífero apresenta cromatina dispersa, que lhes confere um aspecto de menor densidade. Tradicionalmente o núcleo apresenta invaginações que, dependendo do plano de corte, resulta em fendas nucleares ou pseudo-inclusões quando observadas ao microscópio.

Também podem ser encontradas estruturas calcificadas em aproximadamente 40% dos casos de carcinoma papilífero [FELIG, 2001]. Estas estruturas são conhecidas como corpos psamomatosos e são quase exclusivamente encontradas em carcinomas papilíferos. Embora não seja uma característica única deste tipo de carcinoma, sua presença pode indicar suspeita de câncer.

Um sistema de apoio ao diagnóstico capaz de identificar achados patológicos e localizar estruturas importantes em uma imagem de microscopia poderia ajudar o patologista a concluir diagnósticos com o fornecimento de informação extra, podendo evitar procedimentos cirúrgicos desnecessários em pacientes com nódulos benignos e aumentar a sobrevida dos pacientes no caso de câncer com o tratamento precoce.

2.2 Morfometria

Um dos principais meios de diagnosticar doenças durante a análise de imagens microscópicas é a busca por alterações no tecido celular, seja na sua organização ou então as características individuais das células. A morfometria surgiu como um método quantitativo importante na caracterização das células de forma

individual com o uso de métricas e medidas referentes ao núcleo e o citoplasma [JR, 2007]. Embora aplicada com sucesso em lesões de mama, a abordagem morfométrica ainda é pouco explorada em lesões de tireoide dado a subjetividade dos exames [PRIYA; SUNDARAM, 2011].

Entre as características mais estudadas nas células, encontram-se métricas relacionadas com seu raio, perímetro e área, dos quais derivam-se outras métricas como, por exemplo, razão de área núcleo/citoplasma, circularidade e excentricidade. Estes parâmetros celulares podem ser utilizados como informação complementar na Recuperação de Imagens Baseada em Conteúdo, que tradicionalmente emprega técnicas de análise de textura das imagens para descrevê-las.

Em lesões tireoidianas, [PRIYA; SUNDARAM, 2011] conduziu um estudo estatístico utilizando a análise de variância (ANOVA) em imagens de PAAF com parâmetros morfológicos incluindo diâmetro, perímetro e área média nuclear, taxa de circularidade, variação de área nuclear e razão maior/menor diâmetro. Nesse estudo, o carcinoma do tipo folicular obteve uma razão maior/menor diâmetro superior ao bócio adenomatoso e o menor perímetro nuclear, enquanto que o carcinoma anaplásico apresentou o maior perímetro nuclear entre todos os tipos de carcinoma. [DINA, 2000] estudou a importância da morfometria na diferenciação entre a variante "células altas"e o tipo clássico do carcinoma papilífero, obtendo significância estatística para os parâmetros diâmetro e desvio padrão da área nuclear.

Por outro lado, [WRIGHT, 1987] reconheceu a importância da morfometria mas a classificou como inadequada quando empregada como único método para predizer a malignidade de tumores tireoidianos, necessitando de técnicas complementares. Segundo os autores, foi encontrada uma diferença significativa da área e do perímetro nuclear em pacientes com tumor benigno e maligno, porém devido à grande variância encontrada nesses parâmetros a interpretação dos resultados era dificultada.

O estudo de parâmetros morfométricos pode auxiliar o diagnóstico pela quantificação e descrição numérica dos achados patológicos, podendo detectar mudanças morfológicas nas células devido a doenças e assim constituir uma fonte de informação semântica importante na classificação dos nódulos.

Neste trabalho foram consideradas as medidas mais encontradas na literatura em termos de parâmetros morfométricos, sendo eles: raio médio, perímetro, área, Índice de Contorno Nuclear (ICN) e razão maior/-menor raio. O Índice de Contorno Nuclear é uma medida que indica o quanto um núcleo se aproxima de uma circunferência perfeita dado um plano de corte. O ICN é calculado pela Equação 2.1.

$$ICN = \frac{p}{A^2} \tag{2.1}$$

cujo p é o perímetro do núcleo e A sua área. O ICN de um círculo perfeito é aproximadamente 3.544907.

2.3 Recuperação de Imagens Baseada em Conteúdo

Esta seção apresenta as técnicas de processamento de imagens e algoritmos de inteligência artificial empregados neste trabalho. A Subseção 2.3.1 descreve a segmentação de objetos. As duas subseções seguintes descrevem as técnicas de extração de características e as duas últimas apresentam os algoritmos de classificação de padrões. Na Subseção 2.3.2 são descritos o algoritmo de matriz de coocorrência e os descritores de Haralick; na Subseção 2.3.3, transformada discreta de Wavelets; na Subseção 2.3.4, o classificador k-vizinhos próximos e na Subseção 2.3.5, redes neurais multicamada.

2.3.1 Segmentação e Extração de Objetos

A segmentação é uma técnica de processamento de imagens que envolve o particionamento de uma imagem em regiões, partindo do pressuposto que os pixeis de cada uma dessas regiões possuem determinadas propriedades ou características em comum [ACHARYA; RAY, 2005]. Como resultado, a segmentação produz regiões homogêneas de acordo com uma característica ou propriedade que se deseja observar.

Uma das técnicas mais simples de segmentação é o *threshold*, ou limiarização [PETROU; PETROU, 2010]. Essa técnica consiste na criação do histograma do número de pixeis por nível de cinza que representa a função de densidade de probabilidade da imagem em questão. A partir do histograma escolhe-se um ponto de corte, ou limiar, de modo que pixeis com tons de cinza abaixo deste limiar recebem valor 1 e o restante recebe o valor 0. Ao final do processo obtém-se uma máscara binária para a imagem, separando efetivamente o fundo da imagem e as regiões segmentadas.

Neste trabalho, a limiarização foi empregada como estratégia de segmentação para localizar os núcleos celulares. Em uma imagem citológica, os núcleos apresentam tons de cinza mais fortes que os tons do citoplasma ao seu redor, tornando possível o uso de algoritmos de limiarização. O método escolhido para determinar de forma automatizada o limiar de corte foi o Método de Otsu [OTSU, 1975], tradicionalmente empregado para calcular o ponto de corte cuja variância entre o fundo da imagem e as regiões segmentadas é máxima. Segundo o Método de Otsu, o ponto ótimo de corte é o valor t que minimiza a função de variância intra-classe, apresentada na Equação 2.2. Este valor é encontrado por meio de busca exaustiva, variando-se t entre todos os tons de cinza até encontrar a menor variância intra-classe.

$$\sigma^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \tag{2.2}$$

cujo $w_1(t)$ é a quantidade e $\sigma_1^2(t)$ a variância dos pixeis com tons de cinza menor que t e $w_2(t)$ é a quantidade e $\sigma_2^2(t)$ a variância dos pixeis com tons de cinza maior que t.

Após a segmentação, as regiões conexas da máscara binária são identificadas e extraídas da imagem individualmente. Este processo é descrito no Algoritmo 1.

2.3.2 Matrizes de Coocorrência e Descritores de Haralick

As matrizes de coocorrência (*Gray-Level Co-occurrence Matrices* - GLCM) fazem parte de um conjunto de técnicas de processamento de imagens voltadas para descrever estatisticamente a textura de uma imagem, considerando a distribuição e o relacionamento entre seus níveis de cinza [HARALICK, 1973].

Sendo M a GLCM de uma dada imagem, o relacionamento entre um par de níveis de cinza i e j é usualmente calculado em quatro direções (0, 45, 90 e 135 graus) em relação ao eixo x de um pixel (ver Figura 1). As outras direções (180, 225, 270 e 315 graus) podem ser obtidas a partir das quatro principais, pois seus vizinhos serão os mesmos, porém invertidos. Além da direção, uma distância entre o pixel central e sua vizinhança de interesse é estabelecida.

Dada uma distância δ e uma direção θ, a GLCM M obtida a partir de uma imagem I de dimensões n x m e k níveis de cinza é calculada segundo o Algoritmo 2. A partir da soma das matrizes resultantes de cada uma das quatro possíveis direções, obtém-se a GLCM final. Em seguida, esta matriz é normalizada dividindo-a pela soma de todos seus elementos, de modo que a soma final de todos eles seja igual a 1. Após a normalização da GLCM final, são calculadas métricas baseadas nos descritores de Haralick [HARALICK, 1973]. Em seu trabalho original, Haralick propôs quatorze características (descritores), porém devido à correlação existente entre elas este conjunto pode ser reduzido para cinco [CONNERS; HARLOW, 1980].

```
função extracaoObjetos(I, limiar)
  converterTonsCinza(I)

  //limiarização da imagem
  para cada pixel (x,y) da imagem I
    se obterValorPixel(x, y) menor que limiar
      mascaraBinaria[x, y] <- 0
    senão
      mascaraBinaria[x, y] <- 1
    fim se
  fim para

  //identificação das regiões conexas
  para cada posicao (x,y) da mascaraBinaria
    se mascaraBinaria[x, y] igual a 1
      i <- 0
      bordaObjeto <- vetor[]
      ponto <- (x, y)
      pontoVizinho <- encontrarVizinhoValorUm(ponto)

      bordaObjeto[i] <- ponto
      i <- i + 1
      bordaObjeto[i] <- pontoVizinho
      i <- i + 1

      //percorre-se os vizinhos até voltar ao ponto inicial
      enquanto pontoVizinho for diferente de ponto
        pontoVizinho <- encontrarVizinhoValorUm(ponto)
        bordaObjeto[i] <- pontoVizinho
        i <- i + 1
      fim enquanto

      //extração do objeto contido no interior da borda
      para cada posicao (i,j) no interior da bordaObjeto
        objeto[i,j] <- obterValorPixel(i, j)
        mascaraBinaria[i,j] <- 0
      fim para
    fim se
  fim para
fim
```

Algoritmo 1: *Segmentação e extração de regiões conexas de uma imagem*

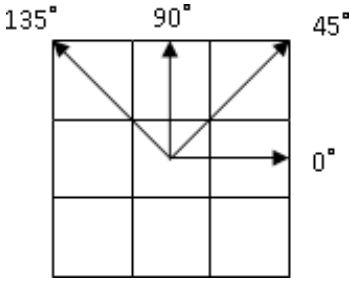

Figura 1: *Direções em torno de um pixel com distância 1*

```
função obterMatrizCooc(I, δ, θ, k)
  para cada pixel (x,y) da imagem I
    i <- obterValorPixel(x, y)
    j <- obterValorPixelVizinho(x, y, δ, θ)

    //construção da matriz de coocorrência
    M[i,j] <- M[i,j] + 1
    M[j,i] <- M[j,i] + 1
  fim para
fim
```

Algoritmo 2: *Cálculo da matriz de coocorrência em uma dada direção e distância*

Os cinco descritores de Haralick mais utilizados são: energia (En), contraste (Con), correlação (Cor), entropia (Ent) e homogeneidade (Hom). As definições desses descritores estão representadas nas Equações 2.3 à 2.7.

$$En = \sum_i \sum_j M_{ij}^2 \tag{2.3}$$

$$Con = \sum_{k=0}^{m-1} k^2 \sum_{|i-j|=k} M_{ij} \tag{2.4}$$

$$Cor = \frac{1}{\sigma_i \sigma_j} \sum_i \sum_j ij M_{ij} - \mu_i \mu_j \tag{2.5}$$

onde μ_i e σ_i são a média e variância horizontal e μ_j e σ_j a vertical.

$$Ent = \sum_i \sum_j M_{ij} \log(M_{ij}) \tag{2.6}$$

$$Hom = \sum_i \sum_j \frac{M_{ij}}{1 + (i-j)^2} \tag{2.7}$$

2.3.3 Transformada Discreta de Wavelets

Outra abordagem comum no processamento de imagens é a espectral, representada pelo uso de transformadas discretas. Esta abordagem tem como objetivo transformar uma imagem (interpretada como um sinal) em uma função de frequência e tempo, denominada espectro.

Uma função bastante utilizada para o processamento de sinais é a Transformada de Fourier, que representa um sinal por meio da soma de funções senoidais, sendo apropriada para sinais periódicos. A Transformada de Fourier e suas transformadas derivadas (por exemplo, a Transformada Discreta do Cosseno) conseguem detectar a variação de tons da imagem através da análise dos picos de frequência, porém não armazenam a informação sobre sua localização dentro do sinal. Devido a este fato, sinais não-periódicos (por exemplo, imagens) podem necessitar de uma quantidade de elementos maior que o próprio sinal original para representá-los, inviabilizando o uso das transformadas.

A Transformada Discreta de Wavelets (*Discrete Wavelet Transform* - DWT) busca contornar os problemas presentes na análise de sinais não-periódicos pela decomposição do sinal em diferentes escalas, utilizando bases de funções que são válidas apenas em determinados intervalos do sinal original. Esta abor-

dagem não só permite que suas variações sejam detectadas, mas também localizadas [DAUBECHIES, 1992].

O funcionamento da DWT consiste na aplicação sucessiva de filtros de passa-alta e passa-baixa, dividindo o sinal original em duas componentes menores de alta e baixa frequência. Em uma imagem, este filtro é aplicado de maneira iterativa entre as linhas e colunas, dividindo-a em quatro imagens de frequências diferentes a cada iteração, representadas na Figura 2.

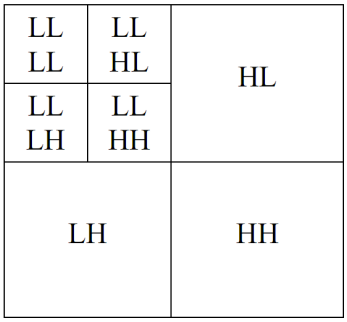

Figura 2: *Representação bidimensional dos filtros passa-baixa (L) e passa-alta (H)*

Existem várias transformadas de wavelets propostas na literatura científica. A DWT de Haar pode ser considerada a mais simples [HAAR, 1910]. Esta transformada divide o sinal de entrada em dois sinais de comprimento duas vezes menor que o original, de acordo com as operações de soma e diferença. O primeiro é resultante da soma de pares sucessivos do sinal original e representa a componente de baixa frequência. O segundo é obtido pela diferença entre os mesmos pares, representando a alta frequência. Repete-se esta sequência de operações para todas as linhas de uma imagem e em seguida para todas as suas colunas. As operações de soma e de diferença podem ser ponderadas, dando origem a operações de média e de diferença média.

Segundo [WANG, 1998], a DWT de Haar não apresenta as propriedades de transição acentuada e rápida atenuação do sinal, impedindo que a imagem seja separada em componentes claras e distintas. Parte dessa limitação é decorrente da simplicidade do filtro, por usar apenas dois pontos sucessivos. A DWT de Daubechies [DAUBECHIES, 1992] é uma extensão da DWT de Haar, buscando preservar uma quantidade maior de informações do sinal original com o uso de janelas maiores que se sobrepõem.

Dada uma janela de tamanho k, a DWT de Daubechies obtida a partir de uma imagem I de dimensões n x m é calculada seguido o Algoritmo 3.

Dada uma janela de tamanho quatro, as equações de soma e de diferença podem ser escritas como:

$$s = h_0 x_i + h_1 x_{i+1} + h_2 x_{i+2} + h_3 x_{i+3} \tag{2.8}$$

$$d = h_3 x_i - h_2 x_{i+1} + h_1 x_{i+2} - h_0 x_{i+3} \tag{2.9}$$

onde i é o índice do elemento cuja janela será calculada e h são os coeficientes da DWT de Daubechies 4-tap: $h_0 = 0.4830$; $h_1 = 0.8365$; $h_2 = 0.2241$; $h_3 = -0.1294$.

2.3.4 K-Vizinhos Próximos

Dentre todos os algoritmos utilizados na classificação e reconhecimento de padrões através de exemplos, o K-Vizinhos Próximos (*K-Nearest Neighbor* - KNN) está entre os mais simples. O funcionamento deste algoritmo está relacionado à expectativa de que elementos de uma mesma classe geralmente estejam mais próximos entre si, considerando uma função de distância e algum atributo para comparação.

```
função obterDWTDaubechies(I, n, m, k)
  //aplicação dos filtros nas linhas da imagem
  para y de 1 a m
    para x de 1 a n passo 2
      s <- somaJanelaHorizontal(I, x, y, k)
      d <- diferençaJanelaHorizontal(I, x, y, k)

      temp[x/2,y] <- s
      temp[n/2+x/2,y] <- d
    fim para
  fim para

  //aplicação dos filtros nas colunas do passo anterior
  para x de 1 a n
    para y de 1 a m passo 2
      s <- somaJanelaVertical(temp, x, y, k)
      d <- diferençaJanelaVertical(temp, x, y, k)

      saída[x,y/2] <- s
      saída[x, n/2+y/2] <- d
    fim para
  fim para
fim
```

Algoritmo 3: *Cálculo da DWT de Daubechies*

Dado um elemento de classe desconhecida, o algoritmo, na primeira tentativa, tenta encontrar os seus k vizinhos mais próximos e utiliza suas classes para determinar as classes candidatas. A quantidade de vizinhos pode ser qualquer número inteiro maior que um, porém é recomendável que seja um número impar para evitar situações de empate quando é considerado a classe majoritária entre as candidatas para determinar a classificação final.

A medida de similaridade entre o elemento desconhecido e um de seus vizinhos é utilizada como peso para a classe correspondente, e se um ou mais vizinhos compartilham a mesma classe, esta classe recebe uma ponderação maior [YANG; LIU, 1999]. Após obter todas as classes candidatas, estas são ordenadas de acordo com o peso recebido e uma é escolhida, usualmente a que atingiu a maior pontuação. A Figura 3 ilustra um possível espaço de classificação, partindo do círculo central e utilizando 3 e 5 vizinhos próximos.

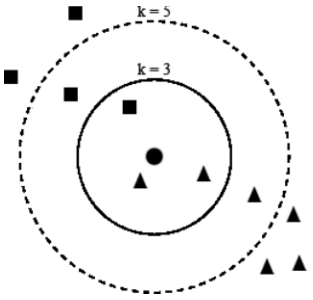

Figura 3: *Classificação via KNN com diferentes raios de abrangência*

A similaridade entre duas imagens pode ser calculada pela seguinte função de distância Minkowski:

$$d_{ab} = \sqrt[p]{\sum_{i=1}^{n} |a_i - b_i|^p} \tag{2.10}$$

onde a e b são os vetores de características das imagens, n o número de atributos e p a norma da distância (para p=2, tem-se a distância Euclidiana). Quanto menor a distância entre duas imagens, maior será a similaridade entre elas. Considerando a distância Euclidiana e uma ponderação baseada no número de classes encontradas, a classificação de um elemento t utilizando um conjunto de exemplos x considerando k vizinhos pode ser feita utilizando o Algoritmo 4.

```
função classificadorKNN(x, t, k)
  a <- vetorCaracterística(t)
  para cada exemplo i
    b <- vetorCaracterística(x)

    dist[i] <- 0
    para cada atributo j
      dist[i] <- dist[i] + potência(a[j] - b[j], 2)
    fim para

    dist[i] <- raizQuadrada(dist)
  fim para

  //retorna a classe mais frequente considerando as k menores distâncias
  classe <- classeMajoritária(dist, k)
fim
```

Algoritmo 4: *Classificador K-Vizinhos Próximos*

2.3.5 Redes Neurais Multicamada

As redes neurais multicamada (*Multi-Layer Perceptron* - MLP), são compostas por elementos de processamento interconectados chamados neurônios [HAYKIN, 2001] dispostos conforme a Figura 4. Classificadores baseados em MLP funcionam por meio do mapeamento de um conjunto de neurônios na camada de entrada (x_n) para um conjunto de neurônios na camada de saída (y_n), utilizando uma ou mais camadas ocultas de neurônios que se conectam por meio de pesos. Cada neurônio i de uma camada é ligado a um neurônio j da camada anterior por um peso w_{ij}, utilizado para representar a conexão entre eles.

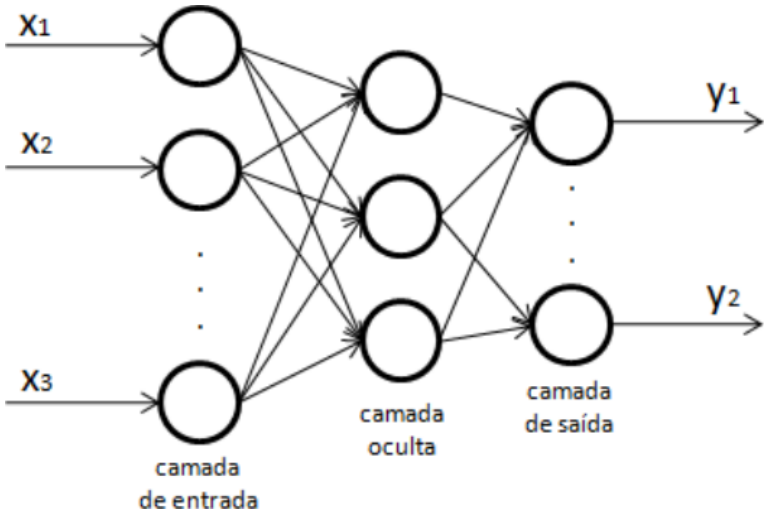

Figura 4: *Esquematização de uma Rede Neural Multicamada*

De maneira geral, a saída y_j de um neurônio j depende de suas conexões de entrada o_i, calculada

através de uma soma ponderada de peso w_{ij}. Após o cálculo da contribuição de cada conexão de entrada, o resultado final é obtido com o uso de uma função de ativação, geralmente uma sigmoidal, conforme descrito na Equação 2.11.

$$y_j = \varphi(u_j(n)) \tag{2.11}$$

onde u_j representa a somatória ponderada e $\varphi(x)$ a função de ativação logística, dadas por:

$$u_j(n) = \sum_i w_{ji} o_i \tag{2.12}$$

$$\varphi(x) = \frac{1}{1 + \exp^{-\alpha x}} \tag{2.13}$$

O treinamento de uma rede MLP é feito em duas etapas: *forward* e *backpropagation*. Na fase *forward*, os pesos w_{ij} são calculados partindo da camada de entrada, até chegar à camada de saída. A energia do erro é calculada no fim da fase *forward* de acordo com a Equação 2.14.

$$E(n) = \frac{1}{2} \sum_j (e_j(n))^2 \tag{2.14}$$

onde e_j (Equação 2.15) é o erro instantâneo dado pela diferença entre o valor desejado x_j e a saída y_j, representando a classe do exemplo fornecido na entrada e a saída do neurônio, respectivamente.

$$e_j = x_j(n) - y_j(n) \tag{2.15}$$

Caso a energia do erro seja maior que um valor pré-definido, todos os pesos são corrigidos na fase *backpropagation* de acordo com a seguinte regra:

$$\Delta w_{ji}(n) = -\eta \delta_j(n) o_i(n) \tag{2.16}$$

onde η é a taxa de aprendizado da rede e δ_j representa o gradiente local do erro. Se o neurônio está na camada de saída, este gradiente é calculado por:

$$\delta_j(n) = \varphi'(u_j(n)) e_j(n) \tag{2.17}$$

e se estiver em uma camada oculta:

$$\delta_j(n) = \varphi'(u_j(n)) \sum_k \delta_k(n) w_{kj}(n) \tag{2.18}$$

A fase de treinamento é finalizada e os pesos finais podem ser usados para calcular a saída de um exemplo de classe desconhecida, uma vez que o erro seja minimizado segundo algum critério (calculando seu erro médio quadrático, por exemplo) e atinja um nível aceitável. Simplificando as fases *forward* e *backpropagation* e considerando um conjunto de exemplos x, o treinamento de uma rede MLP com taxa de aprendizado η pode ser descrito segundo o Algoritmo 5.

```
função MLP(x, η)
  //inicializa os pesos com valores aleatórios
  para cada peso w[i,j]
    w[i,j] <- rand()
  fim para

  para cada exemplo i
    d[i] <- classe(x[i])
  fim para

  enquanto (critério de parada não satisfeito) faça
    //fase forward
    para cada camada k partindo da camada de entrada
      para cada neurônio j da camada k
        calcular saída y[j]
      fim para
    fim para

    //fase backpropagation
    para cada camada k partindo da camada de saída
      para cada neurônio j da camada k
        para cada neurônio i da camada k-1
          atualizar pesos w[j,i]
        fim para
      fim para
    fim para
  fim enquanto
fim
```

Algoritmo 5: *Treinamento de uma Rede Neural Multicamada*

2.4 Processamento de Linguagem Natural

Processamento de linguagem natural (*Natural language Processing* - NLP) é uma tentativa de extrair significado a partir de texto livre [KAO; POTEET, 2007]. Em NLP, é necessário reconhecer as variações linguísticas, que podem ser de ordem morfológica ou semântica, pela normalização da linguagem. Um dos grandes desafios de pesquisadores de NLP é manipular a ambiguidade da linguagem natural, uma vez que palavras e expressões podem possuir diferentes significados dependendo do contexto em que são empregados. As seções a seguir apresentam uma breve discussão sobre estes temas, sendo a normalização linguística discutida na Seção 2.4.1 e a redução de ambiguidade na Seção 2.4.2.

2.4.1 Normalização Linguística

Uma das tarefas mais comuns realizadas antes do processamento de informação textual é sua normalização. A normalização linguística tem como objetivo facilitar o reconhecimento das variações que ocorrem durante a escrita de textos em linguagem natural. Essas variações podem ser de origem morfológica ou semântica.

A variação morfológica ocorre quando várias palavras apresentam um radical em comum, seja pela variação de gênero, número e grau ou até mesmo em suas diferentes formas verbais. *Stemming*, ou radicalização, é uma técnica utilizada para extrair o radical de uma palavra por meio da remoção de sufixos, permitindo que palavras de mesmo significado sejam comparadas independente de sua variação de gênero,

número, grau e flexão. Em [VIERA; VIRGIL, 2007] são apresentados e comparados os principais algoritmos de radicalização para a língua portuguesa, destacando o algoritmo de Orengo [ORENGO, 2004] e a adaptação do algoritmo de Porter [PORTER, 1980].

A variação semântica é observada principalmente nos casos de sinonímia, onde palavras de formas diferentes podem representar o mesmo conceito ou ter um significado semelhante. Sua normalização é feita pela identificação de sinônimos em comum, criando agrupamentos de palavras de significados similares. Isto pode ser feito, por exemplo, através da busca de sinônimos em tesauros. Um tesauro pode ser visto como uma lista de palavras e seus sinônimos, dado um domínio específico.

2.4.2 Redução da Ambiguidade

Um dos maiores desafios na área de Processamento de Linguagem Natural é identificar e resolver problemas de ambiguidade presente nos textos. A ambiguidade surge quando uma expressão apresenta mais de uma interpretação possível, podendo ser principalmente de origem léxica ou estrutural [ULMANN, 1964].

Observa-se a presença da ambiguidade léxica principalmente nos casos de polissemia, onde uma mesma palavra pode ter significados diferentes dado o contexto em que se apresentam. Por exemplo, o termo "célula" pode significar "pequena cavidade", "unidade microscópica fundamental da matéria viva" ou até mesmo "agrupamento de pessoas para fins políticos". Uma abordagem semântica pode ser empregada para determinar o significado de uma palavra dentro do domínio, através do uso de ontologias de domínio.

A ambiguidade estrutural está relacionada às diferentes formas que as palavras podem ser organizadas para gerar uma interpretação, como no seguinte exemplo: "câncer de tireoide ausente". Nesta construção, existem duas interpretações possíveis: (1) Não há câncer de tireoide ou (2) o tipo de câncer é "tireoide ausente". Embora neste caso a primeira opção seja a mais provável de ser verdadeira, não há uma regra para determinar qual é correta, dado que a interpretação é sensível ao contexto. Uma forma de resolver este tipo de ambiguidade é atribuir pesos para cada combinação de palavras, que podem ser calculados segundo a frequência de cada uma na coleção de documentos. Neste cenário, os termos são agrupados através do uso de *n-grams*. Considerando o exemplo anterior, as representações *tri-gram* (onde o número de palavras é igual a três) seriam: "câncer de tireoide" e "de tireoide ausente". Desta forma, se o *tri-gram* "câncer de tireoide" ocorrer com mais frequência que "de tireoide ausente", esta estrutura é preferida em favor da outra.

2.5 Recuperação de Informação

Pesquisadores da área de Recuperação de Informação (*Information Retrieval* - IR) investigam maneiras de manipular a representação, o armazenamento, a organização e o acesso à informação [BAEZA-YATES; RIBEIRO-NETO, 1999], sendo o foco a informação que usuários precisam. Para buscar a informação desejada torna-se necessária a modelagem da coleção de documentos, onde cada documento é representado por um conjunto de palavras-chave representativas (termos de índice). Em geral, os pesquisadores de IR tendem a focar esforços na identificação de estruturas organizacionais, conteúdo e significado. Assim, sistemas tradicionais de IR manipulam dados desestruturados e são incapazes de suportar a manipulação de outros tipos de mídias (por ex. imagens). Este projeto difere do grupo de sistemas de IR tradicionais, uma vez que manipula imagens médicas e laudos radiológicos. Nesse sentido, é necessária uma abordagem que seja capaz de manipular laudos pelo uso de algoritmos de IR e imagens microscópicas por técnicas de CBIR.

2.5.1 Modelo Vetorial

O Modelo Vetorial [SALTON; LESK, 1968] é um modelo utilizado em Recuperação de Informação para representação de informação textual. Neste modelo, os termos de índice de um documento ou de uma expressão de busca são representados por pesos numéricos, ao contrário do Modelo Booleano que apenas registra a presença ou ausência dos termos. A escolha de pesos numéricos torna possível a recuperação parcial de documentos, permitindo o cálculo do grau de similaridade entre documentos e uma expressão de busca.

No modelo vetorial, o documento é representado por um vetor onde cada dimensão corresponde ao peso de um termo em relação ao documento. Caso o termo não ocorra no documento, seu peso atribuído é igual a zero. Existem diversas maneiras de calcular estes pesos, sendo a mais comum a medida TF-IDF (*Term Frequency-Inverse Document Frequency*). Dado uma coleção de D documentos e um termo t_k, a medida TF-IDF pode ser calculada segundo a Equação 2.19.

$$w_{ki} = \frac{n_{ki}}{N_i} log\left(\frac{|D|}{|d : t_k \in d|}\right) \tag{2.19}$$

onde n_{ki} é o número de vezes que o termo t_k aparece no documento d_i, N_i o número total de palavras neste documento, $|d : t_k \in d|$ o número de documentos que apresentam o termo t_k e $|D|$ o número total de documentos na coleção. Existem também outras equações na literatura para o cálculo de TF-IDF.

2.5.2 Similaridade por Cosseno

A similaridade entre documentos desestruturados geralmente é medida pelo grau de similaridade de seus conteúdos, partindo do princípio que documentos com termos em comum têm maior probabilidade de serem semelhantes.

Considerando-se o espaço vetorial gerado pelo Modelo Vetorial, cujo perfil de cada documento é definido como sendo o vetor de pesos dos termos que o compõe, a similaridade entre dois documentos pode ser medida através do cosseno do ângulo formado por seus vetores, dada a equação 2.20.

$$cos\,(d_i, d_j) = \frac{\sum\limits_k w_{ki} w_{kj}}{\sqrt{\sum\limits_k w_{ki}^2}\sqrt{\sum\limits_k w_{kj}^2}} \tag{2.20}$$

onde k é o número de termos presentes na coleção de documentos e w_{ki} e w_{kj} representam a importância do termo t_k para os documentos d_i e d_j, respectivamente, utilizando a medida TF-IDF.

O resultado da Equação 2.20 pode ser interpretado como sendo o cosseno do ângulo formado pelos dois vetores de pesos. Devido ao fato dos pesos calculados segundo a medida TF-IDF serem sempre positivos, o cosseno resultante estará entre 0 e 1. Um cosseno próximo a zero indica que há poucos termos comuns entre os documentos, à medida que quanto mais próximo de 1, maior será a similaridade.

2.6 Trabalhos Relacionados

Foram identificadas na literatura tentativas de construção de ferramentas de apoio ao diagnóstico em câncer de tireoide com o objetivo de classificar nódulos benignos e malignos. As abordagens mais utilizadas são a análise morfométrica e de textura do núcleo das células [DASKALAKIS, 2008; HARMS, 2002; RAJESH, 2004]. Em [GUPTA, 2001] é realizada uma avaliação da eficiência diagnóstica de algoritmos baseados em parâmetros quantitativos de núcleos celulares na discriminação da malignidade em lesões tireoidianas. Os

resultados mostraram que os parâmetros nucleares podem ser úteis na discriminação entre as variantes malignas e benignas de lesões do tipo papilífero.

Em [WANG, 2010], é descrito um método para detecção e classificação de adenoma folicular de tireoide, carcinoma folicular e tecido tireoidiano normal. Essa proposta também utiliza características morfométricas e de textura dos núcleos da célula para a extração de características, empregando como classificador o SVM (*Support vector machine*). Uma das conclusões obtidas pelos autores foi que as características nucleares foram suficientes para classificar automaticamente entre os tipos de lesões do tecido tireoidiano.

Já em [CHEN, 2008], a imagem microscópica de tireoide é analisada como um todo e classificada entre diferentes tipos de tecido, incluindo coloide, estroma, células foliculares e célula de carcinoma papilífero. A classificação foi realizada considerando os atributos presentes no tecido como, por exemplo, brilho, entropia e energia. Verificou-se que nenhum atributo sozinho foi capaz de diferenciar um único tipo de tecido dos demais. No entanto, quando combinados, os atributos utilizados no trabalho apresentavam capacidade para diferenciar os tipos de tecido. Alguns destes atributos são também empregados no presente trabalho.

Entre os algoritmos de classificação, [DASKALAKIS, 2008] explorou uma abordagem de múltiplos classificadores baseado em k-vizinhos próximos, redes bayesianas e redes neurais. Em [RAJESH, 2004], testes estatísticos foram aplicados para identificar variáveis que diferenciassem as características morfométricas dos núcleos de células da tireoide e em [HARMS, 2002] a classificação é feita por meio de árvores de decisão.

Em [NÉVÉOL, 2009] é proposto um método para recuperação de documentos médicos baseado na anotação automática de imagens. Esta anotação automática é realizada pelo voto majoritário de códigos obtidos pela classificação de imagens médicas de textos acadêmicos e o processamento da informação textual de suas respectivas legendas e parágrafos de referência, envolvendo técnicas de NLP e CBIR. Os resultados obtidos indicam que a análise multimodal proposta combinando imagens e informação textual parece ser mais benéfica para a indexação do que a recuperação de documentos médicos. Na tarefa de recuperação de imagens a análise multimodal obteve resultados muito próximos da análise de imagens, indicando que a contribuição da análise de informação textual para a recuperação de imagens pode ter sido pequena.

O presente trabalho difere dos demais por propor um método de análise de imagens que combina características de textura extraídas das imagens, dados morfométricos nucleares e informação textual obtida em laudos. Inicialmente, técnicas de CBIR e Morfometria são utilizadas para a identificação dos componentes como, por exemplo, tipos celulares específicos em uma imagem não diagnosticada, com o objetivo de gerar uma descrição microscópica da imagem. Esta descrição automatizada é utilizada na localização de laudos médicos similares, através dos quais extraem-se informações complementares de modo a auxiliar o patologista no diagnóstico da imagem de interesse.

Capítulo 3

O *Framework* FREDS

O *framework* FREDS (*Framework* para Redução da Descontinuidade Semântica em Imagens Médicas) foi construído com o objetivo de suportar aplicações que necessitem de acesso à componentes de Processamento de Linguagem Natural, Recuperação de Informação e Recuperação de Imagens Baseada em Conteúdo para investigar o mapeamento do conteúdo de imagens médicas e informações textuais em exames de tireoide. Este projeto é uma continuação do *framework* desenvolvido pelo autor em [PESSOTTI, 2008] para análise de imagens radiológicas de tomografia computadorizada.

Para determinar as necessidades e compreender as funcionalidades da proposta desta dissertação, foram conduzidas atividades de análise de requisitos e desenvolvimento de um diagrama de classes utilizando a linguagem UML. Os principais requisitos funcionais do FREDS são: (i) permitir que o patologista submeta imagens microscópicas obtidas em microscópios com câmeras digitais; (ii) identificar regiões de interesse em uma imagem médica; (iii) calcular e apresentar medidas morfométricas nucleares; (iv) retornar laudos cujas descrições sejam similares a expressões de busca obtidas a partir dos rótulos de uma imagem; (v) obter e agrupar os diagnósticos de um conjunto de laudos e (vi) obter os termos mais frequentes de um conjunto de diagnósticos.

O *framework* possui basicamente quatro interfaces de software: Segmentador de Objetos, Extrator de Características, Classificador de Padrões e Recuperador de Informação. Essas interfaces foram definidas com o objetivo de prover um meio de acesso comum às classes do sistema, permitindo que novos componentes sejam implementados futuramente sem a necessidade de readequar os componentes que os utilizam. Para cada interface tem-se as seguintes funcionalidades:

- Segmentador de Objetos: define componentes responsáveis pela segmentação e rotulação de imagens.
- Extrator de Características: define componentes responsáveis pela extração de características de imagens.
- Classificador de Padrões: define componentes responsáveis pela classificação de padrões.
- Recuperador de Informação: define componentes responsáveis pelo processamento de linguagem natural e recuperação de informação.

A interface Segmentador de Objetos é implementada pela classe Threshold, que conduz a segmentação de uma imagem pela técnica de limiarização de Otsu. Seus principais métodos são: (i) segmentarImagem, que obtém a máscara da binária de uma imagem e (ii) extrairObjetos, que identifica os núcleos celulares. Os detalhes do funcionamento da segmentação foram apresentados na Seção 2.3.1.

As interfaces Extrator de Características e Classificador de Padrões são uma adaptação do trabalho desenvolvido pelo autor em [PESSOTTI, 2008]. Nesse novo cenário, as classes foram adaptadas para a leitura das imagens de microscópio e um novo extrator baseado em Morfometria foi desenvolvido. A interface Extrator de Características é implementada pelas classes de Matriz de Coocorrência, Wavelets e Morfometria e seu principal método é gerarVetorCaracterísticas, que obtém a assinatura de uma imagem por meio do cálculo de métricas. Já a interface Classificador de Padrões é implementada pelas classes K-Vizinhos Próximos e Redes Neurais, tendo como principais métodos treinarClassificador, validarClassificador e obterClasse, que realizam as tarefas de treinamento, validação e classificação de padrões, respectivamente.

A interface Recuperador de Informação é implementada pela classe Analisador de Documento e seus principais métodos são (i) obterTermosMaisFrequentes, que retorna os termos de maior frequência em um conjunto de documentos e (ii) buscaSimilaridade, que a partir de uma expressão de busca (*query*) retorna os documentos que apresentam maior similaridade. A classe Analisador de Documento apresenta duas classes auxiliares: Documento, que armazena o conteúdo dos laudos, seus termos e *scores* e Processador de String, que contém métodos de processamento textual

Além destas interfaces, o *framework* também apresenta a classe de software Visualizador que, apoiado pelo pacote de desenvolvimento ImageJ, é responsável pela exibição das imagens e fornece métodos de manipulação como, por exemplo, obter pixeis específicos de uma imagem e alteração de brilho/contraste para facilitar sua visualização. Essas alterações são válidas apenas durante a exibição das imagens, sendo desprezadas nas tarefas de segmentação de objetos e extração de características para evitar a variância entre as imagens. Finalmente, todas interfaces são controladas pela classe controladora do sistema, chamada FREDS, responsável pela integração dos componentes de software e chamada dos métodos para a execução dos experimentos.

Este trabalho explora a característica de modularidade do *framework* inicial de [PESSOTTI, 2008] para modificar e estender seus componentes de modo a poder aplicá-lo à proposta desta monografia. A Figura 5 apresenta as classes de software do FREDS. Os componentes sombreados no diagrama foram adaptados de [PESSOTTI, 2008] para o cenário propostao. Os demais componentes foram desenvolvidos especificamente para este trabalho.

A linguagem de programação escolhida para desenvolver o FREDS foi o Java, por suportar o rápido desenvolvimento por meio do reuso de componentes, a existência de APIs bastante confiáveis e a sua característica multi-plataforma, tornando-o possível de ser executado em diferentes sistemas operacionais [GOSLING; MCGILTON, 1995]. Para o acesso ao conteúdo das imagens, foram definidos plugins de processamento de imagens, codificados utilizando bibliotecas fornecidas pelo software livre ImageJ, construído em Java e desenvolvido pelo National Institutes of Health (NIH) [RASBAND, 2008]. Os classificadores de padrões foram instanciados a partir de componentes presentes no software livre Weka, um ambiente em Java composto por vários algoritmos de aprendizado de máquina [HOLMES, 1994].

O FREDS utiliza um banco de dados relacional para o armazenamento das imagens e laudos necessários na execução do sistema. O Sistema de Gerenciamento de Banco de Dados (SGBD) escolhido foi o SQLite [HIPP; KENNEDY,], por ser um SGBD relacional que não precisa de um servidor para ser acessado. O acesso ao banco é realizado por meio de uma biblioteca SQLite multi-plataforma e as tabelas, bem como seus dados, são armazenados em um único arquivo, facilitando o acesso aos dados. Por ser escrito em Java e utilizar o SGBD SQLite, o FREDS apresenta um grau elevado de independência do sistema operacional utilizado pelo usuário, podendo ser executado em ambientes Windows, Linux e Mac OS X sem a necessidade de versões dedicadas.

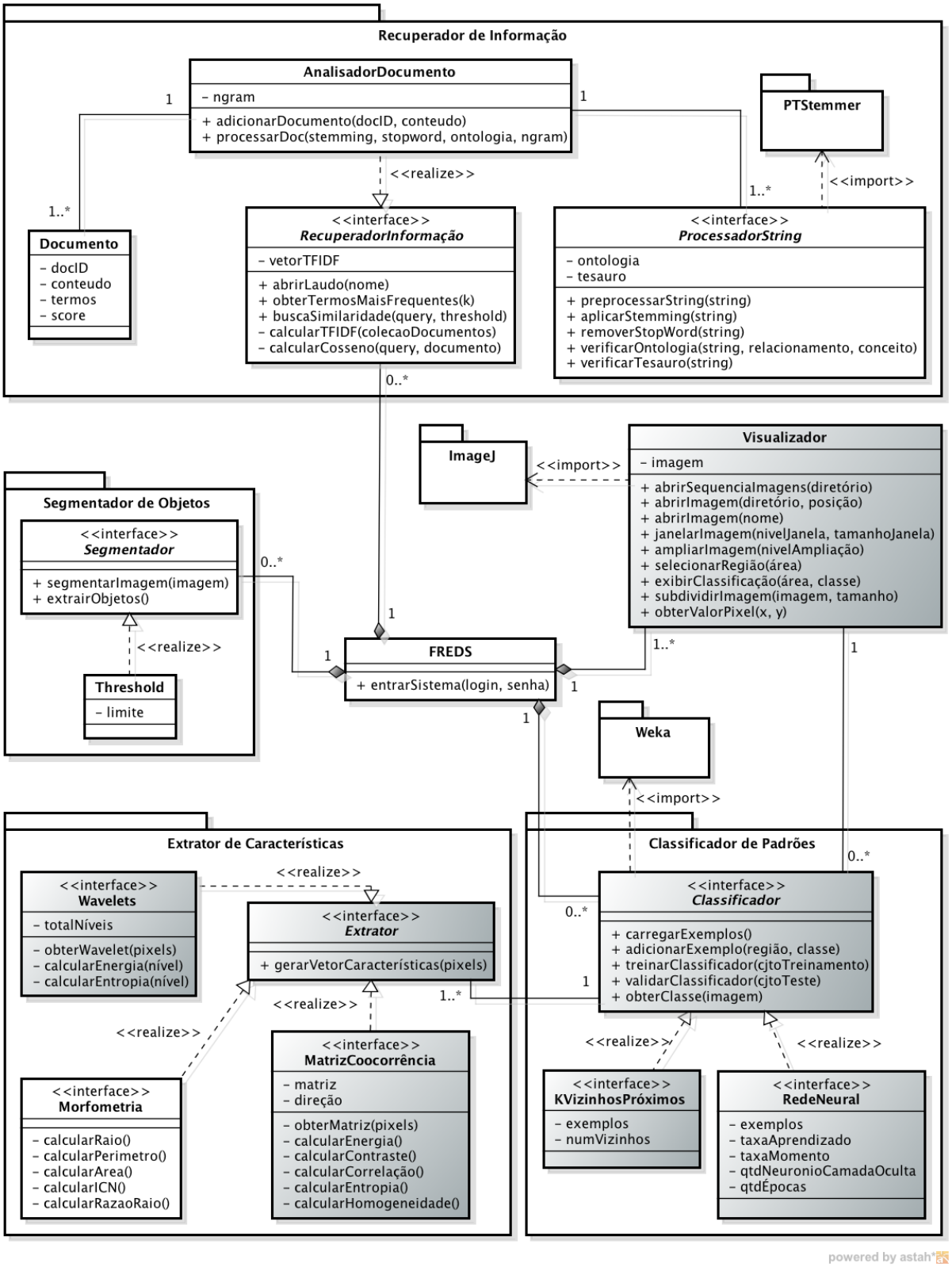

Figura 5: *Classes de Software do FREDS. Em cinza, os componentes propostos originalmente em [PESSOTTI, 2008] e em branco os desenvolvidos neste trabalho.*

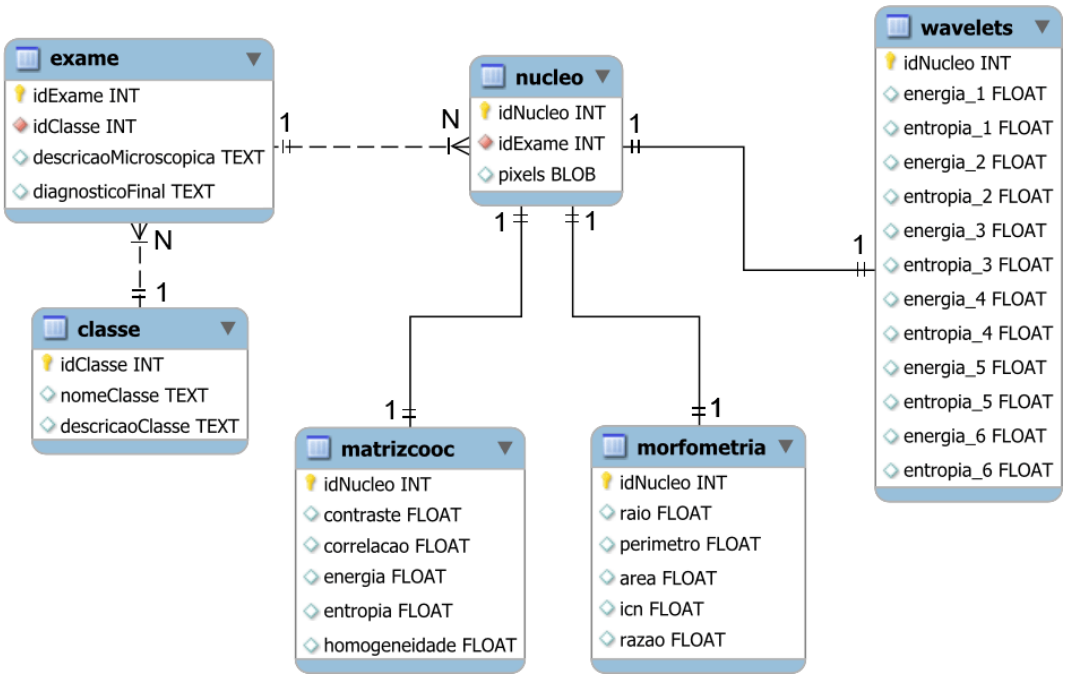

Figura 6: *Diagrama Entidade Relacionamento que representa o banco de dados construído para o FREDS.*

A Figura 6 apresenta o Diagrama Entidade Relacionamento do banco de dados construído, contendo tabelas e relacionamentos definidos para o armazenamento de exames, núcleos celulares segmentados e métricas extraídas pelo FREDS. A tabela exame contém a descrição microscópica e o diagnóstico final presentes no laudo, bem como sua classificação dentre o tipo de lesão tireoidiana. A tabela classe contém as possíveis classificações de exames. A tabela nucleo armazena os pixels de cada núcleo extraído das imagens que compões um exame. As tabelas matrizcooc, wavelets e morfometria armazenam os valores de características e métricas calculadas a partir da análise de um dado núcleo.

A experiência de desenvolvimento do *framework* FREDS foi descrita em artigo no XI Workshop de Informática Médica (WIM), um evento realizado dentro do XXXI Congresso da Sociedade Brasileira de Computação (CSBC) [PESSOTTI, 2011]. O autor está preparando um novo artigo descrevendo o FREDS com suas classes e métodos para a realização do mapeamento conceitual entre conteúdo baseado em análise de imagens e informações morfométricas. Este trabalho será em breve enviado para evento da área.

Respectivamente, nas Seções 3.1 e 3.2, os dois principais processos[1] do FREDS são detalhados: a Rotulação de Imagens e a Seleção de Laudos. A Seção 3.3 apresenta interfaces gráficas construídas para a utilização do FREDS e seus processos.

3.1 Rotulação de Imagens

Rotulação de imagens é o processo de atribuição de valores numéricos para regiões conexas de uma imagem que compartilham uma determinada característica de interesse [ROSENFELD; PFALTZ, 1966]. Este processo consiste na identificação de regiões conexas, por exemplo, por meio da segmentação de componentes de uma imagem. Durante a rotulação, todos os pixeis pertencentes a uma mesma região recebem o mesmo valor numérico, chamado de rótulo, que é atribuído sequencialmente conforme as regiões são identificadas.

A rotulação é necessária para identificar de forma única os componentes de uma imagem microscópica e pode ser conduzida de duas maneiras diferentes: (a) rotulação automatizada com algoritmos de processa-

[1] Processo é um conjunto de passos ou técnicas computacionais para a execução de uma determinada tarefa

mento de imagens para segmentar e identificar os rótulos e (b) rotulação auxiliada por especialista, que é responsável por identificar os achados patológicos e posteriormente rotulá-los.

3.1.1 Rotulação Automatizada

A rotulação automatizada é realizada pelo uso de técnicas de Processamento de Imagens e Recuperação de Imagens Baseada em Conteúdo. Inicialmente deve ser criada uma base de exemplos de achados patológicos, estruturada com o auxílio de um especialista, contendo exemplos de regiões de interesse para a definição de diagnóstico em câncer de tireoide.

A rotulação automatizada pode ser conduzida pela segmentação de uma imagem microscópica por métodos da interface Segmentador de Objetos do *framework* FREDS. Cada componente segmentado é submetido à extração de atributos e posteriormente é realizada sua classificação por similaridade, sendo atribuído um rótulo de acordo com sua similaridade com os achados patológicos armazenados na base de exemplos. Essas tarefas são executadas por métodos das interfaces Extrator de Características e Classificador de Padrões do *framework* FREDS. O processo de rotulação automatizada do FREDS é apresentado no Algoritmo 6. A base de achados patológicos foi construída e é apresentada na Subseção 4.1 desta dissertação.

```
função rotularImagem(I)
  nroRótulo <- 0
  rótulos <- vetor[]
  componentes <- segmentarImagem(I)

  para cada componente da imagem I
    nroRótulo <- nroRótulo + 1
    áreaSegmentada <- matriz[]

    para cada pixel (x,y) do componente
      rótulosImagem[x,y] <- nroRótulo
      áreaSegmentada[x,y] <- I[x,y]
    fim para

    vetorAtributos <- extraçãoAtributos(áreaSegmentada)
    classe <- classificarSimilaridade(vetorAtributos)

    rótulos[nroRótulo] <- obterRótulo(classe)
  fim para
fim
```

Algoritmo 6: *Rotulação automatizada de imagens*

3.1.2 Rotulação Auxiliada pelo Usuário

Na rotulação auxiliada pelo usuário, o patologista é o responsável pela identificação das estruturas e achados patológicos presentes na imagem microscópica. Inicialmente no FREDS, a imagem é segmentada por métodos da interface Segmentador de Objetos do *framework* e cada componente extraído é apresentado ao patologista, que deve classificá-lo entre os possíveis achados patológicos e atribuir manualmente o rótulo, de acordo com o Algoritmo 7.

A rotulação auxiliada é útil nas situações em que o patologista consegue identificar os achados patológicos em uma imagem, mas apresenta dificuldade em formular seu diagnóstico. Esta abordagem também pode servir como *feedback* para a rotulação automática, permitindo ao patologista corrigir eventuais erros

```
função rotularImagemAuxilioUsuario(I)
  nroRótulo <- 0
  rótulos <- vetor[]
  componentes <- segmentarImagem(I)

  para cada componente da imagem I
    nroRótulo <- nroRótulo + 1
    áreaSegmentada <- matriz[]

    para cada pixel (x,y) do componente
      rótulosImagem[x,y] <- nroRótulo
    fim para

    exibirRegiãoImagem(áreaSegmentada)
    classe <- classificaçãoPatologista(áreaSegmentada)

    rótulos[nroRótulo] <- obterRótulo(classe)
  fim para
fim
```

Algoritmo 7: *Rotulação auxiliada de imagens*

durante a classificação e adicionar rótulos que não foram identificados de forma automatizada, podendo levar a um aumento na taxa de precisão da rotulação.

3.2 Seleção de Laudos

A fim de auxiliar o patologista na formulação do diagnóstico, o *framework* FREDS deve ser capaz de identificar laudos semelhantes à imagem sendo visualizada pelo patologista para fornecer uma segunda opinião. Esta tarefa é realizada pela interface Recuperador de Informação. Neste contexto, a seleção de laudos é responsável pela identificação de documentos presentes na coleção de laudos que contém em sua descrição microscópica achados patológicos semelhantes à aqueles presentes na imagem sendo visualizada pelo patologista (Figura 7).

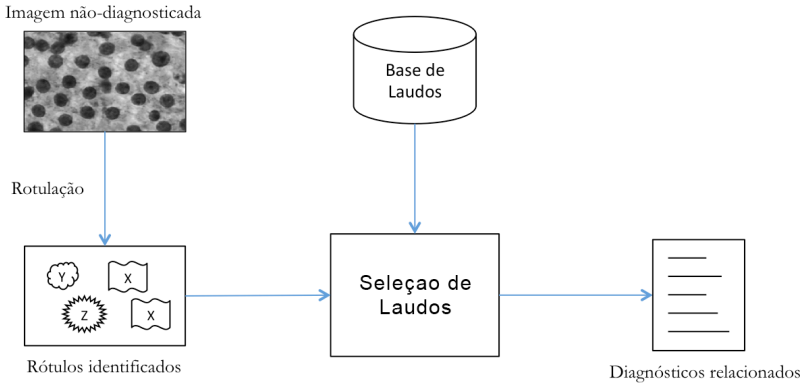

Figura 7: *Fluxograma da experimentação da coleção de laudos*

De acordo com as Figuras 7 e 8, uma vez identificados os rótulos de uma imagem por algum dos métodos descritos na Seção 3.1, estes rótulos são utilizados na formação de expressões de busca (*queries*) para a etapa de seleção de laudos patológicos. Para verificar se os rótulos identificados estão dentro do contexto de câncer de tireoide, foi utilizada uma ontologia de domínio definida para este projeto junto ao especialista. No estágio atual, esta ontologia contém termos associados a achados patológicos e diferentes tipos de câncer de tireoide. Conforme a necessidade, em trabalhos futuros, essa ontologia poderá ser expandida.

Embora existam ontologias de carcinomas que contenham relacionamentos do tipo “achado patológico” [GOLBECK, 2003], não foi encontrada nenhuma especificamente dentro do domínio de tireoide para a língua portuguesa.

A formação das *queries* foi iniciada pela expansão dos termos pelo uso de um tesauro, seguido da combinação desses termos expandidos por conectivos lógicos AND e OR. Estas *queries* foram utilizadas na seleção de laudos patológicos por meio do cálculo de similaridade do cosseno entre a descrição microscópica presente nos laudos e cada combinação formada. Uma visão geral deste processo de busca e seleção de laudos pode ser visualizada na Figura 8.

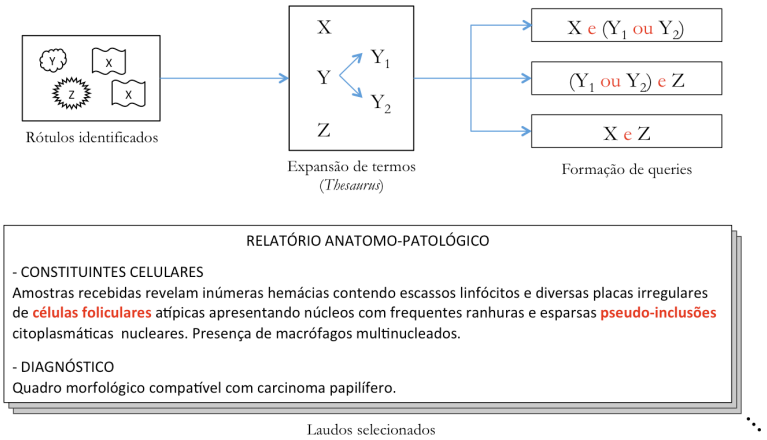

Figura 8: *Seleção de Laudos para Recuperação de Informação*

O resultado da seleção de laudos pode ser apresentado ao patologista em uma das seguintes maneiras: (1) listagem dos termos mais relevantes encontrados nos laudos, utilizando o peso de cada termo como medida de *score* e (2) agrupamento dos diagnósticos presentes nos laudos, utilizando a similaridade cosseno como medida de *score*. Acredita-se que a listagem de termos relevantes possa fornecer auxílio ao patologista durante a escrita do seu diagnóstico final. Já o agrupamento de diagnósticos é mais indicado para o patologista comparar seu diagnóstico com diagnósticos similares presentes na base de laudos.

3.3 Validação do FREDS

Para a validação do *framework* FREDS foi construído um sistema de apoio ao diagnóstico médico. O sistema é composto por um conjunto de interfaces gráficas. A Figura 9 apresenta a tela inicial do sistema FREDS. A tela inicial é composta por um botão para abrir uma imagem que deseja-se classificar; um botão para acessar as opções do sistema; um controle para ajuste fino da segmentação, caso o usuário não esteja satisfeito com a segmentação automática e, por fim, um botão para classificar e retornar os termos mais frequentes associados à imagem aberta.

A Figura 10 apresenta a tela de opções do FREDS, onde o usuário pode selecionar quais algoritmos serão usados nas funcionalidades internas do sistema referentes à extração de características, classificação de padrões e processamento textual. O sistema foi pré-configurado para os parâmetros que atingiram a maior taxa de acertos durante a fase de experimentação, que serão apresentados na Seção 4.2.

A Figura 11 apresenta uma demonstração da execução do sistema. A imagem microscópica em questão foi obtida de um exame de Bócio e submetida ao sistema, que classificou os núcleos celulares de acordo com sua similaridade em relação à exames armazenados na base de dados definida na seção. Os núcleos destacados em azul representam Bócio, em verde, Câncer Papilífero e, em vermelho, Indeterminado. É possível também observar os termos mais frequentes encontrados em imagens similares ao exame submetido,

Figura 9: *Tela inicial do sistema construído*

Figura 10: *Tela de opções do sistema construído*

que indicam que a imagem foi classificada como sendo Bócio.

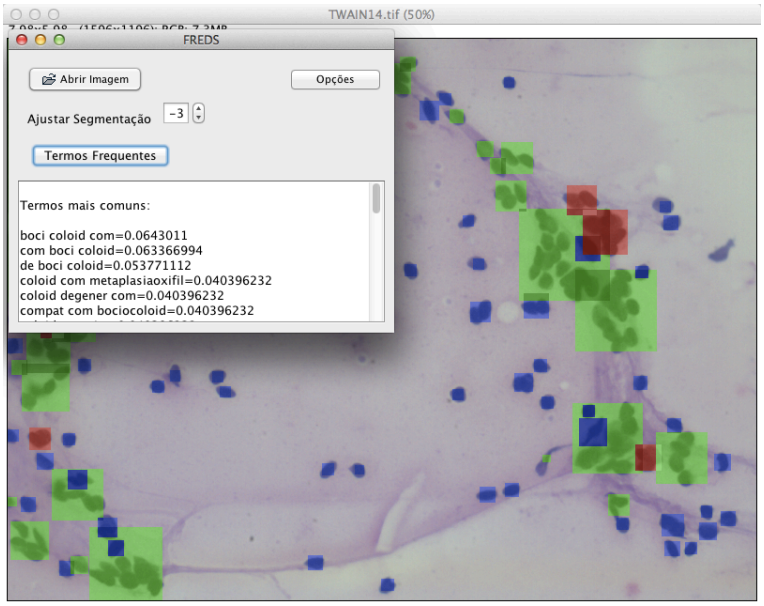

Figura 11: *Tela demonstrando a execução do sistema*

Capítulo 4

Experimentação e Resultados

Este capítulo apresenta a coleção de imagens usadas em experimentos e os resultados obtidos para o uso do *framework* FREDS na coleção de imagens e laudos. Os resultados são descritos em duas seções, sendo a primeira utilizando a coleção de Imagens e a segunda, o componente de Seleção de Laudos.

4.1 Coleção de Imagens e Laudos

A coleção de imagens utilizada na experimentação do FREDS e seus componentes é composta pela combinação de lâminas de Punção Aspirativa por Agulha Fina (PAAF) de tecido tireoidiano de exames realizados no Serviço de Patologia do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP-USP) e uma coleção de exames de PAAF cedidos pela Profa. Dra. Edna Teruko Kimura. As imagens de exames realizados no HCFMRP-USP foram obtidas com permissão e apoio do Prof. Dr. Edson Garcia Soares da Faculdade de Medicina de Ribeirão Preto (FMRP-USP). O projeto tem a aprovação dos Comitês de Ética em Pesquisa da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP-USP) e da FMRP-USP para a publicação dos resultados (Processo SISNEP CAAE-0027.0.222.000-11).

A construção da coleção de imagens e laudos foi realizada em quatro etapas:

1. Levantamento de Exames. Inicialmente, foi feito o levantamento de lâminas de exames de tireoide realizados no Serviço de Patologia do HCFMRP-USP, das quais são obtidas as imagens digitalizadas e o respectivo laudo patológico.

2. Escolha de exames relevantes. Com a ajuda do Prof. Dr. Edson Garcia Soares, foram escolhidas as imagens mais relevantes de cada exame dado o diagnóstico em questão. Considerada-se relevantes as imagens que, segundo o patologista, apresentam todas características de um determinado tipo de lesão.

3. Identificação dos padrões. Após a escolha das imagens mais relevantes, foi solicitado ao patologista que identificasse os padrões presentes em cada imagem. Estes padrões são utilizados no treinamento dos algoritmos do FREDS.

4. Criação da base. A partir das regiões identificadas na etapa anterior obtém-se os limites da região marcada pelo patologista, os valores de pixel em seu interior e o laudo associado ao exame, armazenados posteriormente na base de dados descrita no Capítulo 3.

A coleção de exames cedida pela Profa. Dra. Edna Teruko Kimura foi incorporada posteriormente na base de dados, passando somente pela etapa 4 descrita anteriormente. Isso foi possível devido aos exames já terem sido selecionados e identificados por ela.

As sessões de captura de imagens foram conduzidas no Laboratório de Neurologia Aplicada e Experimental da FMRP-USP, utilizando uma câmera digital Zeiss Axiocam MRc acoplada a um microscópio Zeiss Axiophot com um aumento total de 480x (objetivas de 40x, *optovar* de 1.6x e fator da câmera de 7.5x).

Entre as diferentes sessões de captura houve a preocupação de manter constante o nível de luz do microscópio, o fator de aumento e os parâmetros de captura da câmera, para eliminar possíveis fontes de variância entre as imagens. No entanto, fatores ligados à construção da lâmina não podem ser controlados, pois as lâminas são obtidas por meio de um estudo retrospectivo e não contém a identificação do paciente, o que impossibilita a localização do bloco de parafina para a construção de novas lâminas. Entre os fatores que não podem ser controlados, pode-se citar a espessura do corte do material biológico, a proporção de corantes e reagentes utilizados,

A coleção de imagens estava acompanhada de laudos contendo descrição microscópica e diagnóstico, ambos redigidos pelo patologista. Cada laudo é fundamentado na observação de uma ou mais imagens patológicas. No laudo, a descrição contém uma visão geral da lâmina em termos de achados patológicos e organização celular. O diagnóstico contém a conclusão do laudo, sendo formulado pelo patologista a partir da análise dos achados da lâmina. Os achados patológicos são os objetos visíveis que o patologista julga importantes como, por exemplo, alterações nucleares que possam caracterizar uma lesão.

A coleção de imagens obtida para fim de experimentação do FREDS é constituída de 4 casos de carcinoma papilífero, totalizando 19 imagens capturadas, 4 casos de bócio (19 imagens) e 2 indeterminados (17 imagens). Considerando todas as imagens, foi encontrado um conjunto de 1952 núcleos celulares segmentados, divididos entre 922 núcleos de bócio, 300 indeterminaos e 730 de câncer papilífero. Já a coleção de laudos é composta por 46 laudos patológicos, totalizando 1939 palavras ligadas à descrição microscópica e 690 palavras presentes nos diagnósticos.

A Figura 12 apresenta uma imagem de uma lâmina de tecido tireoidiano corado em HE (Hematoxilina-eosina) e o Quadro 1 apresenta um exemplo de descrição microscópica e diagnóstico de carcinoma papilífero.

RELATÓRIO ANATOMO-PATOLÓGICO
- CONSTITUINTES CELULARES Amostras recebidas revelam diversas placas irregulares de células foliculares atípicas apresentando núcleos com frequentes ranhuras e esparsas pseudo-inclusões citoplasmáticas nucleares.
- DIAGNÓSTICO Quadro morfológico compatível com carcinoma papilífero.

Quadro 1: *Exemplo de laudo patológico com a descrição microscópica e o diagnóstico final.*

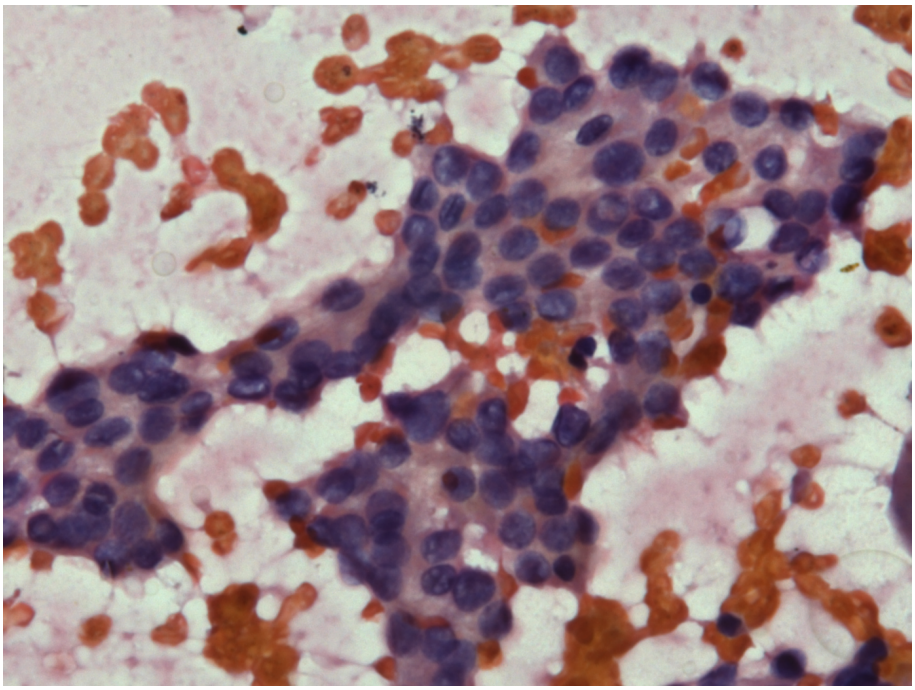

Figura 12: *Exemplo de uma imagem típica de carcinoma papilífero.*

4.2 Experimentação e Resultados

Para avaliar a taxa de acertos (verdadeiros positivos) dos algoritmos de classificação do FREDS foram definidos três cenários com a combinação dos algoritmos de extração de características. O primeiro cenário considera os algoritmos de extração aplicados separadamente (COOC, DWT, MORFO), o segundo considera a combinação dois a dois dos extratores (COOC+DWT, COOC+MORFO, DWT+MORFO) e o terceiro inclui a combinação de todos (COOC+DWT+MORFO), onde COOC é o extrator Matriz de Coocorrência; DWT, Transformada Discreta de Wavelets e MORFO, gerado a partir da análise morfométrica dos núcleos celulares.

A classificação de padrões foi realizada considerando dois algoritmos: K-Vizinhos Próximos (KNN) e Redes Neurais Multicamada (MLP). A definição dos parâmetros dos classificadores foi feito por teste exaustivo de parâmetros que propiciaram a maior taxa de acertos dado as classes Bócio, Indeterminado e Papilífero.

A fim de investigar a influência dos parâmetros dos algoritmos na classificação, foram conduzidos experimentos variando o tamanho da vizinhança para o KNN. De acordo com a Figura 13, observa-se que a taxa de acertos atinge o ponto máximo (84.4%) para K=9. Para vizinhanças maior que K=9, a taxa de acertos decresce conforme a vizinhança aumenta.

Para o classificador MLP, foram verificados os parâmetros de taxa de aprendizado e de taxa de momento para 1000 gerações. Como observado na Tabela 1 e na Figura 14, as melhores taxas de acerto foram obtidas por taxas de aprendizado em torno de 0.05. A melhor taxa de acerto obtida foi 85.71%, utilizando taxa de aprendizado 0.05 e taxa de momento 0.25.

No primeiro cenário, apresentado nas três primeiras colunas das Tabelas 2(a) e 2(b) é possível observar que a Matriz de Coocorrência obteve a maior taxa de acertos entre os algoritmos de extração. Não houve grande diferença entre os algoritmos de classificação KNN e MLP para este cenário. A baixa taxa de acertos dos algoritmos DWT e MORFO na maioria dos casos se deu principalmente pela classe Indeterminado. Por fim, a melhor taxa registrada foi de 81.2% com o uso da Matriz de Coocorrência e o classificador KNN.

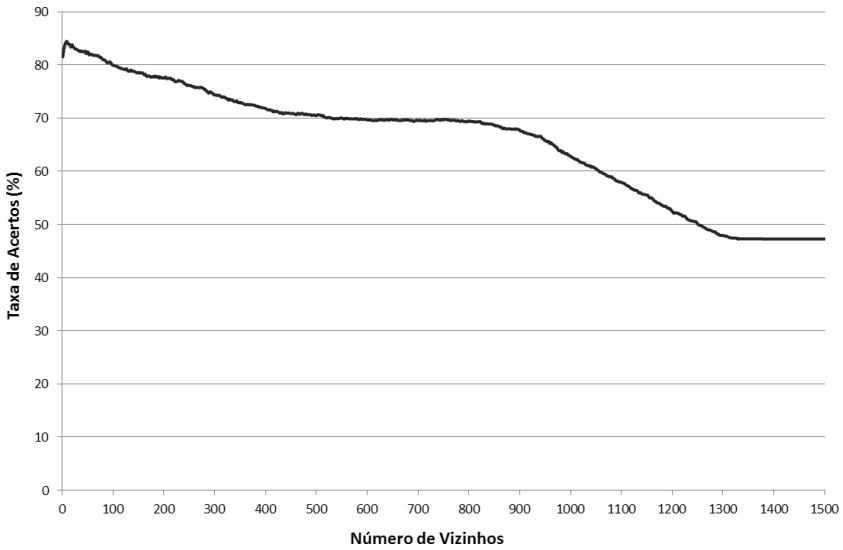

Figura 13: *Taxa de acertos em função do tamanho da vizinhança para o classificador KNN.*

		Taxa de Momento										
		0.00	0.05	0.10	0.15	0.20	**0.25**	0.30	0.35	0.40	0.45	0.50
Taxa de Aprendizado	0.00	84.68	85.04	84.63	84.63	84.27	84.43	84.89	84.73	84.58	84.12	84.43
	0.05	85.50	85.55	85.55	85.40	85.35	**85.71**	85.55	85.19	85.50	85.45	85.56
	0.10	84.94	84.99	85.45	85.55	85.40	85.45	85.09	84.99	84.63	84.78	84.68
	0.15	85.30	85.09	84.43	85.35	85.40	84.48	84.43	84.94	83.91	84.43	84.12
	0.20	84.63	84.22	84.32	84.12	84.22	84.78	84.27	84.07	84.73	84.89	84.07
	0.25	84.68	85.04	84.22	84.27	84.43	84.27	84.73	84.68	84.07	84.78	83.86
	0.30	84.68	85.04	84.63	84.63	84.27	84.43	84.89	84.73	84.58	84.12	84.43
	0.35	84.32	84.89	84.99	84.27	84.58	83.76	83.71	84.27	84.27	84.58	83.81
	0.40	84.78	84.94	84.89	83.76	84.27	84.63	84.78	84.89	84.32	84.02	84.07
	0.45	84.07	84.58	84.02	84.48	84.22	85.04	83.86	84.99	84.89	83.76	83.81
	0.50	84.17	84.38	84.17	84.12	84.07	84.43	84.48	84.17	84.22	84.02	83.97

Tabela 1: *Teste de parâmetros da MLP, mostrando a taxa de acertos dada uma combinação de parâmetros de aprendizado e momento. Em negrito, a melhor taxa de acerto dada a combinação de taxa de aprendizado e momento.*

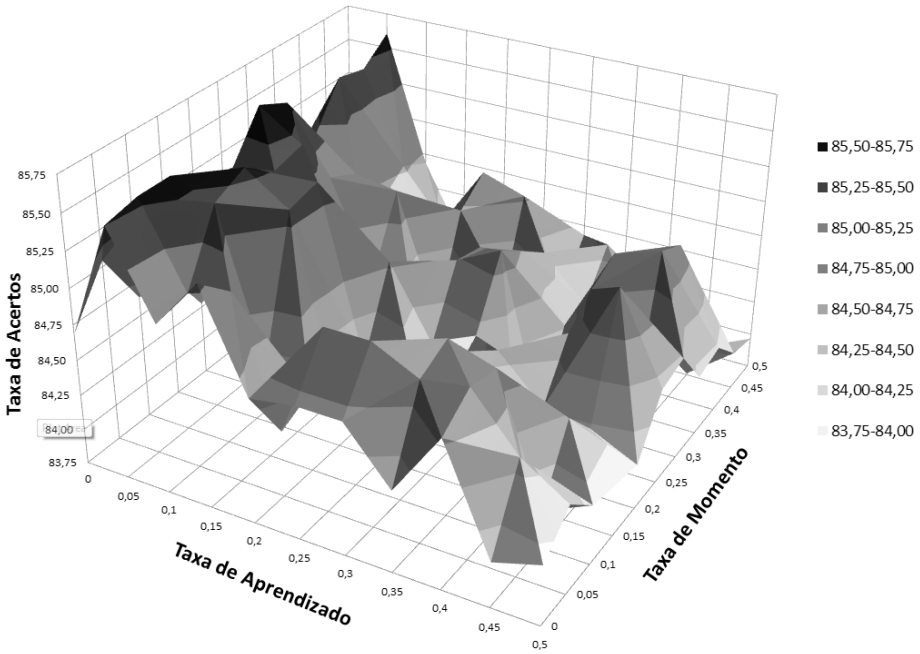

Figura 14: *Taxa de acertos em função das taxa de aprendizado e taxa de momento para o classificador MLP.*

	Extratores de Características						
	COOC	DWT	MORFO	COOC+DWT	COOC+MORFO	DWT+MORFO	COOC+DWT+MORFO
Bócio	0.879	0.790	0.676	0.862	0.867 (-1,37%)	0.747 (-5,44%)	0.863 (0,12%)
Indeterminado	0.553	0.287	0.307	0.567	0.637 (15,19%)	0.407 (41,81%)	0.640 (12,87%)
Papilífero	0.834	0.584	0.768	0.840	0.899 (7,79%)	0.763 (30,65%)	0.904 (7,62%)
Total	0.812	0.635	0.654	0.808	0.843 (3,82%)	0.701 (10,39%)	0.844 (4,46%)

(a) Taxas de acerto para o classificador KNN com algoritmos de extração de características

	Extratores de Características						
	COOC	DWT	MORFO	COOC+DWT	COOC+MORFO	DWT+MORFO	COOC+DWT+MORFO
Bócio	0.871	0.906	0.714	0.874	0.866 (-0,57%)	0.756 (-16,56%)	0.870 (-0,46%)
Indeterminado	0.553	0.033	0.217	0.557	0.647 (17,00%)	0.167 (406,06%)	0.677 (21,54%)
Papilífero	0.826	0.296	0.684	0.858	0.903 (9,32%)	0.699 (136,15%)	0.915 (6,64%)
Total	0.805	0.538	0.626	0.819	0.846 (5,09%)	0.644 (19,70%)	0.857 (4,64%)

(b) Taxas de acerto para o classificador MLP com algoritmos de extração de características

Tabela 2: *Taxa de acerto dos algoritmos de classificação entre as classes avaliadas. Em tons de cinza, os três cenários considerados para a experimentação com os extratores de características: (1) sem combinação; (2) combinação dois a dois; (3) combinação dos três.*

No segundo cenário, composto pela combinação dois a dois dos algoritmos de extração de características e apresentado nas colunas 4 a 6 das Tabelas 2(a) e 2(b), foi observada uma melhora na taxa de acertos quando considerada a combinação COOC+MORFO, alcançando o máximo de 84.6% para o classificador MLP. No último cenário, formado pela combinação dos três algoritmos de extração de características e apresentado na última coluna das Tabelas 2(a) e 2(b), atingiu-se a maior taxa de acertos com o classificador MLP (85.7%). Considerando separadamente as classes, observou-se uma taxa de 91.5% de acertos de núcleos relacionados a Câncer Papilífero, 87.0% de Bócio e 67.7% no caso do grupo Indeterminado.

Corroborando a contribuição esperada com este trabalho, foi verificado que a agregação de informação semântica representada pelo extrator MORFO pode auxiliar na tarefa de classificação. Para a experimentação, foi calculada a variação percentual da taxa de acertos dos extratores com e sem a adição do extrator MORFO. O extrator mais beneficiado com a adição de informação semântica foi o DWT, chegando a um ganho de 19,7% no classificador MLP em relação ao extrator sem a agregação de informação. Já a classe mais beneficiada foi a Indeterminado, no entanto, na classe Bócio foi observado uma queda na taxa de acertos na maioria dos casos.

Para verificar a razão da baixa taxa de acertos na classe Indeterminado, foi gerada a matriz de confusão para o pior caso observado (extrator DWT e classificador MLP). Na Tabela 3, verificou-se que a principal fonte de erros foi a classificação dos núcleos Indeterminados como sendo pertencentes da classe Bócio. Por se tratar de uma classe resultante de núcleos que o patologista não chegou em um consenso diagnóstico e, portanto, não apresenta um padrão de classificação interno, era esperado que a classe de Indeterminados possuísse uma baixa taxa de acertos.

	Bócio	Indeterminado	Papilífero
Bócio	835	0	87
Indeterminado	276	10	14
Papilífero	514	0	216

Tabela 3: *Matriz de Confusão para o pior caso (DWT-MLP). Nas colunas a classe predita e nas linhas a classe real.*

De acordo com a matriz de confusão apresentada na Tabela 4, observou-se que houve uma melhora no número de acertos para a classe Indeterminado resultante da combinação dois a dois dos extratores de características.

	Bócio	Indeterminado	Papilífero
Bócio	798	51	73
Indeterminado	72	194	34
Papilífero	59	12	659

Tabela 4: *Matriz de Confusão para a combinação COOC+MORFO e o classificador MLP, que obteve a melhor taxa de acerto entre as combinações dois a dois dos extratores. Nas colunas a classe predita e nas linhas a classe real.*

De acordo com a matriz de confusão apresentada na Tabela 5, a combinação dos três extratores de característica manteve a melhora na taxa de acertos, assim como foi observado anteriormente na Tabela 4.

	Bócio	Indeterminado	Papilífero
Bócio	802	47	73
Indeterminado	67	203	30
Papilífero	43	19	668

Tabela 5: *Matriz de Confusão para a combinação dos três extratores, considerando o algoritmo de classificação MLP. Nas colunas a classe predita e nas linhas a classe real.*

A avaliação do desempenho da classificação foi feita com base em curvas de Precisão-Revocação, apresentadas nas Figuras 15 a 28. Em sistemas de apoio ao diagnóstico, é desejável obter a maior precisão possível mesmo em um nível mais baixo de revocação, isto é, o classificador deve atingir níveis ótimos de precisão considerando que o conjunto de imagens analisadas pode ser pequeno. Essa necessidade advém do ambiente de trabalho e da sobrecarga imposta ao patologista, que pode usar um sistema de apoio ao diagnóstico suportado pelo FREDS para a redução das imagens que necessitam ser analisadas. Ao analisar as curvas de precisão e revocação, considerou-se como sendo o melhor caso a situação onde as três classes apresentavam simultaneamente curvas de precisão elevadas.

Da Figura 15 a Figura 21 são apresentadas as curvas de Precisão e de Revocação para o classificador MLP. Em cada figura desse intervalo varia a combinação de algoritmos de extração de características. No melhor caso da Figura 21, obtido pela combinação dos três extratores COOC+DWT+MORFO, a curva da precisão das três classes mantiveram-se próximas ao valor ótimo 1, apresentando um F-Measure de 0.856. F-Measure é uma média harmônica ponderada calculada entre a precisão e revocação, utilizada para resumir as duas medidas em um único número que varia de 0 a 1. Quanto mais próximo de 1, melhor é o classificador.

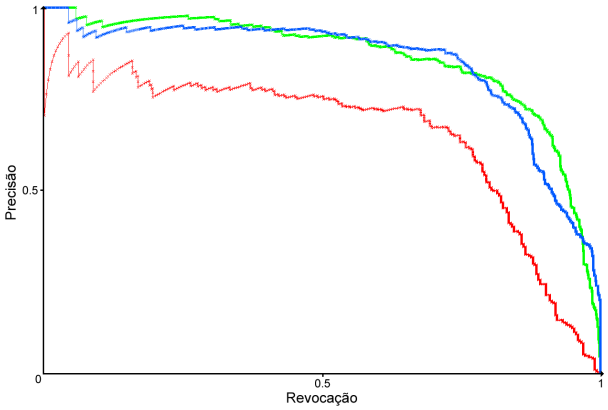

Figura 15: *Curvas de Precisão-Revocação para o extrator COOC e classificador MLP. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

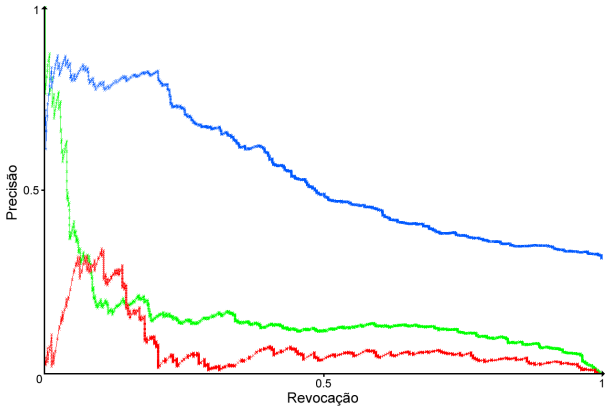

Figura 16: *Curvas de Precisão-Revocação para o extrator DWT e classificador MLP. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

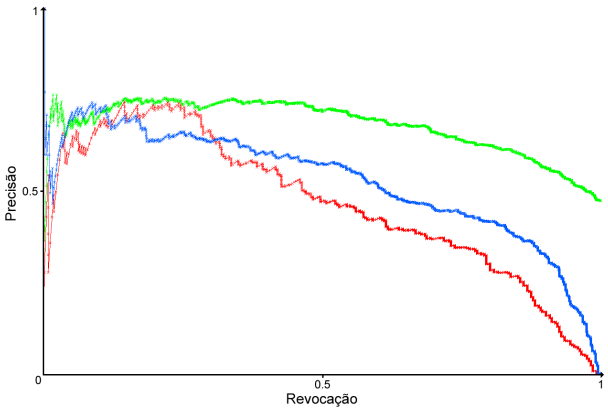

Figura 17: *Curvas de Precisão-Revocação para o extrator MORFO e classificador MLP. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

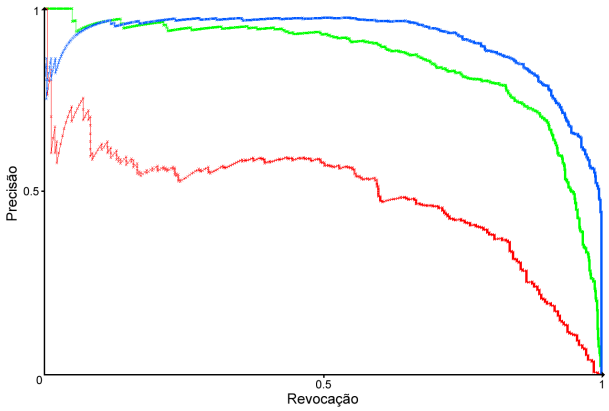

Figura 18: *Curvas de Precisão-Revocação para o extrator COOC+DWT e classificador MLP. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

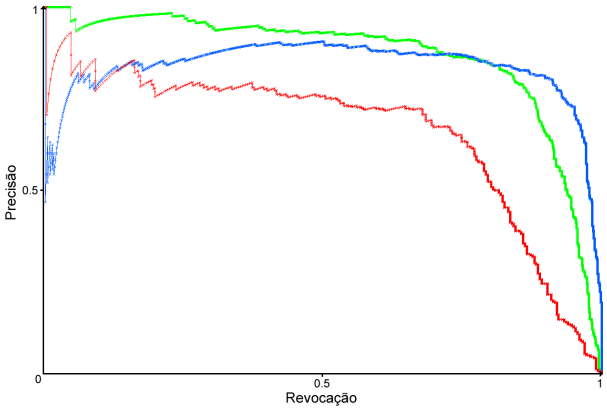

Figura 19: *Curvas de Precisão-Revocação para o extrator COOC+MORFO e classificador MLP. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

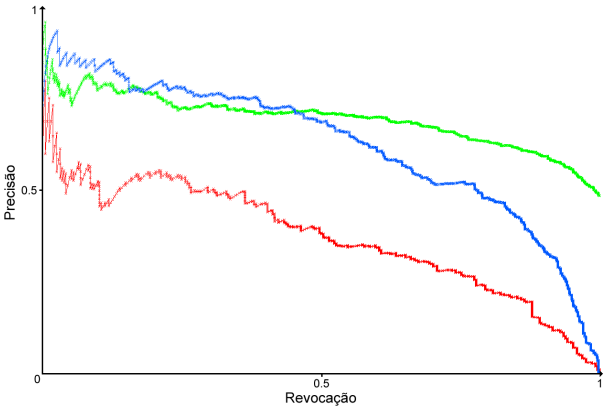

Figura 20: *Curvas de Precisão-Revocação para o extrator DWT+MORFO e classificador MLP. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

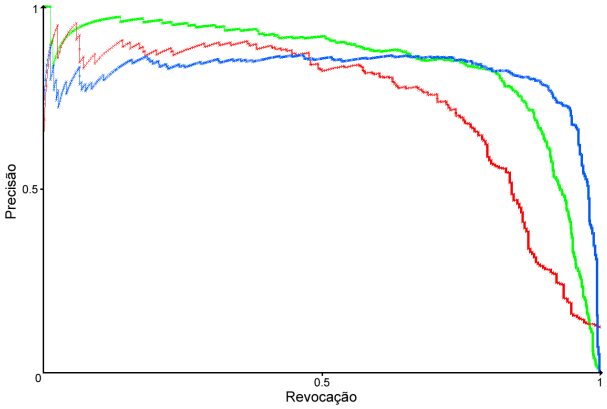

Figura 21: *Curvas de Precisão-Revocação para o extrator COOC+DWT+MORFO e classificador MLP. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

Das Figuras 22 a 28, são apresentadas as curvas de Precisão e de Revocação para o classificador KNN variando os cenários de combinação dos extratores. No melhor caso da Figura 28, representado pela combinação dos extratores COOC+DWT+MORFO, observou-se uma curva semelhante à obtida pelo MLP no seu melhor caso (ver Figura 21), com um F-Measure de 0.843, portanto, abaixo do MLP. No caso KNN, a queda da precisão global do classificador foi causada pelo decremento mais acentuado da classe Indeterminado.

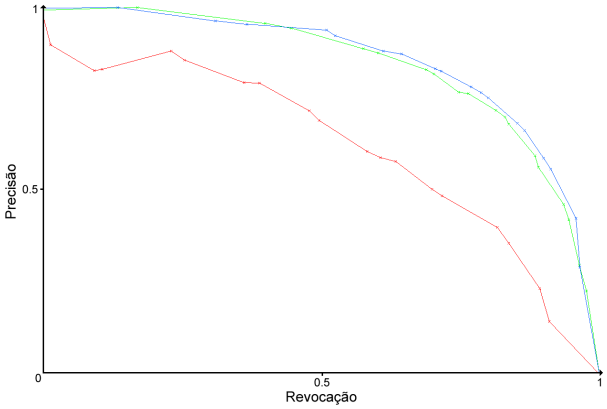

Figura 22: *Curvas de Precisão-Revocação para o extrator COOC e classificador KNN. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

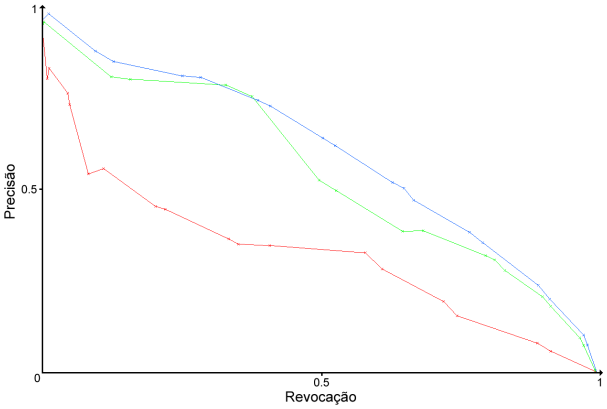

Figura 23: *Curvas de Precisão-Revocação para o extrator DWT e classificador KNN. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

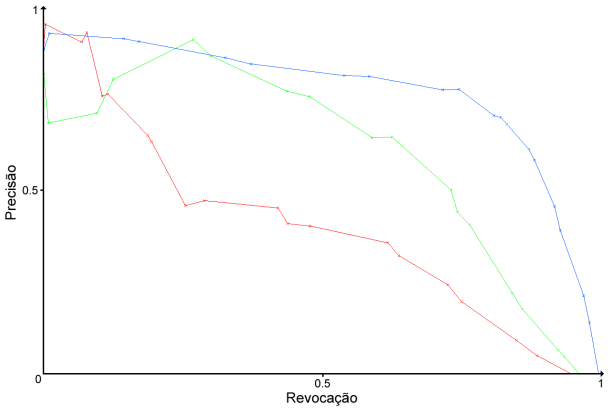

Figura 24: *Curvas de Precisão-Revocação para o extrator MORFO e classificador KNN. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

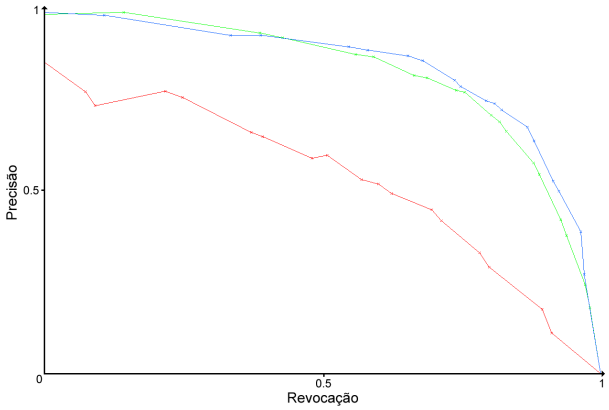

Figura 25: *Curvas de Precisão-Revocação para o extrator COOC+DWT e classificador KNN. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

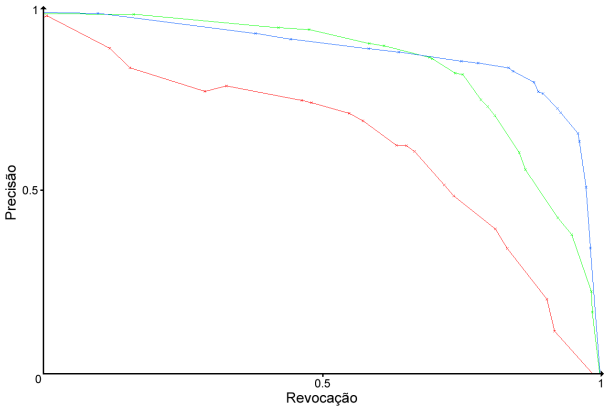

Figura 26: *Curvas de Precisão-Revocação para o extrator COOC+MORFO e classificador KNN. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

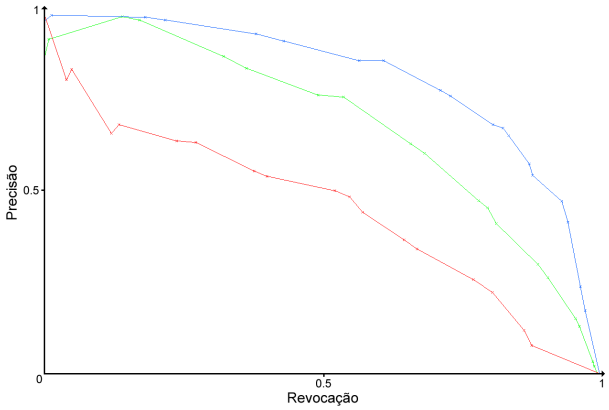

Figura 27: *Curvas de Precisão-Revocação para o extrator DWT+MORFO e classificador KNN. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

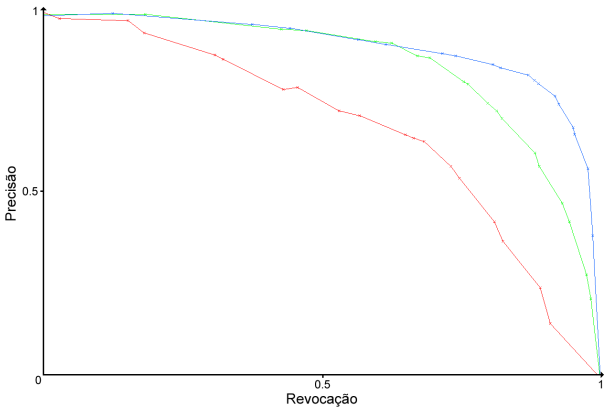

Figura 28: *Curvas de Precisão-Revocação para o extrator COOC+DWT+MORFO e classificador KNN. Em azul, a classe Papilífero, Bócio em verde e Indeterminado em vermelho.*

Após geradas as matrizes de confusão para cada classificador, foi calculado o coeficiente Kappa para cada combinação de extrator e classificador. O coeficiente Kappa de Cohen resume todas informações disponíveis em uma matriz de confusão, medindo o grau de concordância entre um experimento realizado e observações feita ao acaso, isto é, frutos de uma classificação aleatória [COHEN, 1960]. Em uma situação

onde estatisticamente não há diferença entre um experimento e uma observação aleatória, o coeficiente Kappa assume valor $\kappa = 0$, sendo o valor máximo $\kappa = 1$ atribuído somente quando o classificador apresenta total concordância, isto é, não há nenhuma evidência de que a classificação tenha sido fruto do acaso.

Embora não exista um critério único para avaliar o valor do coeficiente Kappa, [LANDIS; KOCH, 1977] define uma escala de interpretação para o coeficiente Kappa, onde valores no intervalo 0.0-0.20 indicam uma pequena concordância; 0.21-0.40 sendo razoável; 0.41-0.60 como moderada; 0.61-0.80 como substancial e por fim, 0.81-1.00 como sendo uma concordância quase total.

A Tabela 6 apresenta os valores de Kappa para todas combinações de extratores e classificadores. O maior coeficiente observado foi de 0.7659 para o cenário de extratores COOC+DWT+MORFO e o classificador MLP, situando-se na faixa substancial e próximo à faixa de concordância quase total, segundo o critério apresentado anteriormente.

	Coeficiente Kappa (κ)	
Extratores	KNN	MLP
COOC	0.6902	0.6818
DWT	0.3787	0.1509
MORFO	0.4208	0.3577
COOC+DWT	0.6853	0.7024
COOC+MORFO	0.7426	0.7465
DWT+MORFO	0.5002	0.3841
COOC+DWT+MORFO	0.7447	0.7659

Tabela 6: *Estatística Kappa gerada a partir de todas combinações de extratores e classificadores.*

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho foi desenvolvido um sistema de apoio ao diagnóstico, instanciado a partir da especificação do *framework* FREDS. Esta especificação foi iniciada no projeto de Iniciação Científica do aluno (Processo FAPESP 2007/08470-5), naquele momento sendo aplicada em lesões pulmonares e imagens radiológicas obtidas por exames de tomografia computadorizada. A especificação foi expandida e modificada durante o presente trabalho (Processo FAPESP 2008/08098-1), acrescentando-se um módulo de Morfometria e ampliando os módulos de Recuperação de Informação, para (i) dar continuidade à proposta inicial de criação de um *framework* modular para processamento de imagens médicas e informação semântica e (ii) acoplar as diferenças de manipular imagens radiológicas e microscópicas. O *framework* FREDS é uma abstração de processos que podem instanciar a utilização de informações semânticas como, por exemplo, dados morfométricos, na análise de imagens.

Uma maneira de construir um mapeamento entre imagens médicas e laudos é a partir da rotulação de uma imagem. Com a rotulação automatizada foi possível identificar os achados patológicos presentes em uma imagem não-diagnosticada, a partir dos quais obtiveram-se termos relacionados a diagnósticos de câncer pela busca textual em uma coleção de laudos. Uma segunda forma é a extração de informações de uma imagem que apresentem significado semântico como, por exemplo, dados morfométricos.

O uso de dados morfométricos torna possível a interpretação da saída dos algoritmos pelo patologista, quantificando informações que ele já está habituado a utilizar, como área e raio nuclear. Os resultados obtidos indicam que o uso de morfometria como informação adicional durante a classificação de núcleos tireoidianos pode ocasionar um aumento na precisão dos algoritmos de processamento de imagens, além de tornar os resultados mais fáceis de serem interpretados pelo usuário. Desta forma, os resultados obtidos neste trabalho podem ser considerado um elo entre o conteúdo da imagem e a descrição do laudo, permitindo que o patologista tenha acesso a informações adicionais referentes a uma imagem além de sua classificação.

Além de auxílio ao diagnóstico, outras oportunidades de aplicação do *framework* FREDS são vislumbradas, por exemplo, nos contextos de aprendizado eletrônico, de sistemas de vigilância e, em especial, de auditoria de exames. Para patologistas com menos experiência ou residentes, o sistema resultante da pesquisa proposta poderá prover um ambiente de aprendizado eletrônico, fornecendo informações relevantes para o diagnóstico como, por exemplo, palavras-chave obtidas de exames similares e dados morfométricos. Outra aplicação do uso de palavras-chaves seria apoiar técnicas de Recuperação de Imagens Baseada em Conteúdo que, além de utilizar os atributos extraídos de uma imagem, pode também utilizar palavras-chaves para aumentar sua precisão na busca de imagens similares.

Outro cenário considerado para a aplicação do *framework* é o desenvolvimento de sistemas de auditoria de conteúdo. Neste caso, o sistema auxiliaria na observação do conteúdo escrito pelo patologista para verifi-

car se este segue o padrão semântico e sintático de informação esperado para aquele tipo de imagem dentro do nível de confiança fornecido pelo sistema. Acredita-se poder realizar esse tipo de auditoria através da comparação de um laudo com laudos já escritos para imagens similares. Uma variação desse sistema pode ser voltada para sistemas de vigilância em saúde. Um mecanismo de vigilância poderia alertar patologistas sobre possíveis erros de diagnóstico ou ausência de informações esperadas para o diagnóstico informado.

Como extensão direta deste trabalho, pretende-se ampliar a experimentação incluindo um número maior de exames e realizar efetivamente uma avaliação de laudos considerando técnicas de Recuperação de Informação e Processamento de Linguagem Natural, que foi iniciada no Anexo A.

Anexo A - Experimentação com o componente de Seleção de Laudos

Os Quadros 2(a), 2(b) e 3 apresentam resultados obtidos por meio da experimentação com a base de laudos. Considerou-se como entrada os rótulos “pseudo-inclusão” e “fendas”. Esses rótulos representam achados patológicos que são frequentemente associados à câncer papilífero.

“papilífero” = 0.1079
“folicular” = 0.0596
“microcarcinoma” = 0.0568

a) n-gram = 1

“compatível microcarcinoma papilífero” = 0.1152
“lesão padrão folicular” = 0.1152
“papilífero recomenda-se confirmação” = 0.0578

b) n-gram = 3

Quadro 2: *Conjuntos de termos mais relevantes e score obtido*

O Quadro 2(a) contém os três termos de maior *score* encontrados em 5 exames retornados pelas expressões de busca formadas pelos rótulos. Já o Quadro 2(b) apresenta a mesma consulta, mas considerando o parâmetro n-gram = 3 nos algoritmos de Recuperação de Informação. O uso de n-gram retornou um conjunto de termos mais significativo segundo o especialista, pois as palavras retornadas sem o uso de n-gram não apresentam significado quando aparecem isoladas. Os termos “papilífero” e “folicular” não demonstram o grau de desenvolvimento do carcinoma, apenas definem o seu tipo. Ao aparecerem associados a outros termos, como “lesão”, “microcarcinoma” ou “carcinoma”, os conjuntos passam a ter um significado maior para o patologista.

O baixo *score* dos termos é explicado pelo fato da base atual de exemplos conter uma baixa variedade de diagnósticos, pois obteve-se acesso apenas aos laudos dos exames selecionados para captura de imagens, o que restringe a coleção de documentos em câncer papilífero, bócio coloide e indeterminados. Consequentemente, os termos aparecem com frequência maior na coleção de documentos e seu *score* é reduzido pela medida TF-IDF, de acordo com a equação (2.19).

“Quadro morfológico compatível com carcinoma papilífero. Requer confirmação histopatológica” = 0.3309
“Quadro morfológico sugestivo de lesão de padrão microfolicular a esclarecer (hiperplásica x neoplásica)” = 0.2514
“Quadro compatível com microcarcinoma papilífero. Recomenda-se confirmação histopatológica com exame per-operatório de congelação” = 0.2065

Quadro 3: *Exemplo contendo três diagnósticos selecionados e score obtido*

O Quadro 3 mostra três diagnósticos que obtiveram o maior *score* durante a seleção automatizada dos laudos a partir dos rótulos de entrada. Segundo o especialista, os documentos retornados estão de acordo com o esperado para os rótulos de imagem considerados (“pseudo-inclusão” e “fendas”), pois representam termos associados à câncer papilífero.

Referências Bibliográficas

ACHARYA, T.; RAY, A. *Image processing: principles and applications*. [S.l.]: Wiley-Interscience, 2005.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval*. [S.l.]: ACM press New York, 1999.

CHEN, Y. T.; HOU, C. J.; LEE, M. W.; CHEN, S. J.; TSAI, Y. C.; HSU, T. The image feature analysis for microscopic thyroid tissue classification. *Conf Proc IEEE Eng Med Biol Soc*, p. 4059–4062, 2008.

COELI, C.; BRITO, A.; BARBOSA, F.; RIBEIRO, M.; SIEIRO, A.; VAISMAN, M. Incidência e mortalidade por câncer de tireóide no Brasil. *Arq Bras Endocrinol Metab*, v. 49, p. 503–509, 2005.

COHEN, J. et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, v. 20, n. 1, p. 37–46, 1960.

CONNERS, R. W.; HARLOW, C. A. A theoretical comparison of texture algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-2, n. 3, p. 204–222, 1980.

DASKALAKIS, A.; KOSTOPOULOS, S.; SPYRIDONOS, P.; GLOTSOS, D.; RAVAZOULA, P.; KARDARI, M.; KALATZIS, I.; CAVOURAS, D.; NIKIFORIDIS, G. Design of a multi-classifier system for discriminating benign from malignant thyroid nodules using routinely h&e-stained cytological images. *Comput. Biol. Med.*, Pergamon Press, Inc., Elmsford, NY, USA, v. 38, n. 2, p. 196–203, 2008. ISSN 0010-4825.

DAUBECHIES, I. *Ten lectures on wavelets*. [S.l.]: Society for Industrial and Applied Mathematics, 1992.

DEVITA, V.; HELLMAN, S.; ROSENBERG, S. et al. *Principles and practice of oncology*. 8. ed. [S.l.]: JB Lippincott, Philadelphia, 1997.

DINA, R.; CAPITANIO, A.; DAMIANI, S. A morphometric analysis of cytological features of tall cell variant and classical papillary carcinoma of the thyroid. *Cytopathology*, v. 11, n. 2, p. 124–128, 2000.

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

FELIG, P.; BAXTER, J.; FROHMAN, L.; BROADUS, A. *Endocrinology and Metabolism*. 4. ed. [S.l.]: McGraw-Hill, 2001.

GABRIEL, J.; EBRARY, I. *The biology of cancer*. [S.l.]: Wiley Online Library, 2007.

GHARIB, H. et al. Fine-needle aspiration biopsy of thyroid nodules: advantages, limitations, and effect. In: *Mayo Clinic Proceedings*. [S.l.: s.n.], 1994. v. 69, n. 1, p. 44.

GOLBECK, J.; FRAGOSO, G.; HARTEL, F. et al. The National Cancer Institute's thesaurus and ontology. *Journal of web semantics*, Citeseer, v. 1, n. 1, p. 75–80, 2003.

GOSLING, J.; MCGILTON, H. The java language environment. *Sun Microsystems Computer Company*, 1995.

GUPTA, N.; SARKAR, C.; SINGH, R.; KARAK, A. Evaluation of diagnostic efficiency of computerized image analysis based quantitative nuclear parameters in papillary and follicular thyroid tumors using paraffin-embedded tissue sections. *Pathology & Oncology Research*, Springer, v. 7, n. 1, p. 46–55, 2001.

HAAR, A. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, Springer, v. 69, n. 3, p. 331–371, 1910.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, v. 3, n. 6, p. 610–621, 1973.

HARMS, H.; HOFMANN, M.; RUSCHENBURG, I. Fine needle aspiration of the thyroid: can an image processing system improve differentiation? *Anal Quant Cytol Histol*, v. 24, n. 3, p. 147–153, 2002.

HAYKIN, S. *Redes Neurais Princípios e Práticas*. [S.l.]: Bookman, Porto Alegre, 2001.

HIPP, D.; KENNEDY, D. *SQLite*. Disponível em: <http://www.sqlite.org/>. Acesso em: 28 fev. 2012.

HOLMES, G.; DONKIN, A.; WITTEN, I. Weka: A machine learning workbench. In: IEEE. *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*. [S.l.], 1994. p. 357–361.

JEMAL, A.; BRAY, F.; CENTER, M. M.; FERLAY, J.; WARD, E.; FORMAN, D. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, Wiley Subscription Services, Inc., A Wiley Company, v. 61, n. 2, p. 69–90, 2011. ISSN 1542-4863. Disponível em: <http://dx.doi.org/10.3322/caac.20107>.

JR, J. P.; MONTEIRO, R.; ZAN, T.; AZOUBEL, R.; SANTI, D. D.; TABOGA, S.; MARTINS, A. Morphometric analysis of nucleus and nucleolar organizer regions (nors) in tongue squamous cell carcinoma (scc). *International Journal of Morphology*, v. 25, n. 3, p. 493–499, 2007.

KAO, A.; POTEET, S. *Natural language processing and text mining*. [S.l.]: Springer Verlag, 2007.

LANDIS, J.; KOCH, G. The measurement of observer agreement for categorical data. *Biometrics*, p. 159–174, 1977.

MAIA, A.; WARD, L.; CARVALHO, G.; GRAF, H.; MACIEL, L.; WAISMAN, M. et al. Consenso brasileiro em nódulos de tireóide e câncer diferenciado de tireóide, Departamento de Tireóide–Sociedade Brasileira de Endocrinologia e Metabologia. *Arq Bras Endocrinol Metab*, v. 51, n. 5, 2007.

MELMED, S.; POLONSKY, K.; LARSEN, P.; KRONENBERG, H. *Williams Textbook of Endocrinology*. 11. ed. [S.l.]: New York, NY: Elsevier, 2008.

MUEEN, A.; ZAINUDDIN, R.; BABA, M. S. Automatic multilevel medical image annotation and retrieval. *J. Digital Imaging*, v. 21, n. 3, p. 290–295, 2008.

NÉVÉOL, A.; DESERNO, T.; DARMONI, S.; GÜLD, M.; ARONSON, A. Natural language processing versus content-based image analysis for medical document retrieval. *Journal of the American Society for Information Science and Technology*, v. 60, n. 1, p. 123–134, 2009.

ORENGO, V. *Assessing relevance using automatically translated documents for cross-language information retrieval*. Tese (Doutorado) — School of Computing Science, Middlesex University, London, 2004.

OTSU, N. A threshold selection method from gray-level histograms. *Automatica*, v. 11, p. 285–296, 1975.

PANNO, J. *Cancer: the role of genes, lifestyle, and environment*. [S.l.]: Facts on File, 2004.

PESSOTTI, H. C.; AZEVEDO-MARQUES, P. M. de; MACEDO, A. A. Framework para classificação automática de tomografias computadorizadas de alta resolução para auxílio ao diagnóstico de lesões intersticiais de pulmão. In: *Anais do XXI Congresso Brasileiro de Engenharia Biomédica (CBEB 2008)*. Salvador, Brazil: [s.n.], 2008.

PESSOTTI, H. C.; MURTA-JUNIOR, L. O.; SOARES, E. G.; MACEDO, A. A. Freds: Framework para redução da descontinuidade semântica em imagens médicas. In: *XI Workshop de Informática Médica (WIM 2011) - em publicação*. Natal, Brazil: [s.n.], 2011.

PETROU, M.; PETROU, C. *Image processing: the fundamentals*. [S.l.]: Wiley, 2010.

PORTER, M. An algorithm for suffix stripping. *Program*, v. 14, n. 3, p. 130–137, 1980.

PRIYA, S.; SUNDARAM, S. Morphology to morphometry in cytological evaluation of thyroid lesions. *Journal of Cytology*, v. 28, n. 3, p. 98, 2011.

RAJESH, L.; SAHA, M.; RADHIKA, S.; RADOTRA, B. D.; RAJWANSHI, A. Morphometric image analysis of follicular lesions of the thyroid. *Anal Quant Cytol Histol*, v. 26, n. 2, p. 117–120, 2004. ISSN 0884-6812.

RASBAND, W. *ImageJ: A public domain Java image processing program. National Institute of Mental Health, Bethesda, Maryland, USA*. 2008.

ROBBINS, S.; KUMAR, V.; ABBAS, A.; COTRAN, R.; FAUSTO, N. *Robbins and Cotran Pathologic Basis of Disease*. 8. ed. [S.l.]: WB Saunders Company, 2009.

ROSENFELD, A.; PFALTZ, J. Sequential operations in digital picture processing. *Journal of the ACM (JACM)*, ACM, v. 13, n. 4, p. 471–494, 1966.

SALTON, G.; LESK, M. Computer evaluation of indexing and text processing. *Journal of the ACM (JACM)*, ACM, v. 15, n. 1, p. 8–36, 1968.

ULMANN, S. *Semântica: uma introdução à ciência do significado*. 3. ed. [S.l.]: Fundação Calouste Gulbenkian, Lisboa, 1964.

VIERA, A.; VIRGIL, J. Uma revisão dos algoritmos de radicalização em língua portuguesa. *Information Research*, v. 12, n. 3, p. 12–3, 2007.

WANG, J.; WIEDERHOLD, G.; FIRSCHEIN, O.; WEI, S. X. Content-based image indexing and searching using Daubechies' wavelets. *Int. J. Digit. Libraries*, v. 1, n. 4, p. 311–328, 1998.

WANG, W.; OZOLEK, J.; ROHDE, G. Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry Part A*, Wiley Online Library, v. 77, n. 5, p. 485–494, 2010.

WRIGHT, R.; CASTLES, H.; MORTIMER, R. Morphometric analysis of thyroid cell aspirates. *Journal of Clinical Pathology*, v. 40, n. 4, p. 443, 1987.

YANG, Y.; LIU, X. A re-examination of text categorization methods. In: ACM NEW YORK, NY, USA. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.], 1999. p. 42–49.